Puretoons: A Multi-Modal Approach to Content Filtering

Muhammad Ibrahim Jawaid (k21-4933)
Muhammad Kashaaf (k21-3380)
Muhammad Moiz (k21-4508)

Department of Computer Science
FAST NUCES
{k214933, k213380, k214508}@nu.edu.pk *

Abstract

The exponential rise of the YouTube channels that are aimed at young audiences has led to concerns about what young viewers consume online. Regardless of the measure taken by the platform, unsuitable content, which is usually masked as child-targeted media, can slip through filters to reach underage audiences. This paper's goal is to improve several natural language processing (NLP), and models of visual deep learning intended to detect unsafe content in the children's videos. With the lack of specialized data for such content, we start gathering a complete dataset of cartoon videos, including the textual and visual portions. Our approach is training several models with this dataset and performing comparative analysis to see which model is best to use to detect harmful content. The results of this study can improve automated content filtering on video-sharing sites, where the digital lives of the children are safer.

Keywords: BERT, Natural Language Processing, Content analyses, Video analyses, Child Safety

*__Supervisor:__ Prof. Shoaib Rauf – __Co-supervisor:__ Dr. M. Shahzad

1 Introduction

With the rapid extension of social media, it has become the most important source of information and entertainment for people of all ages. Among all consumers of online content, children are the most popular and high in proportion. Children mostly consume content in the form of animated cartoons, while these cartoons are often designed to entertain and educate kids, there is a growing concern with the question: Are these cartoons even safe? Recently a high presence of LGBTQ+, excessive violence, sexually suggestive and vulgar themes have been noticed in cartoons. These themes are depicted either visually or through natural language, which can negatively influence children's cognitive and emotional development.

The content moderation systems used by prominent platforms such as Youtube predominantly rely on user reports and metadata. However, it is very difficult and unreliable to depend upon such mechanisms when the content uploaded every minute is huge. In addition, many content moderation techniques focus on visual or auditory in isolation, because of which they are ineffective. Also, the use Long Short-Term Memory (LSTM) networks in detecting themes in nlp, run short to detect contextual themes, due to their inability to retain long term dependencies and long-term context. Additionally, the systems employing either visual or natural language analysis are not accurate in cartoon styled content due to very less datasets available, which brings in another problem to solved, of achieving more accurate results by collecting data for this research along which opening doors for future research in this area. But it is just not limited to this because there is a need to discover which Deep Neural Networks(DNN) technique is the best suited to achieve the most accurate results is also a challenge.

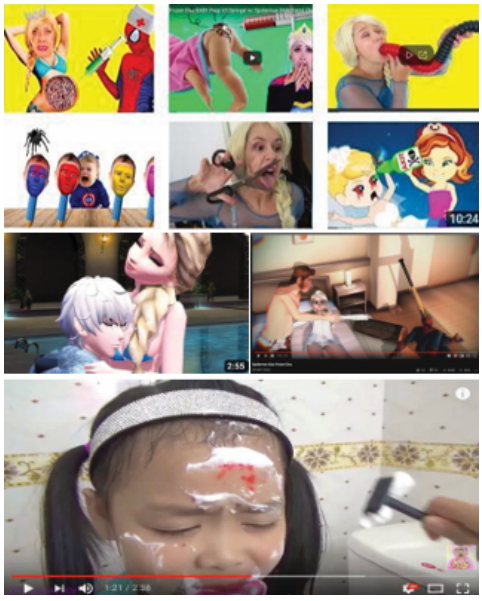

Figure 1

To address these challenges, this research paper proposes a novel multi-modal approach that leverages both natural language and visual analysis in the detection of inappropriate themes in cartoon styled content. At the heart of this system is Transformer-based Natural Language processing model, specifically BERT, which detects context from transcripts, or subtitles. Along with YOLO (You Only Look Once) being employed to detect visual elements across various inappropriate categories.

This research contributes to the field by introducing a comprehensive and context-aware solution in safeguarding children from inappropriate content.

2 Related Work

Recent advancements in content moderation for children's media highlight the limitations of existing approaches. For example, Chuttur et al. [1] used a multi-modal classifier combining BiLSTM and VGGNet for textual and visual analysis, respectively. While their model showed improvements, BiLSTMs have inherent limitations in capturing long-range dependencies, which can hinder their performance in detecting inappropriate content across fragmented or contextually complex cartoon sequences. Also their visual approach had a major limitation that they utilized cartoon characters dataset labeled as inappropriate, but the better approach would be using the clips of the cartoon shows that these characters are part of.

Papadamou et al. [2] employed a four-branch classifier model that utilizes metadata such as titles, tags, thumbnails, and video statistics. However, their approach relies heavily on metadata, which can be incomplete or inaccurate, and does not address the specific challenges of cartoon content.

Binh et al. [3] explored using BERT for subtitle analysis, achieving high accuracy in detecting inappropriate content through textual data alone. While their approach is promising, it focuses solely on subtitles without incorporating visual data, limiting its effectiveness in detecting context-dependent inappropriate content in cartoons. Ahmed et al. [4] compared 3D CNN, LSTM, and VTN for video classification, with VTN showing superior performance. However, their study did not account for the diverse and nuanced nature of inappropriate content, including specific themes like LGBTQ+ issues or substance abuse.

Gkolemi et al. [5] concentrated on detecting channels that upload inappropriate videos, which does not fully address the need for detailed content moderation at the video level, especially in cartoons that may have varied and complex content. In contrast to these methods, our approach leverages Transformer models to capture global context and improve the detection of inappropriate content. Transformers' self-attention mechanisms enable a more nuanced understanding of both visual and textual features, making them well-suited to handle the complexities of cartoon content.

Yousaf et al. [6] adopted EfficientNet-B7 and BiLSTM for detecting inappropriate content in video classification tasks. While this approach achieved competitive accuracy, it was limited to a narrow set of content categories, overlooking crucial elements such as weapons, racist symbols, nudity, and pride flags—content that could have a lasting impact on children. Moreover, applying this method for real-time object detection could increase computational time, particularly in cases where capturing context is unnecessary, such as those in our proposed solution.

As demonstrated by Ramachandra et al. [7], YOLO offers significant advantages over other object detection algorithms. YOLO is designed as a unified model that can be efficiently trained using a simple loss function, in contrast to more complex CNN-based models, making it highly suitable for real-time applications.

3 Methodology

This section outlines the architecture, data processing, and implementation strategies used to develop our multi-modal system for detecting inappropriate content in cartoons. The proposed approach combines Natural Language Processing (NLP) and Computer Vision techniques to analyze both textual and visual elements of cartoon videos. Specifically, we use three different variants of BERT, a Transformer-based language model, to understand the semantic context of subtitles or dialogues, and YOLO, a real-time object detection algorithm, to identify visual indicators of inappropriate content. The outputs from these modules are then integrated to perform final content classification. The following subsections detail the dataset, preprocessing steps, model components, and evaluation procedures used in this research.

3.1 System Overview

In this research, we propose a multi-layered architecture designed to detect inappropriate content in cartoon videos by leveraging both textual and visual information. The system begins with a video input, which is processed in the first stage to extract both individual video frames and the corresponding subtitle transcript. These extracted elements are then forwarded to separate processing pipelines, as illustrated in Figure 2.

The subtitle transcript is passed to the Natural Language Processing (NLP) module, where it is analyzed using a Transformer-based model (BERT) to detect potentially inappropriate textual content. Simultaneously, the extracted video frames are sent to the Visual Processing module, where an object detection algorithm (YOLO) is used to identify visual elements that may be inappropriate for children. The NLP module outputs detected content classes based on semantic analysis of the transcript, while the visual module returns bounding boxes and class labels for objects identified in the frames.

This layered, modular approach enables the system to independently analyze and interpret both language and imagery, which are then combined in a later stage for final classification. A prototype of this architecture has been implemented as a Minimum Viable Product (MVP) to evaluate the feasibility and performance of the proposed method. .

3.2 Dataset Collection and Annotation

As discussed earlier, the shallow dataset availability would have limited this research to less accurate results. To develop a more accurate multi-modal content moderation system, we curated a custom dataset for both Natural Language, and Image processing. The dataset for natural Language was collected manually from transcripts of cartoons, by picking up dialogues and labelling according to the classes they fall in. The dataset was curated in a way such that the trained model is not keyword dependent to detect classes, but gets the context to work. We collected data belonging to the classes: **Vulgar**, **LGBTQ+**, **Hate-speech**, **mature**, and **Neutral**. While the models was trained on initial set of data, it showed key word dependency to detect classes, which increased the amount False Negatives, as when a keyword was used in a good context it still detected it. Keeping this in mind, the next set of data collected was focused on neutral dialogues with keywords, along with

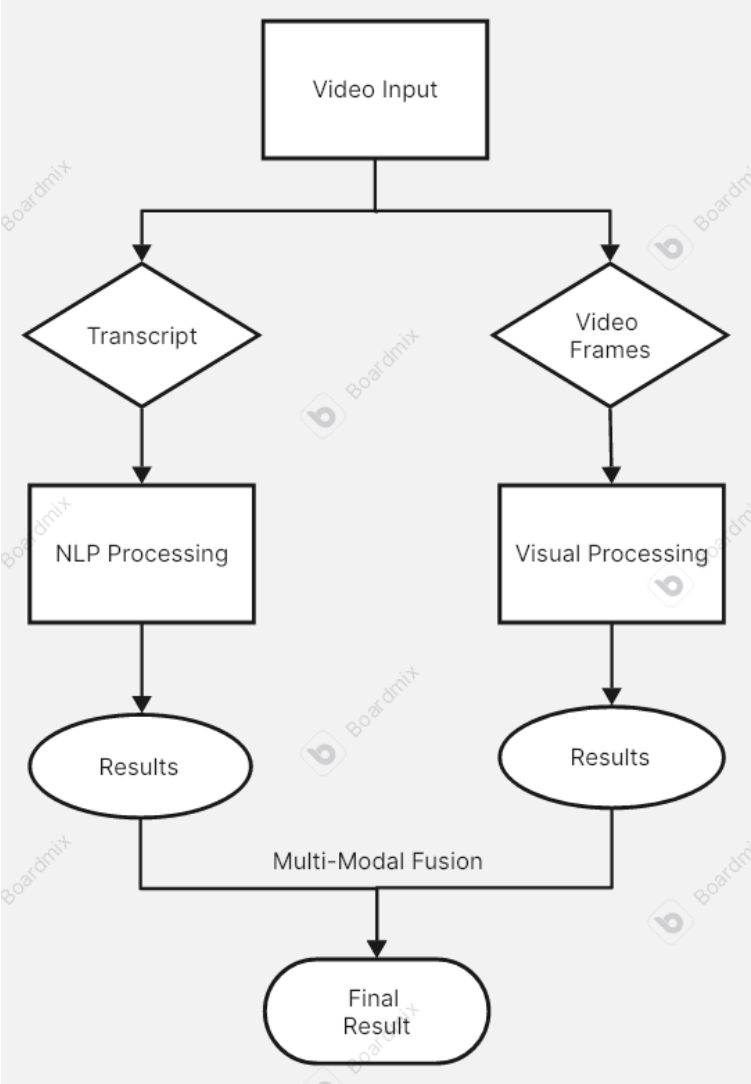

Figure 2: System Architecture

keyword independent dialogues for each class. Textual data has 3 columns, namely dialogue 1, dialogue 2, and class label for each record. Visual dataset was generated using GenAI tools, and was collected in the following classes: **weapon**, **LGBTQ+**, **nudity**, **blood**, and **kiss**. Data was generated and then annotated by creating bounding boxes using roboflow.

3.3 Text Preprocessing and Text Analysis

To perform Natural Language Processing (NLP) analysis on the subtitle transcripts, we utilized **BERT** (Bidirectional Encoder Representations from Transformers), a state-of-the-art pre-trained language model developed by Google in 2018. We experimented with three variants of **BERT** (Bidirectional Encoder Representations from Transformers) namely: **BERT-base**, **Distil-BERT**, and **simple BERT** . It is based on the Transformer architecture and is designed to understand the context of words in a sentence by looking at both the left and

right sides of a word simultaneously. It is trained on massive text corpora which makes it easier to achieve task specific accuracy, even with less data which cannot be achieved if a model is trained from scratch. Data contained dialogues belonging to five distinct classes as shown in Fig.3.3.1. Each model was fine-tuned on this labeled dataset to perform multi-class classification. To further improve model accuracy and reduce the incidence of false positives and false negatives, we implemented a **contrastive learning pipeline**. This involved training the model to differentiate between semantically similar and dissimilar dialogue samples, thus enhancing its ability to detect subtle inappropriate content. The best Model was determined through comparative analysis, and is best explained under the **Results** section.

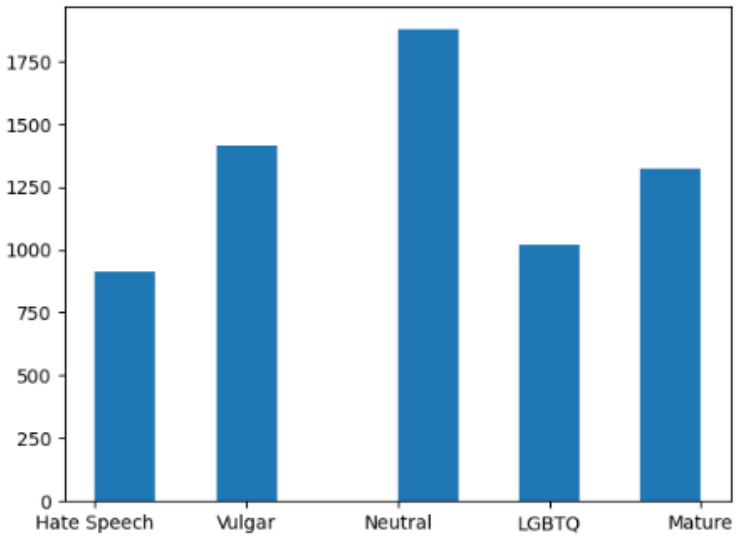

Figure 3: NLP Dataset Count

3.4 Image Preprocessing and Visual Analysis

To train a model which can detect inappropriate classes visually, we generated data through GenAI such as MetaAI, Chatgpt, and etc. The dataset contained annotated cartoon images containing various potentially visual elements in bounding boxes. For object detection, we trained and evaluated three state-of-the-art deep learning models: **Faster R-CNN**, **RF-DETR**, and **YOLOv11**. These models were trained to conduct a comparative analysis for accuracy, and granularity. Each model was trained on our annotated dataset to localize and classify inappropriate content through bounding boxes. Bounding box annotations were created using the **Roboflow** platform, which provided an intuitive interface for labelling and ensured consistency across the dataset. All images were resized to a uniform resolution depending on the model requirements to maintain input consistency. Evaluation metrics such as **mean Average Precision (mAP)**, **Precision**, and **Recall** were used to compare the performance of the models. A comparative analysis of these metrics is presented in the **Results** section, where we identify the most effective model for visual content moderation.

3.5 Multi-modal Feature Fusion

The general framework for our study is presented in Fig.1. Generally when a video is uploaded, it is divided into two parts. One being the video transcript that will serve as a input for the nlp model, which will detect inappropriate classes in language. On the other hand, the other part would contain random video slices to decrease the computational time, which would serve as input to the visual model. This would then give a more accurate and confident result for the video combined. The fusion of the different modalities not only could enhance the classification results, but also could increase the confidence level of the final decision. Once the results are obtained for each modality, then those could be fused to obtain the final results. The fusion can be accomplished using many approaches depending upon the architecture.

3.6 Model Training and Evaluation

Each modality in our system was trained independently before being integrated into a unified decision pipeline. The NLP models—including BERT-base, DistilBERT, and BERT—were trained using a BERT built-in loss function, optimized using the Adam optimizer with a learning rate of $2 \times e^{-5}$. Models were trained for 3 epochs. Accuracy scores and F1-scores were used to evaluate the model.

For the visual stream, object detection models (Faster R-CNN, RF-DETR, and YOLOv11) were trained using annotated frames, with bounding box regression and classification losses. YOLOv11, in particular, was trained for 100 epochs using the COCO loss structure, with a batch size of 16 and input image resolution of 640×640. Evaluation was done using precision, recall and F1 scores.

4 Results

4.1 Comparative Analysis between NLP Models

The accuracy comparison outcomes of the four NLP models—**BERT-Base**, **RoBERTa**, **DistilBERT**, and **Contrastive Learning + BERT-Base** bring forth significant trends of Transformer-based language models for the task of inappropriate content recognition in cartoon transcripts. The results are Shown in (Table 1).

Model	**Accuracy Score**
BERT-Base	0.83
RoBERTa	0.84
DistilBERT	0.82
Contrastive Learning + BERT-Base	0.87

Table 1: Accuracy scores of different Transformer models fine-tuned on the cartoon transcript dataset

The baseline model, BERT-Base, performed very well (at about 83% accuracy) owing to the capability to acquire bidirectional context and semantic relationship information from

big pretraining datasets. RoBERTa, which alters BERT's training dynamics by removing the Next Sentence Prediction (NSP) task, applying dynamic masking, and training from a larger and more diverse dataset, slightly exceeded it at 84% accuracy. These modifications enable RoBERTa to more effectively model long-range dependencies as well as subtle linguistic cues, which are especially pertinent in recognizing inappropriate content. DistilBERT, a smaller version distilled from BERT, scored slightly lower (82%) as anticipated by its smaller size and parameters but still provides a fine compromise between speed and accuracy for low-resource settings.

The standout performer in this test was Contrastive Learning + BERT-Base, with the highest accuracy of about 87%. This performance gain is due to the addition of contrastive learning, a self-supervised method that enhances the model's capacity to separate semantically dissimilar versus similar data points. In a technical sense, contrastive learning achieves this by moving positive pairs (such as related or similar sentence embeddings) towards one another within the latent space and moving negative pairs (dissimilar or unrelated examples) apart. In our instance, it assists the model in learning a more discriminative feature representation, especially useful when inappropriate content manifests in covert linguistic forms that might otherwise be misclassified.

Through an integration of contrastive learning and a pre-trained BERT backbone, the model takes advantage of both contextual richness and feature differentiation robustness. This dual strength substantially eliminates false positives (i.e., misclassified benign content as inappropriate) and false negatives (i.e., not-detected inappropriate content), both important issues in content moderation. The hybrid is hence most effective for our task and justifies the power of high-end representation learning techniques to achieve improved classification performance for sensitive media applications.

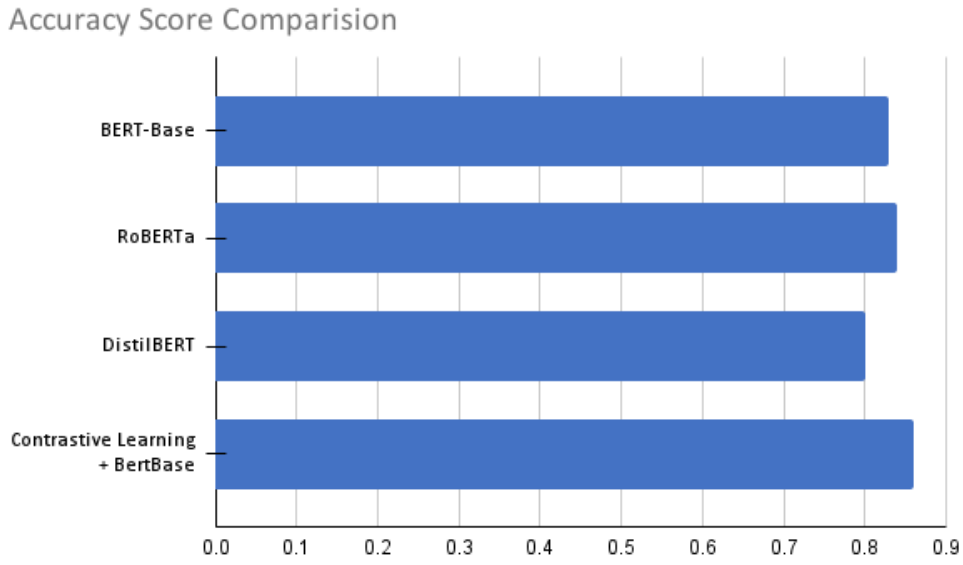

Figure 4: Accuracy Scores

4.2 Comparative Analysis between Visual Models

To measure the performance of trained models for visual detection on the cartoon content, we measured three metrics. Precision, Recall, F1 Score for every of the three models- **YOLO-v11**, **Faster R-CNN**, and **RF-DETR.**. The results are shown in the Table 2.

Table 2: Performance Metrics for Visual Detection Models

Model	**Precision**	**Recall**	**F1 Score**
YOLO-v11	0.83	0.91	0.84
Faster R-CNN	0.60	0.87	0.70
RF-DETR	0.89	0.86	0.86

YOLO-v11 was found to be the best-performing model when it comes to Recall where it obtained the best score (approximately 0.91), implying that **YOLO-v11** model is best suited for identifying most of the relevant objects in the dataset. Its Precision and F1 Score were also high (around 0.83 and 0.84, respectively) showing a good arrangement of trade off between finding true positives and reducing false alarms. The architecture of the YOLO-v11, a speed and end-to-end detection optimized neural network, is helpful in processing complex cartoon scenes with little latency thereof.

RF-DETR closely followed: Precision (0.89) was the highest one of all models. This means that **RF-DETR** had lower instances of false positive detections so it is particularly helpful when mislabeling safe content to be unsafe has to be avoided. Recall and F1 Score (both 0.86) were also high, representing the effectiveness of the model to retain true positive at a price of precision. The decoder of **RF-DETR**, which is based on transformers, arguably performs well due to the global context modeling.

Although Faster remained efficient, it was not quite as efficient as **YOLO-v11** and **RF-DETR**. It was the least Precision and lowest F1 Score (0.70), and although its Recall (0.87) was also competitive. This imbalance indicates that Faster **R-CNN** has a tendency towards over-predicting the objects, which means that it has a high false-positive rate. Its two-stage detection mechanism, while accurate in many traditional object detection tasks, may fail to generalize very well to cartoon data that are stylistically different without any more architectural adaptation or domain adaptation.

All in all, **YOLO-v11** is the best option for circumstances that require high recall (e.g., identifying potentially harmful content), **RF-DETR** is more suitable for conserative detection (minimizing false positives), while some adjustment of Faster R-CNN may be needed to achieve similar effectiveness.

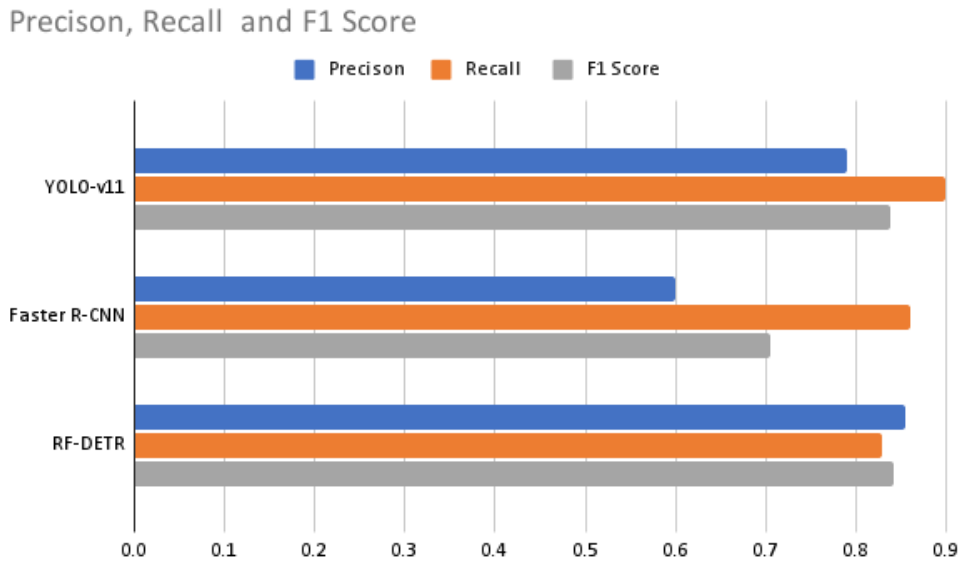

Figure 5: Evaluation Metrics for Visual Model

5 Conclusion

In this project, we proposed a novel multi-modal framework for cartoon video inappropriate content detection by combining approaches of NLP and Computer Vision. Through the use of BERT variants, we processed natural text in the format of transcripts of dialogues, as well as Faster R-CNN, RF-DETR, and YOLOv11 models to recognize sensitive visual content in cartoon images. Our system was designed with the aim of solving the drawback of the already available moderate tools, which are heavily reliant on the human user reporting, or naive classification models.

By a thorough comparative study, we found out that fine-tuning for BERT variants in particular and with the supplementation of contrastive learning resulted in better classification performance in identifying inappropriate text. Similarly, YOLOv11 had the highest rate and accuracy of all the visual models, and it is therefore suitable for real time or mass marketed applications. There is a significant increase in the trustworthiness of the content moderation process from the combination of both modalities in a layered architecture with the cross-verification of evidence from both image as well as textual information.

On balance, our minimum viable product (MVP) confirms that the use of AI-backed, context-aware content moderation software to make children's digital media experience safer and more appropriate is viable.

6 Future Work

Although our system's results are encouraging, there are a few paths to future improvements:

1. **Audio Modality Integration:** Adding raw audio analysis (e.g., tone, ambient noises) to transcript-based NLP can enhance detection credibility, particularly for off-screen or implied information.

2. **Temporal Consistency Modeling:** Incorporating video-based temporal models (e.g., Video Transformers or 3D CNNs) may improve detection accuracy by examining how visual and text context changes over time, instead of on a sentence-by-sentence or frame-by-frame basis.

3. **Bigger and More Varied Dataset:** Increasing the dataset to include a greater array of animation styles, languages, and cultural settings will enhance generalizability and minimize bias in detection.

4. **Real-time Deployment:** Streamlining the pipeline to process in real time—particularly on edge devices such as tablets or smartphones—can make the solution more accessible and more scalable.

By extending the system across these dimensions, this project lays the groundwork for more intelligent and responsible AI tools in the field of child-safe digital media.

References

[1] M. Y. Chuttur and A. Nazurally, "A multi-modal approach to detect inappropriate cartoon video contents using deep learning networks," *Multimedia Tools and Applications*, vol. 81, no. 12, pp. 16 881–16 900, May 2022.

[2] K. Papadamou *et al.*, "Disturbed youtube for kids: Characterizing and detecting inappropriate videos targeting young children," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, May 2020, pp. 522–533.

[3] L. Binh *et al.*, "Samba: Identifying inappropriate videos for young children on youtube," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)*. New York, NY, USA: Association for Computing Machinery, October 2022, pp. 88–97.

[4] S. H. Ahmed, M. J. Khan, H. M. U. Qaisar, and G. Sukthankar, "Malicious or benign? towards effective content moderation for children's videos," in *International FLAIRS Conference Proceedings*, vol. 36, May 2023.

[5] M. Gkolemi, P. Papadopoulos, E. P. Markatos, and N. Kourtellis, "Youtubers not madeforkids: Detecting channels sharing inappropriate videos targeting children," in *14th ACM Web Science Conference*, June 2022, pp. 370–381.

[6] K. Yousaf and T. Nawaz, "An attention mechanism-based cnn-bilstm classification model for detection of inappropriate content in cartoon videos," *Multimedia Tools and Applications*, vol. 83, pp. 31 317–31 340, 2024.

[7] V. Viswanatha, R. K. Chandana, and A. C. Ramachandra, "Real time object detection system with yolo and cnn models: A review," *Journal of Xi'an University of Architecture & Technology*, vol. XIV, 2022.