Deep Multi-Source Visual Fusion with Transformer Model for Video Content Filtering

Senthil Murugan Nagarajan, *Senior Member, IEEE,* Ganesh Gopal Devarajan *Senior Member, IEEE,* Asha Jerlin M, Daniel Arockiam, Ali Kashif Bashir, *Senior Member, IEEE,* Maryam M. Al Dabel

***Abstract*—As YouTube content continues to grow, advanced filtering systems are crucial to ensuring a safe and enjoyable user experience. We present MFusTSVD, a multi-modal model for classifying YouTube video content by analyzing text, audio, and video images. MFusTSVD uses specialized methods to extract features from audio and video images, while processing text data with BERT Transformers. Our key innovation includes two new BERT-based multi-modal fusion methods: B-SMTLMF and B-CMTLRMF. These methods combine features from different data types and improve the model's ability to understand each type of data, including detailed audio patterns, leading to better content classification and speech-related separation. MFusTSVD is designed to perform better than existing models in terms of accuracy, precision, recall, and F-measure. Tests show that MFusTSVD consistently outperforms popular models like Memory Fusion Network, Early Fusion LSTM, Late Fusion LSTM, and multi-modal Transformer across different content types and evaluation measures. In particular, MFusTSVD effectively balances precision and recall, which makes it especially useful for identifying inappropriate speech and audio content, as well as broader categories, ensuring reliable and robust content moderation.**

***Index Terms*—multi-modal Fusion, Transformer Model, Deep Learning, Content Filtering, Speech Enhancement**

I. Introduction

In recent years, the multimedia industry becomes the top service for day-to-day life and its volatile growth occurred when the latest developments emerged for super-fast machines and digital devices. This impact leads to an increase in large volume of multi-modal and sensor data generation. Different types of data are included in this collection of information, such as unstructured text documents, images, audios, videos, and networking statistics. The main content of multi-media-based data generation is occupied by short and long videos that make systems face complex problems toward users [1], [2]. Different types of human emotions and feelings can be triggered by various types of video clips [3], [4]. As video clips can evoke a wide range of human emotions and reactions, effective content filtering and classification have become critical. The integration of sophisticated algorithms and machine learning techniques is essential to address the complexities associated with analyzing such diverse data. This need for advanced solutions highlights the importance of developing robust systems capable of processing and interpreting multi-modal data efficiently. In summary, the growth trajectory of the multimedia industry underscores the need for innovative approaches to manage the vast amounts of data generated daily, particularly video content. These advances not only aim to improve user engagement, but also ensure a safer online environment by effectively moderating potentially harmful content [5]–[7].

Various developments have been introduced to filter the contents of audio, video, and text data. However, different disadvantages such as lack of use experience and different understanding of categories which are often inconsistent make researchers to find new solutions. Despite the various methods proposed for the analysis of information and its extraction, there is still time to pay attention to different modalities [8], [9]. Several researches had analyzed the video categorization based on different genres into various semantic concepts. Due to the high diversity of information or data related to subject, genre, format, quality, and style, this leads to a variety of problems during the process and analysis. The information obtained is from multiple channels and different sources, where it contains not only audio and video, but also the speech and text that were applied during the analysis. SO, representing such information is always a difficult problem and leads to improvements to existing models [10], [11].

Due to the increase in information related to digital commercials, it became the common reason for the expansion of larger repositories. Several advances have been made using neural network models and recent methods, such as BERT and transformer techniques, have been used to increase the performance of multimodality-based fusion models [12]. For this reason, developing fusion-based methods has become prominent for understanding the performance of deep learning models with improving the analysis over different feature extraction techniques. In this paper, we discuss different processes that integrate feature extraction for analyzing not

Senthil Murugan Nagarajan is with Department of Mathematics, Faculty of Science, Technology, and Medicine, University of Luxembourg, Belval Campus, Luxembourg (e-mail: senthil.nagarajan@uni.lu)

(*Corresponding Author) Ganesh Gopal Devarajan is with Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Delhi - NCR Campus, Delhi - Meerut Road, Modinagar, Ghaziabad, Uttar Pradesh - 201204, India (e-mail: dganeshgopal@ieee.org)

Asha Jerlin M is with School of Computer Science and Engineering, Vellore Institute of Technology, Chennai, Kelambakkam, Rajan Nagar, Chennai 600127, India (e-mail: ashajerlin.m@vit.ac.in)

Daniel Arockiam is with CSE-ASET, Amity University, Madhya Pradesh, India(e-mail:danielarockiam@gmail.com)

Ali Kashif Bashir is with Department of Computing and Mathematics, Manchester Metropolitan University, UK, and with Woxsen School of Business, Woxsen University, India, and with Department of Computer Science and Mathematics, Lebanese American University, Beirut, Lebanon (email:dr.alikashif.b@ieee.org)

Maryam M. Al Dabel is with Department of Computer Science and Engineering, College of Computer Science and Engineering, University of Hafr Al Batin, Saudi Arabia (e-mail:maldabel@uhb.edu.sa)

only video and audio but also textual features. Furthermore, multi-modal fusion methods are applied with the transformer technique to fuse the features and analyze the performance of different models.

The main contributions are summarized as follows.

- MFusTSVD model for content classification in YouTube videos based on text, audio, and video.
- Developed a feature extraction method using an advanced hand-crafted technique to extract important information from the data.
- Introduced two innovative BERT-powered fusion schemes B-SMTLMF and B-CMTLRMF that effectively combine features from different modalities while improving the model's capacity to learn self-representations, particularly in audio data.
- Significant improvements in speech-related content separation and classification capabilities, allowing for more accurate detection of inappropriate speech and audio content.
- Achievement of a well-balanced trade-off between precision and recall, ensuring reliability and robustness in content moderation tasks.

In Section 2, a detailed literature of various research findings is discussed. Section 3 presents the methodology and the system model. In addition, in detail, the feature extraction methods for audio data. Section 4 discussed the classification of video data. In Section 5, the results and discussions based on the benchmark dataset are detailed and show different performance metrics. Finally, the conclusion with some future scope of this research is discussed in Section 6.

II. Related Works

Several studies have explored the recommendation of media advertisements and content filtering based on social relevance [13], [14]. The design and production of multimedia content play a crucial role in feature extraction analysis, as high-level patterns in temporal and multimedia features often mimic human cognition. This requires a deep understanding of both applications and content creation [15], [16]. Kuleshov et al. [17] analyzed the impact of filtering during pre-processing, focusing primarily on band-limited input data preparation for audio super-resolution before down-sampling.

Cambria et al. [18] introduced Sentic-Blending, which combines modalities for emotion-based content. Their model combined natural language text and facial expressions to track sentiments over time, using the MMI and FGNET datasets. Paleari et al. [19] applied feature- and decision-level fusion methods, utilizing the eNTERFACE dataset to combine multi-modal information. Paradarami et al. [20] developed a model using content characteristics and reviews through a deep neural network architecture to generate predictive performance using an aggregated function approach.

Chung et al. [21] proposed a multi-modal collaborative recommendation technique using the attention mechanism to represent image features in high order. They used an LSTM model for feature fusion to capture user preferences. Wei et al. [22] applied a cross-attention model to fuse image and text modalities for downstream tasks. Wang et al. [**?**] developed a model based on multi-modal transformer token fusion to detect uninformative tokens and aggregate features.

Zhu et al. [23] introduced a dual-branch attention fusion deep network to classify multi-resolution data. Their end-to-end network model integrated feature fusion for classification. Zhao et al. [24] proposed a collaborative attention network with dual spatial branches to improve the accuracy of classification by improving features and samples. Liu et al. [25] presented GAFNet, a deep group spatial attention fusion network model designed to extract high- and low-level features while preserving local details and global abstraction.

Nina et al. [26] introduced a multi-modal, modality-agnostic fusion transformer that learns to exchange information between multiple modalities, such as video, audio and text, integrating them into a unified multi-modal representation. The model is trained with a combinatorial loss on various combinations of modality, enabling it to process and fuse any number of input modalities at test time. Swalpa Kumar Roy et al. [27] proposed a multimodal fusion transformer network comprising a multi-head cross-patch attention mechanism for hyperspectral image classification. The model utilizes complementary information from different modalities to achieve better generalization, learning distinctive representations in a reduced and hierarchical feature space.

III. Methodology

This section presents a multi-modal Fusion Transformer Safe Video Detection (MFusTSVD) model that performs Content Filtering and inappropriate content detection in YouTube videos. The architecture of the multi-modal Fusion Transformer Safe Video Detection (MFusTSVD) model is shown in Fig. 1.

At the initial stage, video pre-processing is processed and subdivided into video, audio, and text. The audio transcript data from the video is extracted and converted into text data. Next, video pre-processing involves several key steps to enhance the quality and uniformity of the information present in that data. Frame extraction is performed to break down the video input $Vd_N = \{Vd_1, Vd_2, ..., Vd_i | i = i, 1, 2, ..., N\}$ into small fixed-sized video clips $vc_M^N = \{vc_1^i, vc_2^i, ..., vc_j^i | i = 1, 2, ..., N; j = 1, 2, ..., M\}$ each lasting S seconds. Let the total duration of the video Vd_i be τ seconds, then divide it into small fixed video clips using Eqn. 1:

$$Number\ of\ clip(M) = \frac{\tau}{C} \tag{1}$$

Here, C represents the desired clip length in seconds and τ represents the total duration of the video in seconds.

Each clip is sampled into frames taking $\frac{1}{4}^{th}$ of the average frame rate (AFR), which can be calculated using Eqn. 2.

$$No.\ of\ Frames(k) = \frac{1}{4} * AFR \tag{2}$$

The video's average frame rate (AFR) was 23 to 24 frames per second (FPS). So, we sampled each clip into frames as 6 FPS, which is approximately $\frac{1}{4}^{th}$ of the AFR of the video (23-24 FPS). Therefore, each clip vc_j^i is then sampled at the rate

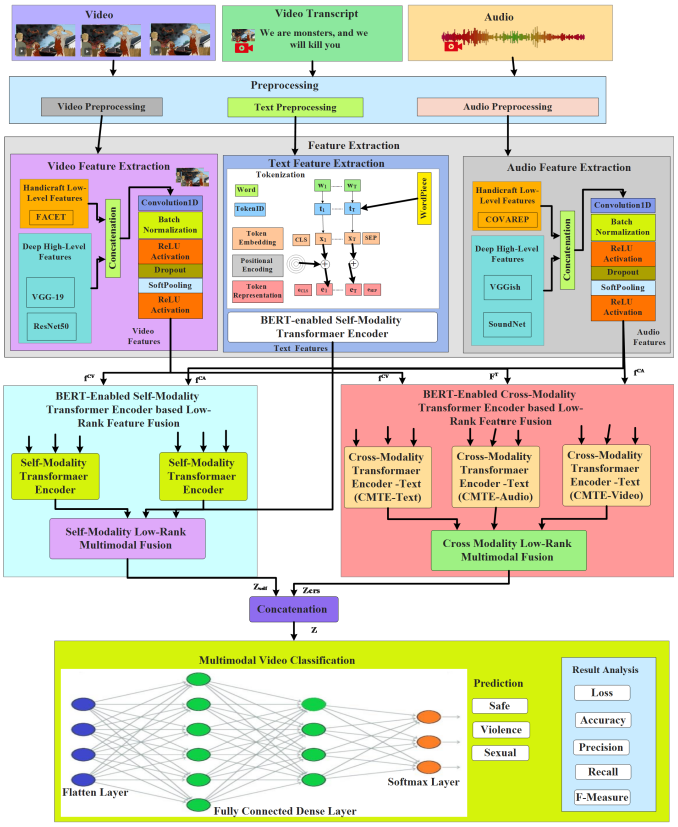


Fig. 1: Architecture of MFusTSVD Model

of 6 frames per second (FPS) and is represented as $vf_k^{i,j} = \{vf_1^{i,j}, vf_2^{i,j}, ..., vf_5^{i,j}\}$, which means that the k^{th} video frame of video clip vc_j^i belongs to the video input Vd_i. Finally, the frame is resized by resizing each frame as 224×224. So, each clip is resized into the frame as 224×224×3 with 6∗S frames.

A. Necessity of BERT Models in MFusTSVD

The inclusion of BERT-powered fusion schemes, BSMTLMF and B-CMTLRMF, in MFusTSVD serves distinct but complementary purposes. The architecture of BERT, based on self-attention mechanisms, is highly effective in capturing contextual relationships and nuanced feature interactions within and between modalities. The reasons why both schemes are necessary:

1) **B-CMTLRMF:** This scheme addresses the interdependence between modalities, such as how audio complements visual data or how text enhances understanding of the other two modalities. By capturing cross-modal interactions, B-CMTLRMF provides a holistic representation that improves overall classification performance.
2) The combined use of these schemes allows the model to leverage both intra-modal depth and intermodal synergies, resulting in significant performance gains across metrics like accuracy, precision, recall, and F1 score.
3) **Computational Cost and Suitability for Real-Time Systems:** We acknowledge that the integration of BERT models increases computational cost due to their complexity. However, several considerations make their inclusion justified and manageable:
4) **Performance Justification:** The improved accuracy and robustness provided by BERT models outweigh the additional cost, particularly in applications where high accuracy is critical, such as content moderation and sensitive material classification.
5) **Optimization Opportunities:** Techniques like model pruning, quantization, and knowledge distillation can reduce the computational overhead of BERT without substantial performance loss, making the system suitable for applications in near-real time. Additionally, hardware accelerators (e.g., GPUs, TPUs) can further optimize inference time.
6) **Scalability:** By modularizing BSMTLMF and B-CMTLRMF, the system can be scaled or adapted to balance performance with computational constraints based on the deployment scenario.
7) Using only one scheme would result in a loss of either intra-modal (BSMTLMF) or intermodal (B-CMTLRMF) insights: Employing only BSMTLMF would limit the model's ability to capture cross-modal dependencies, which are essential for understanding multimodal data holistically. Using only B-CMTLRMF without the contextual enhancement provided by BSMTLMF would reduce the quality of individual modality features, weakening overall fusion quality.

B. Feature Extraction Process

For the task of classifying videos, three data modalities are incorporated: textual, visual, and auditory. Each modality employs a distinct technique for feature extraction, resulting in varied semantic information. Next, we introduce the feature extraction method for each modality.

Textual Features: We employ BERT (Bidirectional Encoder Representations of Transformers) to extract textual features from video transcripts. Tokenization is performed using WordPiece, which segments words into subword tokens $T = \{t_1, ..., t_T\}$ and assigns them token IDs based on a predefined vocabulary. Then these token IDs are assigned to the token embeddings $X = \{x_1, ..., x_T\}$. Since transformers lack inherent word order understanding, Positional Encoding (PEN) is added to embeddings, computed as per Eqn. 3 and 4.

$$PEN(posi, 2i) = sin(\frac{posi}{10000^{\frac{2i}{d}}}) \tag{3}$$

$$PEN(posi, 2i+1) = cos(\frac{posi}{10000^{\frac{2i+1}{d}}}) \tag{4}$$

The final token embedding includes [CLS] at the start and [SEP] at the end, forming the sequence T_E, which is input into BERT's encoder to generate contextualized word embeddings as per Eqn. 5

$$F_C^T = BERT(T_E) \tag{5}$$

Each encoding layer consists of a multihead self-attention sub-layer and a fully connected sub-layer, ensuring effective contextual representation.

Audio Feature Extraction: To maximize precision, we considered two types of feature: Handicraft low-level audio (HLLA) features and deep learning-based high-level audio (DHLA) features.

Handicraft Low-Level Audio Features: We use the COVAREP framework to extract HLLA characteristics f^{HA} such as 1. Pitch & Peak Slope Parameters where it helps to differentiate speech tones, detect emphasis, and identify speaker variations. 2. Mel frequency cepstral coefficients (MFCCs) that capture timbral and phonetic characteristics, making them useful for speech recognition and emotion detection. 3. Maximum proportions of dispersion that provides information on speech rhythm and fluency. 4. Voiced/Unvoiced Segmentation Features, which helps distinguish between speech and non-speech sounds, aiding in content classification. 5. Glottal Source Parameters that analyze voice quality, which can help detect emotional tone and speaker characteristics. These features play a crucial role in our MFusTSVD model by improving its ability to classify speech-related content, detect inappropriate speech patterns, and improve overall audio-based filtering.

Deep Audio Feature: We used the SoundNet and VGGish deep learning model to obtain the DHLA feature maps $f^{DA} = \{f^{sound} \parallel f^{Vggish}\}; f^{vggish} \in R^{1024}, f^{sound} \in R^{128}$. The size of a 3-dimensional feature map is represented as $H \times W \times C$, where H represents the height, w represents the width, and C represents the number of channels of the feature map. In order to finished DHLA features f^{CA}, we concatenate f^{HA} and f^{DA} into feature dimension using Eqn. 6:

$$f^{CA} = f^{HA} \parallel f^{DA} \tag{6}$$

Like audio, we considered two types of visual modalities, including HLLV and Deep Learning-based High-Level Video (DHLV) features. For visual modality, Facet5 extracts a set of low-level visual (HLLV) characteristics of handcrafts f^{HV} including facial action units, facial landmarks, head pose, gaze tracking and HOG characteristics. We applied two deep learning models including VGG-19 and RestNet50 to extract (DHLV) characteristics $f^{DV} = \{f^{Vgg} \parallel f^{ResNet}\}; f^{Vgg} \in R^{4096}, f^{R}esNet \in R^{2048}$, that is. In order to finished DHLV features f^{CV}, we concatenate f^{HV} and f^{DV} into the feature dimension using the following mathematical expression given in Eqn. 7:

$$f^{CV} = f^{HV} \parallel f^{DV} \tag{7}$$

Since the audio and video dimensionality is different from the dimensionality of the text feature set, we employed a convolutional one-dimensional operation (Conv1-D) to achieve the same dimension and enables strong connection with their adjacent element. To compute unimodal input from the audio and video feature, we performed the Conv1-D operation followed by batch normalization, activation ReLU, dropout, soft pooling, and activation ReLU, and mathematically it is expressed in Eqn. 8 and 9.

$$F^{CA} = ReLU\Big(softpooling\big(Drop(ReLU(BN(conv1D(f^{CA}, kr^{CA}))))\big)\Big) \tag{8}$$

$$F^{CV} = ReLU\Big(softpooling\big(Drop(ReLU(BN(conv1D(f^{CV}, kr^{CV}))))\big)\Big) \tag{9}$$

Here, kr^{CA} and kr^{CV} represent the size of the convolution kernel for the audio and video modalities.

Considering the interconnected nature of elements (time steps) within uni-modal sequences and their proximity to neighboring elements, a pooling technique can compress the temporal dimension. Consequently, we employ soft pooling to aggregate the time dimension of individual uni-modal sequences. Soft pooling effectively reduces the length of the uni-modal sequence, thereby refining the information encapsulated within it. Soft-pooling is computed using Eqn. 10.

$$softpooling(f_i^m) = \sum_{i \in R} \frac{e^{f_i^m}}{\sum_{j \in R} e^{f_i^m}} f_i^m \quad ; |m = CA, CV \tag{10}$$

Here, f_i^m refers to the i^{th} element along the temporal axis of the uni-modal sequence after dropout. R denotes the pooling region. In contrast to max-pooling and mean-pooling, soft-pooling allocates non-linear weights to the elements within R, thereby exhibiting greater expressive capability.

C. BERT-enabled Self Modality Transformer based Low-Rank multi-modal Fusion (B-SMTLMF)

The proposed B-SMTLMF model employed the concept of the BERT mechanism to evaluate the importance of correlations between feature characteristics within a singular modality and capture the contextual characteristics of the set of features. Contextual characteristic of the audio and video feature set using the Self Modality Transformer Encoder (SMTE).

The self-Modality Transformer Encoder (SMTE) employs the multi-head self-attention mechanism used in BERT to calculate the self-attention score for multi-headed sub-layer, first compute the key (K), query (Q), and value (V) vectors for multi-modal feature vectors F^m using linear transformation given in Eqn. 11.

$$Q = F^m W_Q^h; K = F^m W_K^h; V = F^m W_V^h \tag{11}$$

Where, m represents modalities including the audio feature set and the video feature set such that m = CA,CV. Then, the attention score $A_i j$ is computed using Eqn. 12.

$$A_{ij}^h = softmax(\frac{QK^T}{\sqrt{d_a}}) \tag{12}$$

Here, $d_a = d_{hs}/h$; d_{hs} represents the hidden state dimension, while h denotes the number of the attention head and $W_K^h \in R^{d_a \times d_{hs}}$, $W_Q^h \in R^{d_a \times d_{hs}}$, $W_V^h \in R^{d_a \times d_{hs}}$ represent weight matrices. Since the self-attention mechanism is performed for a series of hidden states $Hs = \{hs_1, ..., hs_F\}$. Here, hidden state is represented by multi-modal feature vectors F^m. Therefore, the attention score for the hidden state is computed using Eqn. 13.

$$A_{ij}^h = softmax\left(\frac{(W_Q^h hs_i)(W_K^h hs_i)}{\sqrt{d_a}}\right) \tag{13}$$

Based on the attention score, the output of the multi headed self-attention sub-layer $M^H = \{hd_1, ..., hd_H\}$ is given in Eqn. 14.

$$hd_i^h = \sum_{j=1}^{F} A_{ij}^h (W_V^h hs_i) \tag{14}$$

Concatenate the outputs of all heads as $CM^H = [hd_i^1 \parallel hd_i^2 ... \parallel hd_i^h]$ and apply a linear transformation using Eqn. 15.

$$F_{Linear}^m = Linear(CM^H) = W_{mh} M^H + b \tag{15}$$

Where, $W_{mh} \in R^{d_{hs} \times d_{hs}}$ represents the weight matrix, $\parallel$ denotes the concatenation operation. After residual connection, add the original input to the output and apply layer normalization using Eqn. 16.

$$F_{LN}^m = AddNorm(Hs + F_{Linear}^m) = AddNorm(T_E + F_{Linear}^m) \tag{16}$$

Apply a fully connected feed forward network, in position, providing $(\hat{MH})$ as input to the feed forward network that generates the output as $O^{FFN} = \{o_1^{FFN}, ..., o_N^{FFN}\}$ and is expressed in Eqn. 17.

$$F_{FFN}^m = W_2 ReLU(W_1 F_{LN}^m + B_1) + B_2 \tag{17}$$

Where, $W_1, W_2 \in R^{d_{hs} \times d_{hs}}$ represent weight matrices and $B_1, B_2 \in R^{d_{hs} \times d_{hs}}$. Add the output of the input network to the previous output and apply layer normalization to generate the final output using Eqn. 18.

$$F_C^m = AddNorm(F_{Linear}^m + F_{FFN}^m) \tag{18}$$

The final output F_C^m obtained from the last layer is considered a contextual feature embedding matrix for a given audio and video input sequence. We employed the entire stack of H Layers of self-attention multi-headed operation of BERT architecture for estimating feature correlation using the self-attention multi-headed mechanism.

D. Self-Modality Low-Rank Multi-modal Fusion

In this work, we used Low-Rank multi-modal fusion (LRMF) to fuse the characteristics of different modalities, including video, text, and audio, to predict safe video. LRMF is a multi-modal based tensor fusion that correlates the features from different modalities. In this work, we used the concept LRMF to perform multi-modal feature fusion of three different feature vectors $F_C^s = \{F_C^T, F_C^{CA}, F_C^{CV}\}$ extracted from three different modalities including text, video and audio. We consider the low rank decomposition method in LRMF to break down the weight W into distinct factors $\{W^1, ..., W^{d_l}\}$ that align with the modal features z_s and it is expressed mathematically using Eqn. 19.

$$W = \{W^1, ..., W^{d_l}\} \tag{19}$$

Where, d_l represents the dimension of the set of text, audio, and video features, that is, $d_l = d_T \times d_{CA} \times d_{CV}$. W is obtained by applying low-rank decomposition based on the LRMF technique and is expressed in Eqn. 20.

$$W^{self} = \sum_{i=1}^{R} \bigotimes_{s=1}^{S} (\omega)_i^s \tag{20}$$

Here, R represents the rank value of the matrix. We consider $R^{d_s \times t_s}$ as s^{th} modality space which includes the number of modalities S and a random selection of a one-time step is made from the features of each modality denoted $F_C^s \in R^{d_s}$. The input tensor z, formed by the uni-modal representation is calculated using Eqn. 21.

$$Z^{self} = \bigotimes_{s=1}^{S} F_C^s \tag{21}$$

LRMF-based multi-modal feature fusion Z_{fusion}^{self} is obtained using Eqn. 22.

$$Z_{fusion}^{self} = g(Z^{self}; W^{self}, B^{self}) = W^{self}.Z^{self} + B^{self} \tag{22}$$

Here, B^{self} represents bias.

$$\begin{aligned} W^{self}.Z^{self} &= \left(\sum_{R}^{i=1} \bigotimes_{s=1}^{S} (\omega^{sf})_i^s \right) (\bigotimes_{s=1}^{S} F_C^s) \\ &= \sum \left(\sum_{i=1}^{R} \bigotimes_{s=1}^{S} (\omega_{sf})_i^s . \bigotimes_{s=1}^{S} F_C^s \right) \\ &= \sum_{R}^{i=1} \left(\sum \bigotimes_{s=1}^{S} ((\omega^{sf})_i^s . F_C^s) \right) \end{aligned} \tag{23}$$

Here, $\sum$ and. represent the summation of the element in the multiplication of elements in a given direction and can be replaced by $\bigwedge_{s=1}^{S} x_s = x_1.x_2.x_3$ that performs the same summation of the element with the multiplication of elements in a given direction or the input tensor sequence.

$$\begin{aligned} &= \sum_{i=1}^{R} \left(\bigwedge_{s=1}^{S} (\omega_{sf})_i^s . F_C^s \right) = \bigwedge_{s=1}^{S} \left(\sum_{i=1}^{R} (\omega^{sf})_i^s . F_C^s \right) \\ Z_{fusion}^{self} &= \left[\left(\sum_{R}^{i=1} (\omega^{sf})_i^T . F_C^T \right) \circ \left(\sum_{R}^{i=1} (\omega^{sf})_i^{CA} . F_C^{CA} \right) \circ \left(\sum_{R}^{i=1} (\omega^{sf})_i^{CV} . F_C^{CV} \right) \right] + B^{self} \end{aligned} \tag{24}$$

Algorithm 1 represents the BERT-enabled Self-Modality Transformer-based Low-Rank multi-modal Fusion (B-SMTLMF).

E. Cross Modality Model

In this part, we introduce BERT-enabled cross-modal transformer low-rank multi-modal fusion (B-CMTLRMF) for cross-modal feature fusion and their interaction to capture intra-modal complementary information. We employed three Cross-Modality Transformers (CMT), including CMT-Text, CMT-Audio, and CMT-video, that used self-attention multi-headed mechanisms to perform feature fusion from three different modalities, including Text, Video, and Audio.

In CMT-Text, the transformer encoder, we performed the cross-modal feature fusion of the textual feature vectors F_C^T with the audio and video feature matrix f^{CA} and f^{CV}, respectively. First, we compute the attention score for text-audio cross-modality and text-video cross-modality. The text-audio

Algorithm 1 B-SMTLMF Algorithm

Input: Token Embedding $\leftarrow T_E$, Multi-modal Feature $F^m = \{F^{CA}, F^{CA}\}, weight \leftarrow W^{self}$

Output: Self Modality Feature Fusion Z^{self}_{fusion}

Encoding Model
$F^T_C = BERT(T_E)$
$F^m_C = SMTE(F^m)$
Procedure (F^m)
Initialize
$Q = F^m W^h_Q; K = F^m W^h_K; V = F^m W^h_V$
for each feature in F^m **do**
  for each head in hd^h_i **do**
    Compute attention score as per Eqn. 15 and Eqn. 16
  end for
end for
Concatenate heads $CM^H = [hd^1_i \parallel hd^2_i \parallel .... \parallel hd^h_i]$
Apply Linear Transformation as per Eqn. 17
Apply Add and Normalization as per Eqn. 18
Apply FFN $F^m_{FFN} = W_2 ReLU(W_1 F^m_{LN} + B_1) + B2$
Apply Add and Normalization as per Eqn. 20
End Procedure
for $i = 1 toR$ **do**
  for 1 ot s **do**
    $W^{self} = \sum_{i=1}^{R} \bigotimes_{s=1}^{S} (\omega)^s_i$
    $Z^{self}_{fusion} = \left[\left(\sum_R^{i=1} (\omega^{sf})^T_i . F^T_C\right) \circ \left(\sum_R^{i=1} (\omega^{sf})^{CA}_i . F^{CA}_C\right) \circ \left(\sum_R^{i=1} (\omega^{sf})^{CV}_i . F^{CV}_C\right)\right] + B^{self}$
  end for
end for
Return Z^{self}_{fusion}

cross-modality attention score A^{T-CA} is computed using the audio feature matrix F^{CA} as query Q. In contrast, textual feature vectors F^T_C as key K and Value V. Similarly, the Text-Video Cross-Modality Attention Score A^{T-CV} is computed by taking the audio feature matrix F^{CV} as query Q while textual feature vectors F^T_C as key K and Value V. After k^{th} cross-modality multi-head attention layer, we obtained text-audio cross-modality MH^{T-CA^k} and text-video cross-modality MH^{T-CV^k} using Eqn. 25 and 26.

$$MH^{T-CA^k} = CrossM^H(F^{CA^k}, F^{T^k}_C, F^{T^k}_C) = w^k_{cma}[hd^{T-CA^k}_0 \parallel ...hd^{T-CA^k}_p] \quad (25)$$

$$MH^{T-CV^k} = CrossM^H(F^{CV^k}, F^{T^k}_C, F^{T^k}_C) = w^k_{cma}[hd^{T-CV^k}_0 \parallel ...hd^{T-CV^k}_p] \quad (26)$$

where, w^k_{cma} represents the weight matrix; $hd^{T-CA^k}_i = A^{T-CA}(w^k_{Q_i}. F^{CA^k}, w^k_{K_i} F^{T^k}_C, w^k_{V_i} F^{T^k}_C$ and $hd^{T-CV^k}_i = A^{T-CV}(w^k_{Q_i}.F^{CV^k}, w^k_{K_i} F^{T^k}_C, w^k_{V_i} F^{T^k}_C).A^{T-CA}$. (.) represent single attention layer obtained using Eqn. 9 while $hd^{T-CA^k}_i$ represents i^{th} single head attention layer output which is computed using Eqn. 10. Add F^{T-CA^k} and F^{T-CV^k} to generate fused cross-modality FCM^{T,CA,CV^k} using Eqn. 27.

$$FCM^{T,CA,CV^k} = Drop(MHT-CA^k + MHT-CV^k) \quad (27)$$

Here, drop(.) represents the dropout layer. The resultant output FCM^{T,CA,CV^k} was then fed into the residual and normalization layer to generate a set of integrated features based on textual attention T^{CA,CV^k} and mathematically expressed in Eqn. 28.

$$T^{CA,CV^k} = Norm(F^T_C + FCM^{T,CA,CV^k}) \quad (28)$$

In order to obtain the final fused textual feature matrix F^{crs-T} fused with the audio and video feature matrix, the resultant output T^{CA,CV^k} is fed into the feed-forward, residual, and normalization layer, which performs the following operations based on Eqn. 29 and 30.

$$FFN(T^{CA,CV^k}) = Drop(w^k_1(Drop(ReLU(w^k_0.T^{CA,CV^k} + B^k_0))) + B^k_1) \quad (29)$$

$$F^{crs-T} = Norm\left(T^{CA,CV^k} + FFN(T^{CA,CV^k})\right) \quad (30)$$

Where, w^k_0, w^k_1 and B^k_0, B^k_1 represent the weight and bias matrix; Drop(.),FFN(.),Norm(.) and ReLU (.) represent dropout, feedforward, normalization, and the ReLU activation function. Similarly, we obtained the fused audio feature matrix F^{crs-A} fused with the text and video feature matrix, and the fused video feature matrix F^{crs-V} fused with the audio and text feature matrix using the CMT-Audio and CMT-Video encoder. The resultant output F^{crs-A} had been computed by passing the feature matrix F^CA, F^CV and F^T_C to the CMT-Audio transformer encoder and mathematically it is expressed in Eqn. 31 and 32.

$$A^{CA,CV^k} = Norm\left(F^CA + FCM^{CA,T,CV^k}\right) \quad (31)$$

$$F^{crs-A} = Norm\left(A^{CA,CV^k} + FFN(A^{CA,CV^k})\right) \quad (32)$$

The resultant output $F^{cross-V}$ had been calculated by passing the feature matrix f^{CA}, f^{CV} and F^T_C to the CMT-Video transformer encoder and is mathematically expressed in Enq. 33 and 34.

$$V^{CA,CV^k} = Norm\left(f^CV + FCM^{CV,CA,T^k}\right) \quad (33)$$

$$F^{crs-V} = Norm\left(V^{CA,CV^k} + FFN(V^{CA,CV^k})\right) \quad (34)$$

Algorithm 2 represents pseudo-code for BERT-enabled cross-modality transformer-based low-rank multi-modal fusion (B-CMTLMF).

Algorithm 2 Cross Modality Algorithm

Input: Text feature$\leftarrow F_C^T$, Audio Features $\leftarrow F^{CA}$, Audio Features $\leftarrow F^{CV}$, weight $\leftarrow W^{crs}$
Output: Cross Modality Feature Fusion Z_{crs}^{fusion}

Encoding Model
$F_C^T = BERT(T_E)$
$F^{crs-T} = CMTText(F_C^T, F^{CA}, F^{CV})$
$F^{crs-A} = CMTAudio(F^{CA}, F_C^T, F^{CV})$
$F^{crs-V} = CMTVideo(F^{CV}, F^{CA}, F_C^T)$
$ProcedureCMTText(F_C^T, F^{CA}, F^{CV})$
Initialize
$QA = F^{CA}W_{QA}^h; KT = F_C^T W_{KT}^h; VT = F_C^T W_{VT}^h$
$QV = F^{CV}W_{QV}^h; KT = F_C^T W_K T^h; VT = F_C^T W_{VT}^h$
Compute Attention Score
$A^{T-CA}\left(w_{QA_i}^k F^{CA^k}, w_{KT_i.}^k F_C^{T^k}, w_{VT_i.}^k F_C^{T^k}\right)$ and
$A^{T-CV}(w_{QV_i}^k F^{CV^k}, w_{KT_i.}^k F_C^{T^k}, w_{VT_i.}^k F_C^{T^k})$
Compute HEAD Operation
$hd_i^{T-CA^k} = A^{T-CA}\left(w_{QA_i}^k . F^{CA^k}, w_{K_i.}^k F_C^{T^k}, w_{VT_i.}^k F_C^{T^k}\right)$
$hd_i^{T-CV^k} = A^{T-CA}\left(w_{V_i}^k . F^{CA^k}, w_{KT_i.}^k F_C^{T^k}, w_{VT_i.}^k F_C^{T^k}\right)$
Compute cross modality Multi-head Operation
$MH^{T-CA^k} = w_{cma}^k[hd_0^{T-CA^k} \parallel ...hd_p^{T-CA^k}]$
$MH^{T-CV^k} = w_{cma}^k[hd_0^{T-CV^k} \parallel ...hd_p^{T-CV^k}]$
Generate fused cross-modality
$FCM^{T,CA,CV^k} = Drop(MH^{T-CA^k} + MH^{T-CV^k})$
Apply Add and Normalization
$T^{CA,CV^k} = Norm(F_C^T + FCM^{T,CA,CV^k})$
Apply FFN, Add and Normalization
$FFN(T^{CA,CV^k}) = Drop(w_1^k(Drop(ReLU(w_0^k.T^{CA,CV^k} + B_0^k))) + B_1^k)$
$F^{crs-T} = Norm(T^{CA,CV^k} + FFN(T^{CA,CV^k}))$
Return F^{crs-T}
End Procedure
Completion of CMTAudio and CMTVideo Fusion
for i = 1 to R **do**
  for s = 1 to S **do**
    $Z_{fusion}^{crs} = [(\sum_{i=1}^{R}(\omega^{cr})_i^T . F^{crs-T}) \circ (\sum_{i=1}^{R}(\omega^{cr})_i^A . F^{crs-A}) \circ (\sum_{i=1}^{R}(\omega^{cr})_i^V . F^{crs-V})] + B^{crs}$
  end for
end for
Return Z_{fusion}^{crs}

IV. Multi-modal fusion based Video classification and Los

s LRMF enables self and cross-modality based multi-modal feature fusion set Z_{fusion}^{self} and Z_{fusion}^{crs} had been concatenated to obtained final feature matrix Z_{fusion}^{cat}. The operation of concatenation is mathematically expressed in Eqn. 35.

$$Z_{fusion}^{cat} = W_f[Z_{fusion}^{self} \parallel Z_{fusion}^{crs}] \tag{35}$$

Where, W_f represents the learnable weight parameter and $\parallel$ represents the concatenation parameter. The concatenated output Z_{fusion}^{cat} is then fed as input to the dense layer and the SoftMax classification layer to classify the video into safe, violence, and sexual. The video classification is formulated using Eqn. 36, 37, and 38:

$$g_i = ReLU(W_{cl}.Z_{fusion_i}^{cat} + B_{cl}) \tag{36}$$

$$pd_i = softmax(W'_{cl}.g_i + B'_{cl}) \tag{37}$$

$$y'_i = \arg\max_q(pd_i) \tag{38}$$

Where, g_i represents the dense layer output, pd_i represents the prediction probability of the given input video vd_i, W_{cl}, W'_{cl} and B_{cl}, B'_{cl} represent the learnable weight and bias parameter; y'_i represents the predicted class. To compute loss, we used the L2-Norm-based cross-entropy function and trained the proposed model with minimum loss using Eqn. 39.

$$Loss = \frac{1}{\sum_{k=1}^{N} n(k)} \sum_{i=1}^{N} \sum_{j=1}^{n(i)} y_{ij} log y'_{ij} \tag{39}$$

Where, N represents the number of videos, n(i) represents the number of video clips in the i^{th} video, y_{ij} and y'_{ij} represent the actual and predicted class of video clips j^{th} in the i^{th} video. Algorithm 3 represented pseudo-code for multi-modal fusion-based Video classification and Loss.

Algorithm 3 multi-modal fusion-based Video classification and Loss

Input: Self modality fused features Z_{fusion}^{self}, Cross modality fused features Z_{fusion}^{crs}, model parameter W_f, W_{cl}, B_{cl}, W'_{cl}, B'_{cl}
Output: Classification Result and Loss

for each feature in Z_{fusion}^{self} **do**
  for each feature in Z_{fusion}^{crs} **do**
    $Z_{fusion}^{cat} = W_f[Z_{fusion}^{self} \parallel Z_{fusion}^{crs}]$
  end for
end for
for each feature in Z_{fusion}^{cat} **do**
  $g_i = ReLU(W_{cl}.Z_{fusion}^{cat_i} + B_{cl})$
  $pd_i = softmax(W'_{cl}.g_i + B'_{cl})$
  $y'_i = \arg\max_q(pd_i)$
  $Loss = \frac{1}{\sum_{k=1}^{N} n(k)} \sum_{i=1}^{N} \sum_{j=1}^{n(i)} y_{ij} log y'_{ij}$
end for
Return Classification Result and Loss

V. Results And Discussion

A. Experimental Settings

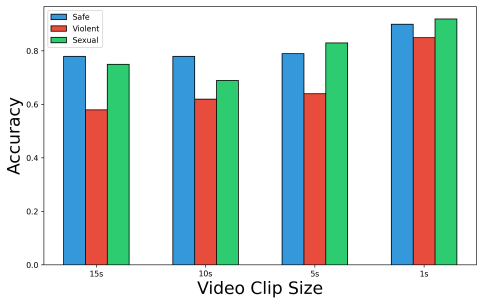
Numerous video data sets are available for research purposes. Google has introduced the YouTube-8M benchmark dataset, which includes more than 8 million video IDs associated with labels from 4 716 classes. Furthermore, various other video benchmarks focus on specific categories such as face recognition (YouTube Celebrities, YTF), sports (UCF-101,

Sports-1M), sentiment analysis, action recognition (Kinetics, HMDB51) and video captioning (MSR-VTT, MSVD). However, none of the existing benchmarks specifically addresses the proposed video classification problem. Data sets including the Elsagate dataset and the NPDI cartoon dataset. Rather, Elsagate dataset is a publicly available dataset that contains cartoon video and best candidate dataset to address our issue in which each video is classified as safe or unsafe, but in this dataset even clear frame is also labeled as unsafe that misleads to mis-classification. Moreover, it lacks the intricate behaviors associated with sensitive content. However, the NPDI dataset consists of only 900 images, which is too small and unsuitable for our deep learning-based video classification task. Anime videos serve as a suitable dataset for our experiments, being animated videos containing intermittent indecent content. Each anime series consists of a varying number of episodes, typically lasting 20 to 25 minutes. As our focus is on fine-grained detection, the dataset needs to include short-duration video clips. Consequently, we divided each episode into one-second duration clips, resulting in 109,835 video clips, encompassing all episodes in that series. To establish ground truth, a video annotation portal was developed. The annotators, which consisted of ten undergraduate and graduate students aged 20 to 25, both male and female, received detailed instructions on the annotation task. Upon logging in, they were presented with a list of videos to watch. During viewing, they were tasked with categorizing each video clip as safe, depicting violence, or containing sexual content. Clips featuring other acts (e.g., extreme bloodshed or violence, smoking, drug use, frightening or horror scenes, etc.) were excluded from this dataset. The manual annotation process resulted in a total of 111,561 video clips, including 57,908 clips classified as safe, 27,003 clips in the sexual-nudity class, and 26,650 clips in the fantasy violence class.

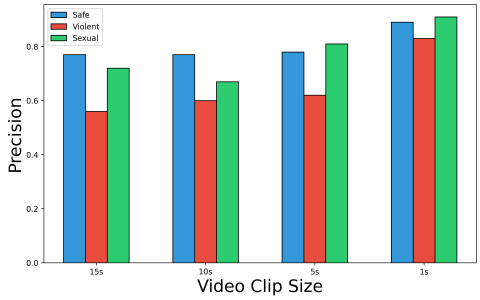
In the realm of multi-class video classification, the accurate classification of diverse classes, such as safe, violence, and sexual content, is a critical task with implications for content moderation, child safety, and platform compliance. This study undertakes a comprehensive comparative analysis of seven distinct categories of models, each designed for distinct fusion methodologies. The objective is to discern their efficacy in accurately classifying videos in multiple classes. Our experimental design meticulously evaluates the proposed methodology through a multi-pronged approach, focusing on:

- Granular Performance Analysis: We systematically vary video clip size and class distribution to quantify the methodology's sensitivity to data granularity and inherent class complexities. This analysis provides valuable insights into its robustness and adaptability in diverse video contexts.
- Bench-marking against Established Techniques: We perform a rigorous comparison of our methodology against leading video descriptors and classifiers. This benchmark sheds light on its competitive advantages and potential areas for further optimization, informs future research directions, and contributes to the advancement of the field.

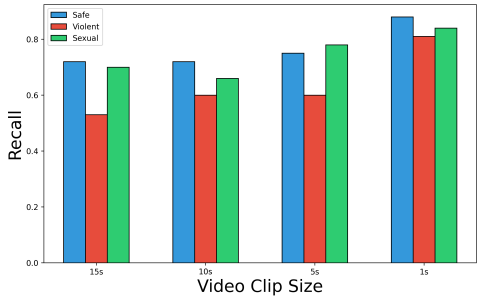
In the constant fight to protect children online, understanding the intricacies of child safety detection models is crucial. This research dives into the relationship between temporal granularity and the performance of our proposed MFusTSVD model, specifically its effectiveness in pinpointing subtle child safety concerns. Using meticulously segmented video clips of varying lengths (12, 7, 4, and 2 seconds), we trained separate MFusTSVD instances to capture the intricacies of temporal context at different scales. To maximize individual learning, we employ a transfer learning approach, initializing the encoder weights of a final, fine-tuned classifier with those of each individual model.



(a) Accuracy



(b) Precision

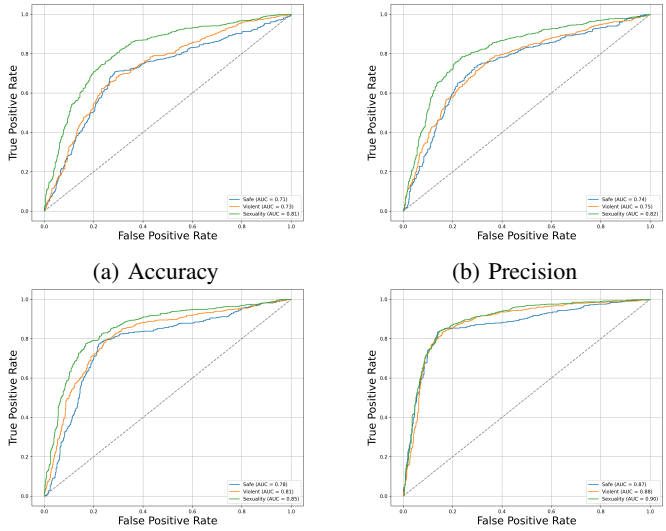


(c) Recall

Fig. 2: Fine-grained Comparative Analysis for Varying Clip Sizes

As depicted in Figure 3, we observed a significant upward trend in precision, recall, and AUC values as clip size decreased. This suggests that the MFusTSVD model thrives on shorter durations, effectively harnessing localized temporal context to pinpoint child safety concerns with remarkable precision. Figure 5 further strengthens this observation, showcasing individual ROC curves for each clip size. We witnessed a steady upward ascension with decreasing clip lengths, particularly pronounced for the "sexual" class of harmful

content. This potentially indicates an advantage of focusing on immediate actions within smaller windows for specific types of child exploitation, potentially avoiding extraneous information or redundancy that might be introduced by broader contexts. Balancing granularity with feasibility, we judiciously set the minimum clip size at 2 seconds. This provides sufficient temporal context for accurate differentiation while remaining practical for annotation and resource management.



(a) Accuracy (b) Precision

(c) Recall (d) Recall

Fig. 3: ROC Analysis for Different Instances

B. Bench-marking against Established Techniques

The chosen models include prominent architectures such as CNNs (VGG-19, ResNet, Combined), RNNs (Memory Fusion Network (MFN), Early Fusion LSTM (EF-LSTM), Late Fusion LSTM (LF-LSTM)), Transformer-based architectures (multi-modal Transformer (MulT), Interpretable multi-modal Routing (IMR)), Tensor fusion-based methods (Tensor Fusion Network (TFN), Low-rank Modality Fusion (LMF)), Graph Fusion Methods (multi-modal Graph), Transforme-LMF (LMF-MulT; Fusion-Based-CM-Attn-MulT), and our proposed model (DL-Transformer-LMF - DSC2LMVFT). To obtain the results for each baseline, we first fine-tune each model by conducting a fifty-times random grid search on the hyper-parameters. Then, we train each model again with the best hyper-parameters five times and calculate the mean results as the final result.

The comparative analysis of unsuitable content detection and classification methods, as illustrated in the provided table, reveals insightful findings regarding the performance of various benchmark approaches and the proposed MFusTSVD. The recall values for MFusTSVD are competitive, indicating its ability to effectively capture relevant instances of Safe, Violent, and Sexual content. The method's commendable average recall underscores its capability to identify instances across different content classes, ensuring comprehensive coverage. Furthermore, balanced F-measure values highlight the ability of MFusTSVD to strike a harmonious trade-off between precision and recall, further emphasizing its robust performance in content classification.

These results have significant implications for content filtering on platforms such as YouTube, suggesting that MFusTSVD could play a pivotal role in creating a safer online environment for users. The method's superior performance positions it as a promising solution for enhancing content moderation mechanisms, not only on YouTube but potentially across various domains requiring accurate and comprehensive content classification. The success of MFusTSVD in this comparative analysis underscores its potential as an advanced and effective approach to the detection and classification of unsuitable content.

VI. Conclusion

The proliferation of YouTube content has necessitated advanced content filtering mechanisms to ensure a safe and user-friendly environment. To address this, we proposed an MFusTSVD model that takes video clips as input. The input is then converted into three modalities, including text, audio, and video image. We employed a handicraft method and deep learning to extract feature from audio and video image data while the BERT transformer was used to extract textual characteristics. We also proposed a BERT-enabled low-rank multi-modal fusion-based self-modality B-SMTLMF and cross-modality B-CMTLRMF feature fusion model to perform feature fusion obtained from three different modalities. Our proposed approach, named MFusTSVD, aims to exceed existing benchmarks by achieving superior accuracy, precision, recall, and F-measure. The integration of visual fusion and transformer models offers a holistic solution to the complex task of content classification, promising greater efficiency and accuracy. The results of our comparative analysis highlight the exceptional performance of MFusTSVD when compared against established benchmarks, including the Memory Fusion Network (MFN), Early Fusion LSTM, Late Fusion LSTM, multi-modal Transformer (MulT), and others. In particular, MFusTSVD consistently outperforms these methods in terms of accuracy, precision, recall, and F-measure across various content classes. This demonstrates the model's efficacy in accurately detecting and classifying inappropriate content in YouTube videos. The balanced trade-off between precision and recall further underscores the reliability and robustness of MFusTSVD. In the future, we should focus on optimizing and fine-tuning MFusTSVD for efficiency and scalability. Large-scale real-world evaluations will provide information on its performance in diverse contexts and in different languages and cultures.

References

[1] J. Chen, Y. Hu, Q. Lai, W. Wang, J. Chen, H. Liu, G. Srivastava, A. K. Bashir, and X. Hu, "Iifdd: Intra and inter-modal fusion for depression detection with multi-modal information from internet of medical things," *Information Fusion*, vol. 102, p. 102017, 2024.

[2] T. Gerkmann, C. Breithaupt, and R. Martin, "Improved a posteriori speech presence probability estimation based on a likelihood ratio with fixed priors," *IEEE transactions on audio, speech, and language processing*, vol. 16, no. 5, pp. 910–919, 2008.

TABLE I: Comparative Analysis of Unsuitable Content Detection and Classification

Benchmark Methods	Classes	Accuracy	Average Accuracy	Precision	Average Precision	Recall	Average Recall	F-Score	Average F-Score
Memory Fusion Network (MNF)	Safe	0.86	0.86	0.86	0.87	0.88	0.88	0.86	0.87
	Violent	0.86		0.89		0.89		0.87	
	Sexual	0.87		0.88		0.88		0.88	
Late-Fusion LSTM	Safe	0.85	0.85	0.86	0.86	0.87	0.87	0.86	0.86
	Violent	0.85		0.86		0.87		0.85	
	Sexual	0.84		0.87		0.88		0.86	
Interpretable multi-modal Routing	Safe	0.86	0.86	0.85	0.85	0.88	0.88	0.86	0.86
	Violent	0.86		0.85		0.88		0.86	
	Sexual	0.87		0.84		0.87		0.87	
Early Fusion LSTM	Safe	0.89	0.89	0.91	0.91	0.90	0.89	0.92	0.92
	Violent	0.89		0.90		0.89		0.93	
	Sexual	0.90		0.89		0.89		0.92	
Tensor Fusion Network	Safe	0.86	0.86	0.87	0.87	0.89	0.89	0.87	0.87
	Violent	0.86		0.86		0.88		0.87	
	Sexual	0.85		0.87		0.89		0.87	
Proposed Method	Safe	0.92	0.93	0.91	0.91	0.94	0.94	0.93	0.93
	Violent	0.92		0.91		0.94		0.93	
	Sexual	0.91		0.90		0.93		0.92	

[3] H. Zhu, M.-D. Luo, R. Wang, A.-H. Zheng, and R. He, "Deep audio-visual learning: A survey," *International Journal of Automation and Computing*, vol. 18, no. 3, pp. 351–376, 2021.

[4] B. J. Borgström and A. Alwan, "Improved speech presence probabilities using hmm-based inference, with applications to speech enhancement and asr," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 5, pp. 808–815, 2010.

[5] M. A. Hassan, M. U. G. Khan, R. Iqbal, O. Riaz, A. K. Bashir, and U. Tariq, "Predicting humans future motion trajectories in video streams using generative adversarial network," *Multimedia Tools and Applications*, vol. 83, no. 5, pp. 15 289–15 311, 2024.

[6] S. Sulun and M. E. Davies, "On filter generalization for music bandwidth extension using deep neural networks," *IEEE Journal of Selected Topics in Signal Processing*, vol. 15, no. 1, pp. 132–142, 2020.

[7] W. Kim and R. M. Stern, "Mask classification for missing-feature reconstruction for robust speech recognition in unknown background noise," *Speech Communication*, vol. 53, no. 1, pp. 1–11, 2011.

[8] S. Poria, E. Cambria, N. Howard, G.-B. Huang, and A. Hussain, "Fusing audio, visual and textual clues for sentiment analysis from multimodal content," *Neurocomputing*, vol. 174, pp. 50–59, 2016.

[9] G. H. Park, M. W. Park, S.-C. Lim, W. S. Shim, and Y.-L. Lee, "Deblocking filtering for illumination compensation in multiview video coding," *IEEE Transactions on Circuits and systems for video technology*, vol. 18, no. 10, pp. 1457–1461, 2008.

[10] J. P. Cabral, K. Richmond, J. Yamagishi, and S. Renals, "Glottal spectral separation for speech synthesis," *IEEE Journal of Selected Topics in Signal Processing*, vol. 8, no. 2, pp. 195–208, 2014.

[11] N. Vedula, W. Sun, H. Lee, H. Gupta, M. Ogihara, J. Johnson, G. Ren, and S. Parthasarathy, "Multimodal content analysis for effective advertisements on youtube," in *2017 IEEE international conference on data mining (ICDM)*. IEEE, 2017, pp. 1123–1128.

[12] S. E. Eskimez, K. Koishida, and Z. Duan, "Adversarial training for speech super-resolution," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 2, pp. 347–358, 2019.

[13] B. Wang, J. Wang, and H. Lu, "Exploiting content relevance and social relevance for personalized ad recommendation on internet tv," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 9, no. 4, pp. 1–23, 2013.

[14] B. Yang, T. Mei, X.-S. Hua, L. Yang, S.-Q. Yang, and M. Li, "Online video recommendation based on multimodal fusion and relevance feedback," in *Proceedings of the 6th ACM international conference on Image and video retrieval*, 2007, pp. 73–80.

[15] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 2, no. 1, pp. 1–19, 2006.

[16] Y. Tanahashi and K.-L. Ma, "Design considerations for optimizing storyline visualizations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2679–2688, 2012.

[17] V. Kuleshov, S. Z. Enam, and S. Ermon, "Audio super resolution using neural networks," *arXiv preprint arXiv:1708.00853*, 2017.

[18] E. Cambria, N. Howard, J. Hsu, and A. Hussain, "Sentic blending: Scalable multimodal fusion for the continuous interpretation of semantics and sentics," in *2013 IEEE symposium on computational intelligence for human-like intelligence (CIHLI)*. IEEE, 2013, pp. 108–117.

[19] M. Paleari and B. Huet, "Toward emotion indexing of multimedia excerpts," in *2008 International Workshop on Content-Based Multimedia Indexing*. IEEE, 2008, pp. 425–432.

[20] W. Shafqat and Y.-C. Byun, "A context-aware location recommendation system for tourists using hierarchical lstm model," *Sustainability*, vol. 12, no. 10, p. 4107, 2020.

[21] Y.-H. Chung and Y.-L. Chen, "Social recommendation system with multimodal collaborative filtering," in *2021 IEEE Global Communications Conference (GLOBECOM)*. IEEE, 2021, pp. 1–7.

[22] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 941–10 950.

[23] W. Ma, J. Zhao, H. Zhu, J. Shen, L. Jiao, Y. Wu, and B. Hou, "A spatial-channel collaborative attention network for enhancement of multiresolution classification," *Remote Sensing*, vol. 13, no. 1, p. 106, 2020.

[24] W. Ma, J. Shen, H. Zhu, J. Zhang, J. Zhao, B. Hou, and L. Jiao, "A novel adaptive hybrid fusion network for multiresolution remote sensing images classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2021.

[25] X. Liu, L. Li, F. Liu, B. Hou, S. Yang, and L. Jiao, "Gafnet: Group attention fusion network for pan and ms image high-resolution classification," *IEEE Transactions on Cybernetics*, vol. 52, no. 10, pp. 10 556–10 569, 2021.

[26] N. Shvetsova, B. Chen, A. Rouditchenko, S. Thomas, B. Kingsbury, R. S. Feris, D. Harwath, J. Glass, and H. Kuehne, "Everything at once-multi-modal fusion transformer for video retrieval," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2022, pp. 20 020–20 029.

[27] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–20, 2023.