ToxVidLM: A Multimodal Framework for Toxicity Detection in Code-Mixed Videos

Krishanu Maity[1,⋆], **A.S. Poornash**[1,⋆], **Sriparna Saha**[1] **and Pushpak Bhattacharyya**[2]
[1]Department of Computer Science and Engineering, Indian Institute of Technology Patna
[2]Department of Computer Science and Engineering, Indian Institute of Technology Bombay
{krishanu_2021cs19, poornash_2101cs01, sriparna}@iitp.ac.in, pb@cse.iitb.ac.in

Abstract

In an era of rapidly evolving internet technology, the surge in multimodal content, including videos, has expanded the horizons of online communication. However, the detection of toxic content in this diverse landscape, particularly in low-resource code-mixed languages, remains a critical challenge. While substantial research has addressed toxic content detection in textual data, the realm of video content, especially in non-English languages, has been relatively underexplored. This paper addresses this research gap by introducing a benchmark dataset, the first of its kind, consisting of 931 videos with 4021 code-mixed Hindi-English utterances collected from YouTube. Each utterance within this dataset has been meticulously annotated for toxicity, severity, and sentiment labels. We have developed an advanced Multimodal Multitask framework built for **Tox**icity detection in **Vid**eo Content by leveraging Language Models (**LM**s), crafted for the primary objective along with the additional tasks of conducting sentiment and severity analysis. *ToxVidLM* incorporates three key modules – the Encoder module, Cross-Modal Synchronization module, and Multitask module – crafting a generic multimodal LM customized for intricate video classification tasks. Our experiments reveal that incorporating multiple modalities from the videos substantially enhances the performance of toxic content detection by achieving an Accuracy and Weighted F1 score of 94.29% and 94.35%, respectively.

Disclaimer: The article contains profanity, an inevitable situation for the nature of the work involved. These in no way reflect the opinion of the authors.

1 Introduction

In an age where social media platforms empower users to become content creators, the digital landscape has witnessed an unprecedented proliferation of information dissemination. By 2023, it is estimated that 82% of internet traffic will be video content (Wilson, 2022). As a result, platforms like YouTube and Dailymotion have become major sources of information. A remarkable statistic underscores the colossal impact of these platforms: on YouTube alone, users collectively view more than a billion hours of video content each day[2]. The viral nature of video content is a double-edged sword: it facilitates rapid news propagation yet simultaneously accelerates the dissemination of toxic speech. We adhere to the definition of toxic speech provided by Dixon et al. (2018), which characterizes it as *"discourteous, disrespectful, or unreasonable language likely to compel someone to exit a discussion"*.

This expansive realm of videos on platforms like YouTube encompasses an array of topics, with the majority of content being innocuous. However, there exists a darker side – videos that blatantly contravene community guidelines and foster harmful narratives (O'Connor, 2021). The non-removal of toxic content from these platforms can have severe repercussions, including the formation of hostile online environments with echo chambers of hateful users, potential loss of revenue, fines, and legal entanglements[3]. While some platforms deploy human moderators to identify and remove harmful content, the sheer volume of daily user-generated content poses an overwhelming challenge. Facebook, for instance, engages approximately 15,000 moderators to review content flagged by both AI algorithms and users but still faces approximately

[⋆] Denotes an equal contribution to this work by the respective authors and are jointly the first authors.

The code and dataset will be made available at https://github.com/justaguyalways/ToxVidLM_ACL_2024

[2] https://blog.youtube/press/

[3] https://www.wsj.com/articles/germany-to-social-networks-delete-hate-speech-faster-or-face-fines-1498757679

Findings of the Association for Computational Linguistics ACL 2024, pages 11130–11142
August 11-16, 2024 ©2024 Association for Computational Linguistics

300,000 content moderation mistakes every day[4]. Furthermore, the toll on human moderators is not limited to their professional obligations but extends to the risk of emotional and psychological trauma. This issue is further compounded by legal regulations that mandate the swift removal of hateful content. Non-compliance with these laws could result in substantial fines.

Current research in the domain of toxic speech detection primarily focuses on text-based models (Kennedy et al., 2020; Roy and Mali, 2022; Obaid et al., 2023; Maity et al., 2022b; Das et al., 2022; Maity et al., 2023), with limited exploration of image-based methodologies (Yang et al., 2019; Gomez et al., 2020; Kiela et al., 2020; Maity et al., 2022a) and very few works on video data (Wu and Bhandary, 2020; Rana and Jha, 2022; Das et al., 2023; Jha et al., 2024). Detecting harmful actions in videos requires the fusion of multi-frame video and speech processing signals, making direct adaptation of image-based hate detection methods inadequate. Existing toxic content detection methods predominantly rely on text-based modalities, with limited exploration of video content and a focus on monolingual languages like English. However, the surge in code-mixed language use, especially in multilingual countries like India, where people frequently blend Hindi and English in their communication (known as code-mixing (Myers-Scotton, 1997)), presents a unique challenge for machine learning tool development, as highlighted by (Vyas et al., 2014). While studies have addressed toxic content detection in code-mixed social media texts, a significant research gap remains in code-mixed videos.

Main Contributions: This paper strives to address these challenges by introducing a comprehensive approach for detecting toxic speech in video content, leveraging the multi-modal nature of video data and advanced deep learning techniques. Through the development of efficient models, it aims to contribute to the creation of safer online environments and facilitate compliance with evolving legal regulations concerning toxic content. Our contributions are twofold:

i) We introduce *ToxCMM*, an openly accessible dataset extracted from YouTube that is meticulously annotated for toxic speech, with utterances presented in code-mixed form. Each sentence within the videos is annotated with three crucial labels, namely Toxic (Yes / No), Sentiment (Positive / Negative / Neutral), and Severity levels (Non-harmful / Partially Harmful / Very Harmful). This extensive dataset comprises 931 videos, encompassing a total of 4021 utterances. The release of the *ToxCMM* dataset is intended to foster further exploration in the realm of multi-modal toxic speech detection within low-resource code-mixed languages.

ii) We have innovated ToxVidLM, a multimodal multitask framework for detecting toxic videos and analyzing their sentiment and severity. ToxVidLM integrates three key modules: the Encoder module, the Cross-Modal Synchronization Module, and the Multitask module, to create a versatile Multimodal LM tailored for video classification tasks. Our framework incorporates a sophisticated gated modality fusion mechanism, empirically proven to outperform standard fusion techniques in ablation studies. We propose a method for synchronizing the text modality with other modalities, yielding promising results as demonstrated in our studies. Notably, our framework is adaptable to various publicly available pre-trained models, serving as modality encoders, making it applicable to diverse problem statements. Our most effective multitask model achieves notable weighted F1-Scores of 94.35%, 86.84%, and 83.42% for Toxicity detection, Severity levels, and Sentiment identification, respectively.

2 Related Works

The widespread availability of multi-modal data has led to the utilization of multi-modal deep learning techniques, enhancing the accuracy of diverse tasks such as visual question answering (Singh et al., 2019), summarization (Ghosh et al., 2024b,a) and the detection of fake news and rumors (Khattar et al., 2019). In recent times, multi-modal hate speech detection has gained traction, where text posts are augmented with additional contextual information such as user and network data (Founta et al., 2019) or images (Yang et al., 2019; Gomez et al., 2020; Kiela et al., 2020; Maity et al., 2022a) to bolster detection accuracy. These multi-modal approaches often involve the utilization of unimodal methods like CNNs, LSTMs, or BERT for text encoding and deep CNNs like ResNet or InceptionV3 for image encoding. Subsequently, multimodal fusion is performed through techniques like simple concatenation, gated summation, bilinear

[4] https://www.forbes.com/sites/johnkoetsier

transformation, or attention-based methods. Additionally, the application of multi-modal transformers such as ViLBERT and Visual BERT has been explored (Kiela et al., 2020).

While research on the detection of offensive or toxic videos is scarce, particularly in the context of languages such as Portuguese (Alcântara et al., 2020), Thai (Maity et al., 2024) and English (Wu and Bhandary, 2020; Rana and Jha, 2022; Das et al., 2023), it is worth noting that the existing work in this area predominantly revolves around monolingual languages. Maity et al. (2024) leverages textual data to propose a dual-channel network to classify hate and sentiment in low-resource settings. Notably, the two studies (Alcântara et al., 2020; Wu and Bhandary, 2020) exclusively considered textual features by extracting video transcripts for classification. In contrast, (Rana and Jha, 2022)'s research takes both textual and audio features for offensive video detection. However, this study confronts issues related to dataset accessibility, insufficiently detailed data curation and annotation processes, and a lack of precise dataset statistics. Das et al. (2023) developed a more comprehensive approach by integrating all three modalities (text, Image, and audio) for hate video detection in English.

To the best of our knowledge, our study pioneers the introduction of a multi-modal toxic video dataset in the context of low-resource code-mixed languages, further distinguished by the annotation of sentence-level labels. We are confident that our dataset, along with the benchmark models developed using it, will significantly assist content moderators in distinguishing genuine cases of hateful content while concurrently reducing false alarms.

3 Toxic Code-Mixed Multimodal (*ToxCMM*) Dataset Creation

Data Collection: We selected YouTube as our primary data source, given its popularity as a video hosting platform. Our focus was on code-mixed language conversations, primarily in Hindi and English. To collect relevant content, we utilized the YouTube API to scrape Indian web series and Hindi "roasted" videos. We subdivided the downloaded videos into smaller sub-videos to annotate them at the sentence level and maximize the inclusion of toxic content. Initially, we obtained 1023 videos, but after a thorough review, we retained 931 videos as the remaining ones were mostly in English, not the intended Hindi-English code-mixed format. To generate transcripts for each video, we used the Whisper (Radford et al., 2023) transcribing model, configured with word timestamps from the OpenAI library. We then manually improved the transcript quality by removing unclear words and symbols resulting from speech disruptions or stammering. Extracting individual utterances from the videos involved cataloging their start and end times.

3.1 Data Annotation

To better clarify the annotation process, we split the annotation section into two subsections: (i) Annotation Training and (ii) Main Annotation.

Annotation training: Three PhD scholars oversaw the annotation process, well-versed in toxic and offensive content, and the actual annotations were conducted by three undergraduate students proficient in both Hindi and English. Initially, we hired a group of masters students in linguistics who volunteered via our department email list and compensated them with gift vouchers and an honorarium. To train our annotators, we required gold standard samples with annotations for toxicity, severity, and sentiment labels. Our expert annotators randomly selected 150 samples (a small video of one sentence) and assigned suitable target classes. We considered two toxicity classes (Non-toxic/toxic), three sentiment classes (positive/neutral/negative), and the severity score on a three-point scale (0, 1, 2) for each video sample. Score 0 signifies that there is no indication of toxicity and 1 indicates that the post contains indications of toxicity. However, they are not severe, and a score of 2 indicates that the post contains strong evidence of toxicity (e.g., physical threats or excitement to commit suicide). Expert annotators engaged in discussions to resolve any differences and created 150 gold-standard samples with rationale and target annotations. These 150 annotated examples were divided into three sets, each containing 50 samples, to facilitate a three-phase training process. After each phase, expert annotators collaborated with novice annotators to correct any inaccuracies in the annotations, and the annotation guidelines were updated as needed. Following the conclusion of the third round of training, the top three annotators were selected to annotate the entire dataset containing 4021 samples.

Main annotation: We began with a small batch of 100 samples, gradually increasing it to 500 as an-

Table 1: Class-wise data statistics of our developed *ToxCMM* dataset

# Video	#Utterances	Toxicity		Severity			Sentiment		
		Non-Toxic	Toxic	0	1	2	Positive	Neutral	Negative
931	4021	2324	1697	2324	834	863	469	1401	2151

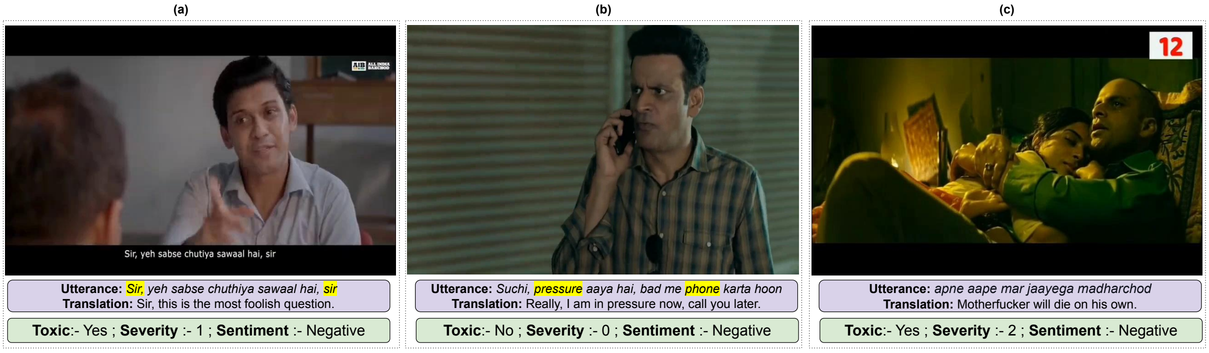

(a)
Sir, yeh sabse chutiya sawaal hai, sir
Utterance: Sir, yeh sabse chuthiya sawaal hai, sir
Translation: Sir, this is the most foolish question.
Toxic:- Yes ; Severity :- 1 ; Sentiment :- Negative
(b)
Utterance: Suchi, pressure aaya hai, bad me phone karta hoon
Translation: Really, I am in pressure now, call you later.
Toxic:- No ; Severity :- 0 ; Sentiment :- Negative
(c)
12
Utterance: apne aape mar jaayega madharchod
Translation: Motherfucker will die on his own.
Toxic:- Yes ; Severity :- 2 ; Sentiment :- Negative

Figure 1: Some samples from annotated *ToxCMM* dataset; The yellow highlighted words are in English.

notators improved. To ensure consistency, we corrected errors from previous batches, and for final labels, majority voting was employed. In cases of disagreements among annotators, expert input was sought. Annotators were instructed to be unbiased in their assessments. The quality of annotations was assessed using Fleiss' Kappa scores (Fleiss, 1971), resulting in IAA scores of 0.74 for toxicity classification, 0.67 for sentiment classification, and 0.64 for severity detection, confirming dataset quality and reliability. Figure 1 shows some samples from annotated *ToxCMM* dataset. Sample (b) shows a non-toxic video with negative sentiment. In contrast, both samples (a) and (c) are identified as toxic videos with negative sentiments. However, sample (a) is deemed less severe, while sample (c) is considered more severe due to its explicit use of profanity, aggressive language, and the wish for harm or death upon someone.

3.2 Dataset Statistics

The *ToxCMM* dataset comprises a total of 4021 utterances, with 1,697 categorized as toxic and the remaining 2,324 labelled as non-toxic. Class-wise statistics for the *ToxCMM* dataset are provided in Table 1. Each utterance in this dataset contains an average of 8.68 words, with an average duration of 8.89 seconds. On average, each utterance in the dataset contains about 68.20% Hindi words, which means that more than two-thirds of the words are in Hindi, while the remaining words are in English.

4 Methodology

Problem Formulation: We formulate our problem as follows: Given an utterance video clip denoted as V, our task is essentially a classification problem. We aim to determine whether the video contains toxic content, as well as assign sentiment and severity labels to it. Each video, V, is expressed as a sequence of frames, $F = \{f_1, f_2, \dots f_n\}$, accompanied by its associated audio A that is sampled at 16kHz to construct a sequence of features $A = \{a_1, a_2, \dots a_l\}$ and a transcript of the video, $T = \{w_1, w_2, \dots w_m\}$, which consists of a sequence of words. Our goal is to construct a deep learning-based video classifier, denoted as $C : C(T; F; A) \rightarrow y$, where y signifies the actual label of the video for a given task.

In this section, we describe our developed LM-based multimodal-multitask framework *ToxVidLM* (see Figure 2) for toxic video detection and its sentiment and severity analysis. To enhance comprehension of our proposed method, we partition it into three distinct components: namely, the Encoder module, Cross Modal Synchronization Module, and the Multitask module.

4.1 Encoder Module

Current transformed-based language models (LMs) exhibit substantial capability but are commonly constrained to processing textual data exclusively. In this section, we elucidate our approach to encoding information from diverse modalities.

Audio Encoder We conducted experiments using two state-of-the-art (SOTA) models, namely

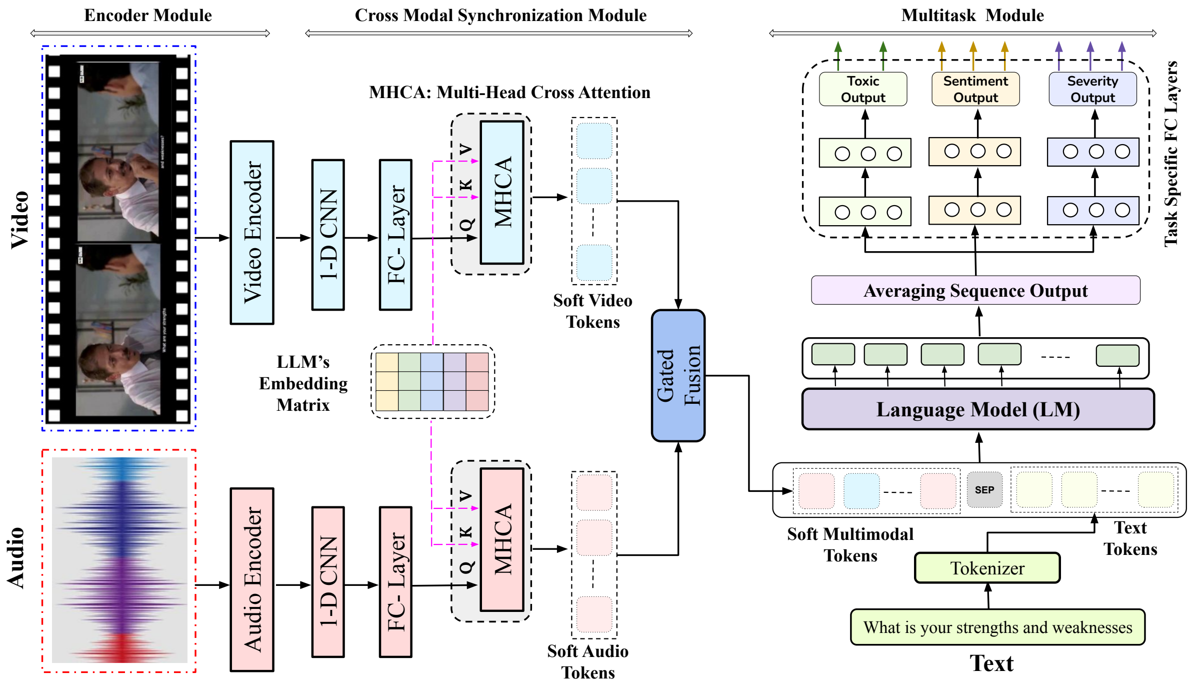


Figure 2: Architecture of proposed *ToxVidLM* model.

Whisper (Radford et al., 2023) and MMS (Pratap et al., 2023), to encode the audio signals and derive meaningful representations from the audio data.

Video Encoder: To handle the spatial and temporal information in the videos, we consider two vision-based models, VideoMAE (Tong et al., 2022) and Timesformer (Bertasius et al., 2021).

Text Encoder: To generate text embeddings, we leverage the BERT (Vaswani et al., 2017) family, renowned for its high effectiveness in various NLP tasks. Given our focus on Hindi-English code-mixed text, we have conducted experiments with three various models in HingBERT (Nayak and Joshi, 2022) family, like HingMBERT, HingRoBERTa, HingGPT and IndicBERT (Kakwani et al., 2020). These models are pre-trained on Hindi-English code-mixed Roman text.

As per the results obtained, we utilize pre-trained models such as VideoMAE and Whisper to encode the input video (V) and audio (A) inputs, respectively, as outlined below:

$$Z_v = \text{VideoMAE}(V) \tag{1a}$$

$$Z_a = \text{Whisper}(A) \tag{1b}$$

Also, $Z_v \in \mathbb{R}^{SL_v \times d_v}$ and $Z_a \in \mathbb{R}^{SL_a \times d_a}$ respectively. Here, SL_v and SL_a represent the sequence lengths of video and audio inputs, respectively.d_v and d_a represent the embedding dimension of encoded audio and video, respectively. Please see Appendix B for more details on audio, video and text encoders used in our study.

4.2 Cross Modal Synchronization Module

Modality encoders are typically trained independently, resulting in discrepancies among the generated representations. Consequently, it becomes imperative to synchronize these distinct representations within a unified space to enhance the overall coherence and effectiveness of multimodal processing. In this section, we present a detailed methodology for aligning these representations.

Modality Synchronization The synchronization strategy aims to effectively correlate features extracted from multiple modalities, such as audio and video, with a primary focus on textual features. This emphasis is due to the heightened significance of textual information in addressing our specific problem statement. Textual features are preferred over auditory and visual signals due to their comparatively lower susceptibility to noise. This empirical observation is consistently reflected in the obtained results, emphasizing the importance of robust textual representation.

The procedure for modality synchronization is delineated as follows:

(1) Abstract Feature Extraction: To mitigate computational expenses and limit the token count in the prefix, we utilize a 1-D convolutional layer

(*Conv*) to compress the length of multi-modal features to a condensed and consistent value. Following this, a linear layer (*FC*) is employed to modify the hidden size of the features, aligning it with the dimensions of the token embeddings in the LMs, as described below:

$$C_v = FC(Conv(Z_v)) \tag{2a}$$

$$C_a = FC(Conv(Z_a)) \tag{2b}$$

where $C_v \in \mathbb{R}^{SL^{'} \times d_t}$ and $C_a \in \mathbb{R}^{SL^{'} \times d_t}$ are the abstract features with a fixed length of $SL^{'}$. Here, d_t represents the dimensionality of the embedding matrix $E_{llm} \in \mathbb{R}^{V \times d_t}$ associated with the textual LMs (i.e., HingRoberta or HingGPT in our work) with vocabulary size V.

(2) Multi-Head Cross Attention (MHCA) To establish a unified representation space guided by the text modality, we implemented Multi-Head Cross-Attention (MHCA) on the abstract video and audio features obtained from the preceding layer. MHCA involves scaled dot-product attention applied to three inputs: Query (Q) from one modality, Key (K) and Value (V) from another. This attention mechanism computes attention weights by comparing the queries Q with the keys K, updating the query representations via a weighted sum of the values V, as described below:

$$MHCA(Q, K, V) = Softmax\left(\frac{QK^{\,T}}{\sqrt{d_k}}\right)V \tag{3}$$

where d_k is the dimensionality of the key and query vectors. Using the attention mechanism in Equation 3, we propose to align the audio and visual representations with the textual embedding space as follows:

$$\begin{aligned} C_v^s &= MHCA(C_v, E_t, E_t), \\ C_a^s &= MHCA(C_a, E_t, E_t) \end{aligned} \tag{4}$$

In our research, we view the attenuated representations of visual (C_v^s) and audio (C_a^s) modalities derived from Equation 4 as the soft tokens utilized by the LM, acting as the input to the Multitask module.

(3) Gated Fusion To combine soft video and audio tokens, we employ a gated fusion strategy. Unlike concatenation or directly assigning weights to each vector, the gate fusion mechanism enables varying contributions to the prediction from different positions of vectors. The joint representation resulting from the gate fusion is computed as follows:

$$\begin{aligned} \alpha &= \sigma(\mathbb{P}_v C_v^s \,+\, \mathbb{P}_a C_a^s \,+\, b_g), \\ J_{va} &= \alpha\, C_a^s \,+\, (1-\alpha) C_v^s \end{aligned} \tag{5}$$

Here, $\mathbb{P}_v$ and $\mathbb{P}_a$ represent weight matrices for the visual and acoustic modalities, while b_g denotes scalar bias and σ is the sigmoid activation function.

4.3 Multitask Module

Typically, the LM model processes the input transcript T, generating a text token embedding E_t with dimensions $SL_t \times d_t$. Here, SL_t is the maximum sequence length of the transcript, and d_t is the embedding dimension. Here we have added the joint multimodal soft tokens (J_{va}) obtained from the Cross-Modal Synchronization Module, appended with the text tokens separated by the special token [SEP], thereby creating a multimodal input to the LM for a more comprehensive understanding of the input video. Subsequently, the sequence output from the LM undergoes averaging and is passed through three task-specific fully connected layers, followed by an output softmax layer, facilitating the concurrent solution of three tasks: toxicity, severity, and sentiment detection from a video.

4.4 Loss Function

The loss function used in all tasks is categorical cross-entropy. The final loss function ($Loss_f$) is a weighted sum of individual task-specific losses ($Loss_s$) for M tasks, where the contribution of each task's loss to the overall loss is determined by the loss weight β as shown in Equation (6).

$$Loss_f = \sum_{k=1}^{M} \beta_k Loss_s^k \tag{6}$$

Where the parameters β_i are learnt end-to-end, signifying task contribution from task i to the multitask loss, enabling differential importance for parameter updates across tasks.

5 Experimental Results and Analysis

Experimental Settings: All experiments were conducted on a machine equipped with an Intel Xeon Gold 5218 CPU featuring 64 cores and 128 threads, coupled with four Nvidia Tesla V100 GPUs with VRAM memory of 30 GB per GPU card. For the experiments' preparation, the dataset was partitioned into testing, validation, and training sets at ratios of 10%, 10%, and 80%, respectively. To ensure robustness, the models were trained ten times with different random splits, and the average performance was reported. Several network configurations were tested, with the best results achieved using the Adam optimizer (Kingma and Ba, 2014) with a Cosine Annealing Learning Rate scheduler (Loshchilov and Hutter, 2016), batch size of 2, a learning rate set to $1e^{-5}$, and training for 30 epochs. All models were implemented in the PyTorch framework[5].

Baseline Setup We have implemented the baseline model as outlined in the study by Das et al. (2023) which introduced a multimodal dataset designed for Hate-speech classification. In their proposed architecture, each modality is processed separately through a transformer-based encoder, followed by modality-specific fully connected layers. Subsequently, a Fusion Layer concatenates the modality-specific representations, which are then forwarded to a fully connected classification layer. The output dimension of this final layer corresponds to the number of classes in the classification task. The entire model is trained using a single cross-entropy loss function. We have exclusively conducted single-task experiments for the three tasks (Toxicity detection, Severity detection, and Sentiment classification) in unimodal/bimodal/trimodal settings. A diagram illustrating baseline models is presented in Figure 3

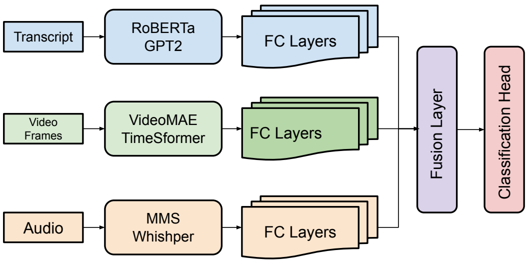


Figure 3: A schematic of baselines models as mentioned in (Das et al., 2023)

Table 2: The outcomes of various transformer-based baseline models are presented within different modalities, i.e., Video (V), Audio (A), and Text (T) configurations for three tasks (Toxicity, Severity, and Sentiment). The results are measured in terms of weighted-average-F1 score (F1) and Accuracy (Acc) values. Bold-faced values represent the maximum scores attained; TF - Timesformer, VM - VideoMAE, WP - Whisper, M-BERT - HingMBERT, GPT2 - HingGPT, RT - HingRoberta.

M	Model	Toxic		Severity		Sentiment	
		F1	Acc	F1	Acc	F1	Acc
	Baselines : Unimodal						
V	TF	66.23	66.47	60.38	62.88	57.12	58.24
	VM	68.67	68.73	61.42	64.26	58.69	60.79
A	MMS	75.81	75.89	67.21	67.35	64.28	64.14
	WP	76.18	76.17	68.97	67.99	65.45	65.01
T	I-BERT	81.72	81.86	73.22	73.44	68.25	68.48
	M-BERT	83.28	83.23	75.35	76.17	71.26	71.15
	GPT2	85.67	85.71	75.02	75.93	73.11	72.72
	RT	**86.98**	**86.95**	**77.23**	**76.42**	**73.41**	**74.44**
	Baselines : Bimodal						
V + A	TF + MMS	76.49	76.53	68.74	68.89	64.31	64.24
	VM + MMS	77.25	77.48	69.42	69.28	65.18	65.49
	TF + WP	78.72	78.94	70.34	70.57	67.85	67.41
	VM + WP	79.23	79.41	71.79	72.71	69.45	69.47
T + V	GPT2 + TF	85.87	85.84	75.13	75.91	73.12	73.91
	GPT2 + VM	86.71	86.84	76.12	76.98	74.03	74.45
	RT + TF	87.01	87.93	76.92	77.19	73.92	74.56
	RT + VM	87.11	87.08	77.68	78.11	74.11	74.59
T + A	GPT2 + MMS	86.27	86.35	75.03	76.21	73.15	72.87
	GPT2 + WP	86.77	86.84	75.73	76.42	74.06	73.94
	RT + MMS	87.18	87.19	77.22	78.15	74.08	74.69
	RT + WP	**87.26**	**87.33**	**77.42**	**78.41**	**74.22**	**74.93**
	Baselines : Trimodal						
T + V + A	GPT2+TF+MMS	86.92	87.02	75.85	76.61	74.27	74.69
	GPT2+VM+MMS	86.72	87.12	75.72	76.48	74.23	74.87
	GPT2+TF+WP	86.88	87.27	75.89	76.64	74.31	74.75
	GPT2+VM+WP	87.21	87.34	77.14	77.17	74.43	74.78
	RT+TF+MMS	87.29	87.39	77.93	78.47	75.09	75.65
	RT+VM+MMS	87.58	87.79	77.96	78.72	75.03	75.49
	RT+TF+WP	87.82	87.67	78.16	78.36	75.11	75.66
	RT+VM+WP	**88.09**	**88.08**	**78.19**	**78.66**	**75.78**	**75.82**

5.1 Findings from Experiments

Table 2 shows the results of toxicity, severity and sentiment classification tasks with different baseline models. Results of the proposed *ToxVidLM* model are shown in Table 3. From all these reported results, we can conclude the following:

(1) Our experimentation involved four text encoders (HingRoberta, HingGPT, IndicBERT, HingMBERT), two video models (VideoMAE, Timesformer), and two audio encoders (Whisper, MMS) to discern optimal performers for toxic video detection in Hindi-English code-mixed language. Analysis of the outcomes, presented in Table 2, reveals the superior performance of HingRoberta/HingGPT, VideoMAE, and Whisper as the best encoders for text, video, and audio modalities, respectively. These top-performing models are subsequently incorporated into our proposed framework.

(2) Across all three tasks, unimodal baselines demonstrate that the text modality consistently outperforms video and audio modalities. This underscores the paramount importance of text modality

[5] https://pytorch.org/.

Table 3: Results of proposed *ToxVidLM* framework with different modality configurations for three tasks (Toxicity, Severity and Sentiment classification) in single and multitask settings; TF - Timesformer, VM - VideoMAE, WP - Whisper, M-BERT - HingMBERT, GPT2 - HingGPT, RT - HingRoberta.

Modality	**Encoder**	**Toxicity**		**Severity**		**Sentiment**	
		F1	**Acc**	**F1**	**Acc**	**F1**	**Acc**
Single Task							
T+V	GPT2 + VM	90.15	90.09	83.35	82.99	75.64	75.26
	RT + VM	91.11	91.03	83.38	83.87	80.46	80.09
T+A	GPT2 + WP	91.27	91.06	84.49	83.89	76.61	76.19
	RT + WP	92.64	92.58	84.46	85.06	81.16	81.27
T+V+A	GPT2 + VM+ WP	92.14	91.98	84.86	84.92	78.14	78.23
	RT + VM + WP	93.85	93.64	86.28	86.51	82.76	82.87
Multi-Task (Toxic + Severity)							
T+V	GPT2 + VM	91.53	91.31	84.68	84.19	-	-
	RT + VM	92.07	92.38	84.88	85.56	-	-
T+A	GPT2 + WP	92.47	92.28	85.79	85.28	-	-
	RT + WP	93.61	93.58	85.87	86.43	-	-
T+V+A	GPT2 + VM+ WP	93.48	93.29	86.54	86.12	-	-
	RT + VM + WP	94.12	93.87	86.56	86.82	-	-
Multi-Task (Toxic + Sentiment)							
T+V	GPT2 + VM	91.64	91.25	-	-	76.09	76.44
	RT + VM	92.33	92.66	-	-	81.57	81.72
T+A	GPT2 + WP	92.51	92.34	-	-	77.73	77.27
	RT + WP	93.78	93.71	-	-	82.98	82.64
T+V+A	GPT2 + VM+ WP	93.45	93.23	-	-	79.79	79.92
	RT + VM + WP	94.06	93.94	-	-	83.05	82.18
Multi-Task (Toxic + Severity + Sentiment)							
T+V	GPT2 + VM	92.72	92.65	85.89	85.59	78.22	77.59
	RT + VM	93.51	93.01	85.48	85.76	82.38	82.49
T+A	GPT2 + WP	92.81	92.59	85.97	85.52	78.19	77.38
	RT + WP	93.88	93.73	85.91	86.01	82.45	82.66
T+V+A	GPT2 + VM+ WP	93.72	93.56	86.71	86.39	80.03	80.21
	RT + VM + WP	**94.35**	**94.29**	**86.84**	**87.12**	**83.42**	**82.43**

in toxicity detection within videos. Notably, employing the HingRoberta model for the text modality yields the highest F1 score of 86.98% in toxicity detection, while VideoMAE and Whisper models achieve accuracy values of 68.67% and 76.18%, respectively, in video and audio modalities. Similar trends are observed in the other two tasks, namely severity and sentiment analysis. Hence we prioritize text modality as a base in our proposed model's performance analysis.

(3) In the context of bimodal baselines, the text+audio configuration consistently demonstrates superior performance compared to the other two combinations across all tasks. Notably, the most effective baseline considering three modalities (RT+VM+WP) achieves the highest accuracy of 88.08%, 78.66%, and 75.82% for toxicity, severity, and sentiment tasks, respectively.

(4) In both single-task and multitask scenarios, our proposed model (*ToxVidLM*) consistently outperforms all baselines by a substantial margin. In the three-task settings, the (RT + VM + WP) variants exhibit enhancements in accuracy, surpassing the best baseline model by 6.21%, 8.46%, and 6.61% for toxicity, severity, and sentiment tasks, respectively. These notable improvements underscore the effectiveness of our proposed model, attributed to the incorporation of an innovative modality synchronization module compared to baseline approaches.

(5) The multitask (MT) variants of our proposed model consistently surpass its single-task (ST) variant across all tasks within the same encoding setting. Notably, in a two-task scenario utilizing GPT2+VM encoders, the MT variant outperforms the ST counterpart, demonstrating improvements in F_1-score of 2.57%, 2.54%, and 2.58% for toxicity, severity, and sentiment tasks, respectively. These results suggest that incorporating sentiment and severity knowledge enhances the performance of the toxicity detection task and contributes to overall model improvement. Conversely, no substantial improvements are observed when comparing multitask settings with two tasks versus three tasks.

Statistical Analysis: A statistical t-test was conducted on the values from ten runs of both the proposed models and baseline models, yielding p-values below 0.05, indicating the statistical significance of the results. The t-test was implemented using functions from the scipy library[6]. We have highlighted (gray color) the results in Table 2 and 3 which are statistically significant.

5.1.1 Ablation Study

Table 4: Ablation study to show the effect of gated fusion (GF), multi-head cross attention (MHCA) in proposed *ToxVidLM* model

Model	**Toxic**		**Severity**		**Sentiment**	
	F1	**Acc**	**F1**	**Acc**	**F1**	**Acc**
ToxVidLM	**94.35**	**94.29**	**86.84**	**87.12**	**83.42**	**82.43**
- GF	92.63	92.84	84.11	84.59	80.29	80.54
- MHCA	89.87	89.92	81.45	81.56	77.38	77.62
- MHCA - GF	87.72	87.86	78.22	78.44	75.25	75.48

We conducted an ablation study (see Table 4) on our proposed model, *ToxVidLM*, to elucidate the impact of gated fusion (GF) and multi-head cross attention (MHCA) in toxic video detection. The removal of GF from *ToxVidLM* results in a discernible decrease of 1.72%, 2.73%, and 3.13% in F1-score for toxicity, severity, and sentiment tasks, respectively. This decline underscores the crucial role of the gated fusion module in effectively fusing video and audio modalities, thereby enhancing overall performance. Upon removing

[6] https://docs.scipy.org/doc/scipy-1.6.3/reference/generated/scipy.stats.ttest_ind.html

the MHCA component from *ToxVidLM*, a substantial performance drop is observed across all tasks, affirming the significant impact of MHCA in our proposed model. This proves MHCA is instrumental in generating text-guided audio and video features. Notably, the simultaneous exclusion of both GF and MHCA components results in a substantial drop of 6.63% in the F1 score for the toxicity task, with similar performance reductions observed in other tasks. This considerable decline underscores the pivotal role of the cross-modal synchronization module, emphasizing its capacity to align representations from three distinct modalities within a unified space. Please see the qualitative analysis of our proposed framework in Appendix A.

6 Conclusion and Future Works

In an ever-evolving internet landscape, where videos have become the predominant form of content, the challenge of detecting toxic content, especially in low-resource code-mixed languages, is more critical than ever. We introduce *ToxCMM*, a pioneering benchmark dataset featuring code-mixed videos for toxic content detection. Our proposed LM-based advanced multimodal framework (*ToxVidLM*) achieved remarkable results, emphasizing the significance of combining text, audio, and video modalities. It is worth noting that, among the individual modalities, transformer encodings of text prove to be particularly effective in detecting toxic videos. Beyond toxicity, the *ToxCMM* dataset includes two additional labels, sentiment, and severity, offering a comprehensive resource for further exploration in sentiment analysis within low-resource code-mixed videos. Our research emphasizes AI's role in fostering a respectful online environment and promoting civility against toxic speech in video data.

7 Limitations

Our endeavour aimed to construct a multimodal framework and introduce a benchmark dataset, ToxCMM, tailored for detecting toxic video content within code-mixed language. However, it is crucial to acknowledge certain inherent limitations in our proposed approach and dataset, including:

1. In this study, we did not consider the context of the video clip; we treated a single utterance as the input post. Future investigations will incorporate the entire video clip as input, recognizing the pivotal role of context in deciphering the actual meaning of the utterance.

2. Implicit or indirect toxic expressions were excluded from this study, primarily focusing on explicit markers. Future work will address the development of datasets and models capable of detecting implicit/indirect toxic posts.

3. The proposed *ToxVidLM* model fine-tunes two encoder modules, an LM, and additional modules, requiring a considerable amount of GPU memory for training. Due to computational limitations, we were unable to experiment even with parameter-efficient fine-tuning methods (PEFT) like LoRA (Hu et al., 2021) or Quantized-LoRA (Dettmers et al., 2023) for models containing billions of parameters like OpenHathi-7B[7], Airavata-7B (Gala et al., 2024), Llama 2-7B (Touvron et al., 2023), or Mistral-7B (Jiang et al., 2023). However, since our model is versatile, those with ample GPU resources can easily substitute larger models into the textual side, potentially achieving enhanced performance specifically for video classification tasks.

References

Cleber Alcântara, Viviane Moreira, and Diego Feijo. 2020. Offensive video detection: dataset and baseline results. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4309–4319.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4.

Mithun Das, Somnath Banerjee, and Animesh Mukherjee. 2022. Data bootstrapping approaches to improve low resource abusive language detection for indic languages. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 32–42.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1014–1023.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

[7] https://huggingface.co/sarvamai/OpenHathi-7B-Hi-v0.1-Base

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Antigoni Maria Founta, Despoina Chatzakou, Nicolas Kourtellis, Jeremy Blackburn, Athena Vakali, and Ilias Leontiadis. 2019. A unified deep learning architecture for abuse detection. In *Proceedings of the 10th ACM conference on web science*, pages 105–114.

Jay Gala, Thanmay Jayakumar, Jaavid Aktar Husain, Mohammed Safi Ur Rahman Khan, Diptesh Kanojia, Ratish Puduppully, Mitesh M Khapra, Raj Dabre, Rudra Murthy, Anoop Kunchukuttan, et al. 2024. Airavata: Introducing hindi instruction-tuned llm. *arXiv preprint arXiv:2401.15006*.

Akash Ghosh, Arkadeep Acharya, Raghav Jain, Sriparna Saha, Aman Chadha, and Setu Sinha. 2024a. Clipsyntel: clip and llm synergy for multimodal question summarization in healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22031–22039.

Akash Ghosh, Arkadeep Acharya, Prince Jha, Sriparna Saha, Aniket Gaudgaul, Rajdeep Majumdar, Aman Chadha, Raghav Jain, Setu Sinha, and Shivani Agarwal. 2024b. Medsumm: A multimodal approach to summarizing code-mixed hindi-english clinical queries. In *European Conference on Information Retrieval*, pages 106–120. Springer.

Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring hate speech detection in multimodal publications. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1470–1478.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Prince Jha, Krishanu Maity, Raghav Jain, Apoorv Verma, Sriparna Saha, and Pushpak Bhattacharyya. 2024. Meme-ingful analysis: Enhanced understanding of cyberbullying in memes through multimodal explanations. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 930–943, St. Julian's, Malta. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. *arXiv preprint arXiv:2005.02439*.

Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*, pages 2915–2921.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in neural information processing systems*, 33:2611–2624.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

Krishanu Maity, Raghav Jain, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2023. Genex: A commonsense-aware unified generative framework for explainable cyberbullying detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16632–16645.

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022a. A multitask framework for sentiment, emotion and sarcasm aware cyberbullying detection from multi-modal code-mixed memes. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1739–1749.

Krishanu Maity, A. S. Poornash, Shaubhik Bhattacharya, Salisa Phosit, Sawarod Kongsamlit, Sriparna Saha, and Kitsuchart Pasupa. 2024. Hatethaisent: Sentiment-aided hate speech detection in thai language. *IEEE Transactions on Computational Social Systems*, pages 1–14.

Krishanu Maity, Sriparna Saha, and Pushpak Bhattacharyya. 2022b. Emoji, sentiment and emotion aided cyberbullying detection in hinglish. *IEEE Transactions on Computational Social Systems*.

Carol Myers-Scotton. 1997. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Ravindra Nayak and Raviraj Joshi. 2022. L3cube-hingcorpus and hingbert: A code mixed hindi-english dataset and bert language models. *arXiv preprint arXiv:2204.08398*.

Mohammed Hussein Obaid, Shawkat K Guirguis, and Saleh M Elkaffas. 2023. Cyberbullying detection and severity determination model. *IEEE Access*.

J. O'Connor. 2021. Building greater transparency and accountability with the Violative View Rate — blog.youtube. [Accessed 28-10-2023].

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2023. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Aneri Rana and Sonali Jha. 2022. Emotion based hate speech detection using multimodal learning. *arXiv preprint arXiv:2202.06218*.

Pradeep Kumar Roy and Fenish Umeshbhai Mali. 2022. Cyberbullying detection using deep transfer learning. *Complex & Intelligent Systems*, 8(6):5449–5467.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pretraining. *Advances in neural information processing systems*, 35:10078–10093.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 974–979.

A. Wilson. 2022. video marketing statistics you simply can't overlook. [Accessed 28-10-2023].

Ching Seh Wu and Unnathi Bhandary. 2020. Detection of hate speech in videos using machine learning. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 585–590. IEEE.

Fan Yang, Xiaochang Peng, Gargi Ghosh, Reshef Shilon, Hao Ma, Eider Moore, and Goran Predovic. 2019. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proceedings of the third workshop on abusive language online*, pages 11–18.

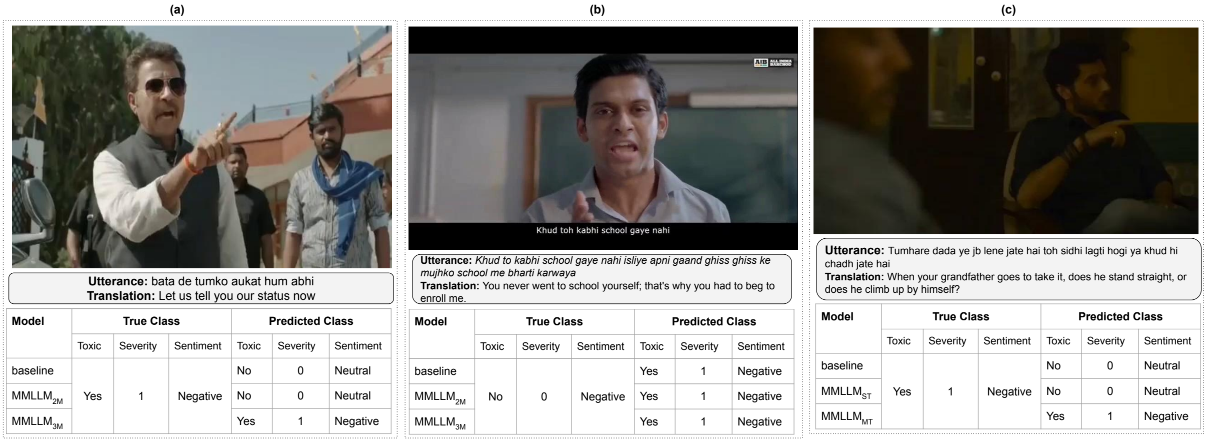
A Qualitative Analysis

Figure 4 illustrates qualitative results comparing the insightfulness of various tasks between ground truth and model predictions.

(i) In the first sample (a), the textual modality ("Let us tell you our status now") lacks offensive words but constitutes a toxic utterance with negative sentiment and a severity score of 1 based on acoustic and visual expression. The user is threatening someone. Both the best baseline model and our proposed model with two modalities settings (ToxVidLM$_{2M}$) make incorrect predictions across all tasks. However, considering three modalities variants (ToxVidLM$_{3M}$) allows accurate prediction of all classes. This observation underscores our proposed model's enhanced comprehension of diverse modalities, demonstrating that incorporating audio and visual cues with text provides a superior understanding of video data.

(ii) The second example's true labels are non-toxic with negative sentiment and a severity score of 0. Regrettably, both baseline and proposed models mispredict toxicity and severity labels. Although the surface sentiment appears negative, and some negative words and angry facial expressions are present, understanding the actual implicit meaning requires knowledge of the video clip's previous context. Since this study focuses on stand-alone utterance labels without considering previous context, all models make incorrect predictions. Integrating context represents a potential future direction for this work.

(iii) In the third example, both baseline and single-task variants of the proposed model (ToxVidLM$_{ST}$) inaccurately predict all labels, while the multitask model (ToxVidLM$_{MT}$) correctly identifies all classes. This example is toxic due to offensive language and derogatory assumptions about someone's grandfather. The terms "sidhi lagti hogi" and "khud hi chadh jate hai" are disrespectful, implying negative sentiment. Task-specific layers in the multitask framework aid in correctly identifying true sentiment, leading to the accurate identification of toxicity and severity labels in the given post.



(a)

Utterance: bata de tumko aukat hum abhi
Translation: Let us tell you our status now

Model	True Class			Predicted Class		
	Toxic	Severity	Sentiment	Toxic	Severity	Sentiment
baseline	Yes	1	Negative	No	0	Neutral
MMLLM$_{2M}$				No	0	Neutral
MMLLM$_{3M}$				Yes	1	Negative

(b)

Utterance: *Khud to kabhi school gaye nahi isliye apni gaand ghiss ghiss ke mujhko school me bharti karwaya*
Translation: You never went to school yourself; that's why you had to beg to enroll me.

Model	True Class			Predicted Class		
	Toxic	Severity	Sentiment	Toxic	Severity	Sentiment
baseline	No	0	Negative	Yes	1	Negative
MMLLM$_{2M}$				Yes	1	Negative
MMLLM$_{3M}$				Yes	1	Negative

(c)

Utterance: Tumhare dada ye jb lene jate hai toh sidhi lagti hogi ya khud hi chadh jate hai
Translation: When your grandfather goes to take it, does he stand straight, or does he climb up by himself?

Model	True Class			Predicted Class		
	Toxic	Severity	Sentiment	Toxic	Severity	Sentiment
baseline	Yes	1	Negative	No	0	Neutral
MMLLM$_{ST}$				No	0	Neutral
MMLLM$_{MT}$				Yes	1	Negative

Figure 4: Human annotation Vs. model's prediction for qualitative analysis with different settings

B Encoder Description

Audio Encoders: We conducted experiments using two state-of-the-art (SOTA) models, namely Whisper Radford et al. (2023) and MMS(Pratap et al., 2023), to encode audio signals and extract meaningful representations from the audio data. In our investigation, Whisper consistently demonstrated superior performance across all experimental settings compared to MMS. The reasons behind Whisper's superiority could be manifold. It might possess a more refined architecture tailored for audio processing, incorporating domain-specific optimizations or leveraging advanced techniques such as self-attention mechanisms or convolutional layers optimized for audio data. Additionally, the training procedure, hyperparameter settings, or data preprocessing techniques employed for Whisper could contribute to its superior performance over MMS.

(i) Massively Multilingual Speech (MMS) (Pratap et al., 2023): Developed by Facebook AI, MMS is a comprehensive multilingual pre-trained model for speech. It undergoes pretraining with Wav2Vec2's self-supervised training objective on an extensive dataset comprising approximately 500,000 hours of speech data across more than 1,400 languages.

(ii) Whisper: Radford et al. (2023) introduced Whisper, a novel multilingual speech recognition model. Whisper is trained on an extensive audio dataset, incorporating weak supervision for improved performance.

Video Encoders: We explored the efficacy of two transformer-based video models equipped with spatiotemporal context by uniformly sampling 16 frames from each video clip and feeding them into these encoders. Our experiments focused on comparing the performance of VideoMAE Tong et al. (2022) and TimeSformer (Bertasius et al., 2021), with VideoMAE consistently outperforming TimeSformer across all settings.

Despite the advanced design and capabilities of TimeSformer, our experimental results indicate that VideoMAE consistently outperforms it across all evaluated settings. The reasons behind this superiority could stem from various factors such as the efficiency of VideoMAE's learning approach, its ability to capture subtle temporal dependencies, or the effectiveness of its feature representation in downstream tasks. Additionally, factors like model architecture, training strategies, and hyperparameter settings may also influence the comparative performance of these models.

(i) VideoMAE: Tong et al. (2022) introduced a data-efficient learning approach for self-supervised video pre-training. It utilizes Masked Autoencoders to efficiently learn representations from video data, demonstrating effectiveness in enhancing model performance.

(ii) TimeSformer: (Bertasius et al., 2021) This model is a novel architecture designed for video understanding tasks. It extends the Transformer architecture to capture temporal relationships in videos by incorporating a spatiotemporal attention mechanism, demonstrating state-of-the-art performance in various video analysis applications.

Text Encoders: (Nayak and Joshi, 2022) introduces several transformer-based models pretrained on L3Cube-HingCorpus, the first large-scale real Hindi-English code-mixed dataset in Roman script: HingBERT, HingMBERT, HingRoBERTa, and HingGPT. These models, evaluated on tasks like sentiment analysis, POS tagging, NER, and language identification, demonstrate the effectiveness of real code-mixed data. They also present HingBERT-LID for language identification and HingFT for code-mixed word embeddings, making all resources publicly available for further research in Hinglish NLP.

(i) HingBERT: HingBERT is a BERT-based model pre-trained on the Hinglish corpus using masked language modeling objectives. It is evaluated on downstream tasks such as code-mixed sentiment analysis, POS tagging, NER, and language identification (LID) from the GLUECoS benchmark.

(ii) HingMBERT: HingMBERT is a variant of the multi-lingual BERT model pre-trained on the Hinglish corpus. It is trained using both Roman and Devanagari scripts and is assessed on various downstream tasks including code-mixed sentiment analysis, POS tagging, NER, and LID.

(iii) HingRoBERTa: HingRoBERTa is a RoBERTa-based model trained on the Hinglish corpus. It has versions trained on Roman script and a combination of Roman + Devanagari scripts. The model is evaluated on downstream tasks such as code-mixed sentiment analysis, POS tagging, NER, and LID.

(iv) HingGPT: HingGPT is a generative transformer model based on the GPT-2 architecture. It is trained on the Hinglish corpus to generate full tweets in code-mixed Hinglish.

We initially utilized all four pre-trained models for the unimodal baselines and filtered the top 2 models namely HingGPT and HingRoBERTa for the rest of the experiments. According to the results obtained from our main framework, HingRoBERTa outperforms HingGPT in all the corresponding settings, proving it to be the best candidate for our framework among all the considered models.