BanVATLLM and BanTSS: A Multimodal Framework and a Dataset for Detecting Toxic Speech in Bangla and Bangla-English Videos

Mohammad Shariful Islam
Computer Science & Telecommunication Engineering, Noakhali Science & Technology University, Bangladesh
shariful.ces43@gmail.com

Mohammad Abu Tareq Rony
Department of Statistics, Noakhali Science & Technology University, Bangladesh
abutareqrony@gmail.com

Abstract

The rise of video content on social media has led to toxic speech spread, necessitating effective moderation. This study addresses detecting toxic speech in Bangla and Bangla-English videos using multimodal data and deep learning techniques. The BanTSS dataset, with 431 videos and 2021 annotated utterances, supports this research. We propose the BanVATLLM framework, a multimodal architecture integrating audio, video, and text data. Utilizing advanced models like Whisper, MMS, VideoMAE, Timesformer, and ChatGPT-3.5, BanVATLLM shows high accuracy in classifying toxicity, sentiment, and severity, with high F1-scores and Fleiss' Kappa scores. Results include 95.78% F1 and 95.72% accuracy for toxicity, 88.27% F1 and 88.55% accuracy for severity, and 84.85% F1 and 83.86% accuracy for sentiment, enhancing detection in low-resource languages.

1 Introduction

In today's digital age, social media platforms empower users to create content, leading to increased information sharing. The rise of video content has transformed information dissemination, bringing challenges like harmful content and toxic speech. Effective moderation, whether human or AI-driven, is crucial to mitigate these risks. By 2023, video content is projected to account for 82% of internet traffic, highlighting the influence of platforms like YouTube and Dailymotion on public discourse (Wilson, 2022). YouTube[1] users watch over a billion hours of video daily. The viral nature of video content facilitates rapid news dissemination but also accelerates toxic speech spread. Toxic speech, defined by (Dixon et al., 2018) as "discourteous, disrespectful, or unreasonable" language, can drive individuals from discussions. Although much content is harmless, a significant portion violates guidelines and promotes harmful narratives (O'Connor, 2021). Failure to remove toxic content can create hostile online environments and echo chambers, leading to financial losses, fines, and legal issues for platforms[2]. Human moderators face emotional and psychological trauma due to the content they review.

1.1 Contributions

This paper presents a comprehensive approach for detecting Bangla and Banglish toxic speech in videos using multimodal data and deep learning techniques. The key contributions are:

1. **BanTSS Dataset**: A dataset from YouTube with 2021 code-mixed utterances annotated for Toxicity, Sentiment, and Severity, fostering research in low-resource languages.

2. **BanVATLLM Framework**: A multimodal multitask framework for toxic video detection, sentiment analysis, and severity assessment, featuring advanced modules for synchronization and modality fusion.

2 Related Work

(Wu and Bhandary, 2020) only used textual features from video transcripts, whereas (Rana and Jha, 2022) combined text and audio features for offensive video detection. This study, however, faced challenges with dataset accessibility, insufficient data curation, and annotation processes. This study (Pannerselvam et al., 2024) examines the effectiveness of the Sentence Transfer Fine-tuning (Set-Fit) method with logistic regression for detecting offensive content in Tamil-English code-mixed language, outperforming five NLP models.

[1] https://www.youtube.com/

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

These results highlight Set-Fit's potential for enhancing NLP systems in code-mixed language contexts. (Maity et al., 2024a) introduced ToxCMM, a benchmark dataset for code-mixed videos, and proposed ToxVidLLM, a multi-modal framework combining text, audio, and video, achieving significant results.(Das et al., 2023) offered a comprehensive approach by integrating text, image, and audio for hate video detection in English. Our study is the first to introduce a multi-modal toxic video dataset in low-resource code-mixed languages, particularly Bangla, with sentence-level annotations. The dataset and standard models will assist content moderators in detecting harmful content more effectively while minimizing false positives.

3 Multimodal (BanTSS) Dataset Creation

We collected 431 Bangla and Bangla-English videos from YouTube, transcribing them with Whisper (Radford et al., 2023). Three graduates and three bilingual undergraduates annotated 2021 samples for toxicity, severity, and sentiment, refining guidelines through discussion. Fleiss' Kappa scores indicated good annotation quality: 0.81 for toxicity, 0.72 for sentiment, and 0.76 for severity. The BanTSS dataset includes 1324 non-toxic and 697 toxic utterances.

4 Methodology

The BanVATLLM architecture classifies Bangla and Banglish video content by detecting offensive content and assigning sentiment and severity labels. Each video U consists of frames R, audio B, and transcript S. The classifier D: $D(S; R; B) \rightarrow z$ outputs label z. The BanVATLLM framework (Figure 1) includes an Encoder, Cross-Modal Synchronization Module, and Multitask module (Maity et al., 2024b).The Audio Encoder uses Whisper (Radford et al., 2023) and MMS (Pratap et al., 2023), the Video Encoder uses VideoMAE (Wang et al., 2023) and Timesformer (Bertasius et al., 2021), and the Text Encoder uses ChatGPT-3.5 (Niyogi and Bhattacharya, 2024).

Features from audio, video, and text are integrated, prioritizing text. Video and audio are encoded using pre-trained VideoMAE and Whisper models:

$$Y_v = \text{VideoMAE}(W) \quad (1)$$
$$Y_a = \text{Whisper}(X) \quad (2)$$

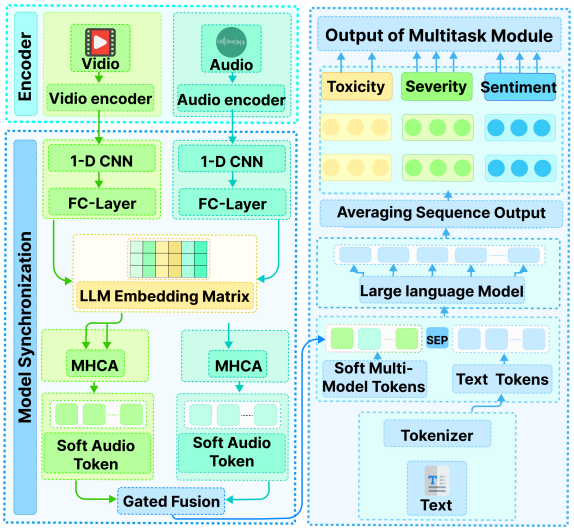

Figure 1: Proposed BanVATLLM model's Architecture

Compressed features are obtained through 1-D convolution and linear layers:

$$F_v = FC(\text{Conv}(Y_v)) \quad (3)$$
$$F_a = FC(\text{Conv}(Y_a)) \quad (4)$$

Multi-Head Cross-Attention (MHCA) aligns audio and visual features with text embeddings:

$$F_v^s = MHCA(F_v, E_t, E_t) \quad (5)$$
$$F_a^s = MHCA(F_a, E_t, E_t) \quad (6)$$

Synchronized representations F_v^s and F_a^s act as soft tokens for the LLM.

A gated fusion strategy merges video and audio tokens:

$$\beta = \sigma(Q_v F_v^s + Q_a F_a^s + b_g) \quad (7)$$
$$H_{va} = \beta F_a^s + (1 - \beta) F_v^s \quad (8)$$

The LLM processes the transcript T, generating a text token embedding E_t. Multimodal soft tokens J_m are appended to text tokens, creating a comprehensive input for the LLM. The output sequence undergoes averaging and task-specific layers for toxicity, severity, and sentiment detection.

The final loss function is a weighted sum of individual task-specific losses:

$$\text{Loss}_f = \sum_{k=1}^{M} \beta_k \text{Loss}_k \quad (9)$$

where β_i are learned end-to-end, reflecting task importance.

5 Experimental Result Analysis

5.1 Experimental Setups

.

The experiments of the study were performed on Google Colab Pro+, leveraging a cloud-based setup. The environment featured an NVIDIA A100 GPU with 40 GB VRAM, 52 GB RAM, and an Intel Xeon CPU (Abu Tareq Rony et al., 2024). The dataset was divided into training, validation, and testing sets in a ratio of 8:1:1. Models were trained ten times with different random splits.For implementing the experiments we used the PyTorch[3] framework.

5.2 Result Analysis

We evaluated various text, video, and audio encoders to determine the most effective models for detecting toxic content in videos. Table 1 (see Appendix) highlights the superior performance of the GPT3.5 text encoder, the VideoMAE video encoder, and the Whisper audio encoder. The results show that combining multiple modalities and leveraging multitask learning significantly enhances model performance for toxicity, severity, and sentiment detection in videos. Our proposed multimodal framework, using the best-performing encoders (GPT3.5, VM, and WP), outperforms single-modality models, demonstrating the advantages of a comprehensive approach to video content analysis.

5.2.1 Performance comparison across different encoders and modalities for toxicity, severity, and sentiment tasks.

Single-Modality, Single-Task Output: Text modality outperforms video and audio in all tasks. GPT3.5 for text achieves 87.10% F1 and 87.14% accuracy in toxicity detection, while VideoMAE and Whisper reach 70.16% and 77.60% accuracy for video and audio. These findings guided the selection of the VM and WP encoders. Similar trends are observed in severity and sentiment analysis.

Multiple-Modality, Single-Task Output: Combining text and video (T+V) with GPT3.5 and VM improves performance, achieving 91.58% F1 and 91.52% accuracy in toxicity detection. Adding audio (T+V+A) with WP further enhances results to 93.57% F1 and 93.41% accuracy.

Multi-task Output (Toxicity, Severity, Sentiment): Combining text, video, and audio (T+V+A) with GPT3.5, VM, and WP in a multitask setting yields 95.15% F1 and 95.06% accuracy for toxicity detection and excels in severity and sentiment analysis.

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

Table 1: Performance comparison across different encoders and modalities for toxicity, severity, and sentiment tasks.

Modality	Encoder	Toxicity		Severity		Sentiment	
		F1	Acc	F1	Acc	F1	Acc
Single-Modality, Single-Task output							
V	TF	67.66	67.90	61.81	62.85	58.55	59.67
V	VM	70.10	70.16	62.85	66.30	61.20	61.75
A	MMS	77.24	77.32	68.64	68.78	65.56	65.51
A	WP	77.55	77.60	70.40	69.42	67.15	67.24
T	GPT3.5	87.10	87.14	77.76	77.36	74.54	74.15
Multiple-Modality, Single-Task output							
T+V	GPT3.5 + VM	91.58	91.52	84.78	84.42	77.07	76.69
T+A	GPT3.5 + WP	92.70	92.49	85.89	86.49	78.04	77.62
T+V+A	GPT3.5 + VM + WP	93.57	93.41	86.29	86.35	79.57	79.66
Multi-task output (Toxic + Severity + Sentiment)							
T+V	GPT3.5 + VM + WP	95.49	95.37	84.79	84.55	84.48	83.61
T+A	GPT3.5 + WP	94.24	94.02	87.40	86.55	79.62	78.81
T+V+A	GPT3.5 + VM + WP	95.15	95.06	87.74	87.82	81.46	81.64

5.2.2 BanVATLLM's modification on Multitask

We studied the impact of Gated Fusion (GF) and Multi-Head Cross-Attention (MHCA) modifications on toxicity, severity, and sentiment tasks (see Table 2). Removing the GF module from BanVATLLM decreased F1-scores by 1.72% for toxicity, 2.73% for severity, and 3.13% for sentiment, highlighting GF's role in integrating video and audio modalities. Removing MHCA resulted in F1-score drops of 4.50%, 6.12%, and 7.11% for toxicity, severity, and sentiment, respectively, underscoring MHCA's importance in aligning text-guided audio and video features. Simultaneous removal of both GF and MHCA led to a 7.03% decline in toxicity F1-score, with similar drops in other tasks, emphasizing the critical role of cross-modal synchronization.

Table 2: Performance metrics for different model modifications on Multi-tasks output

Modality	Toxicity		Severity		Sentiment	
	F1	Acc	F1	Acc	F1	Acc
BanVATLLM	**95.78**	**95.72**	**88.27**	**88.55**	**84.85**	**83.86**
- GF	94.06	94.27	85.54	86.02	81.72	81.97
	-1.72%	-1.45%	-2.73%	-2.53%	-3.13%	-2.25%
- MHCA	91.28	91.33	82.87	82.98	78.81	79.05
	-4.50%	-4.39%	-6.12%	-6.29%	-7.11%	-5.81%
- MHCA - GF	89.05	89.19	79.65	79.87	76.70	76.93
	-7.03%	-6.83%	-9.75%	-9.77%	-9.59%	-8.28%

6 Conclusion

In today's ever-changing online environment, where videos have become the primary type of content, identifying toxic content, particularly in low-resource code-mixed languages, is more critical than ever. We present "BanTSS," an innovative benchmark dataset that contains code-mixed videos for detecting toxic content. Our proposed advanced multimodal framework, "BanVATLLM," based on LLM, delivered impressive results, highlighting the importance of integrating text, audio, and video modalities. Among the individual modalities, transformer encodings of text are particularly effective in identifying toxic videos. In addition to toxicity, the "BanTSS" dataset includes two additional categorizations: sentiment and severity, providing a thorough resource for further research in sentiment analysis within low-resource code-mixed videos.

6.1 Limitations

The study did not account for the context of the video clips, analyzing single utterances in isolation. Future work will incorporate entire video clips to capture context better. Our study focused on explicit toxic markers, excluding implicit or indirect toxic expressions. Future research will aim to develop datasets and models to detect these subtler forms of toxicity. The BanVATLLM model requires significant GPU memory for training due to its fine-tuning of two encoder modules, an LLM, and additional modules. Computational constraints prevented us from experimenting with parameter-efficient fine-tuning methods (PEFT) like LoRA or Quantized-LoRA for large models.

Ethics Statement

This study utilized public YouTube videos, ensuring compliance with terms of service and anonymity. Annotators received training and support to handle sensitive content. Bias mitigation included guideline reviews and majority voting. Our aim is to enhance toxic speech detection ethically, promoting safer online spaces without infringing on free speech.

Acknowledgement

We extend our gratitude to Noakhali Science and Technology University (NSTU) for funding the dataset curation process. We also recognize and appreciate the hard work and dedication of the data annotators and collectors. Their support was crucial in making this work possible.

References

Mohammad Abu Tareq Rony, Mohammad Shariful Islam, Tipu Sultan, Samah Alshathri, and Walid El-Shafai. 2024. Medigpt: Exploring potentials of conventional and large language models on medical data. *IEEE Access*, 12:103473–103487.

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

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.

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.

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

Krishanu Maity, AS Poornash, Sriparna Saha, and Pushpak Bhattacharyya. 2024b. Toxvidllm: A multimodal llm-based framework for toxicity detection in code-mixed videos. *arXiv preprint arXiv:2405.20628*.

Mitodru Niyogi and Arnab Bhattacharya. 2024. Paramanu: A family of novel efficient indic generative foundation language models. *arXiv preprint arXiv:2401.18034*.

J. O'Connor. 2021. Building greater transparency and accountability with the violative view rate — blog.youtube. Accessed 28-10-2023.

Kathiravan Pannerselvam, Saranya Rajiakodi, Sajeetha Thavareesan, Sathiyaraj Thangasamy, and Kishore Ponnusamy. 2024. Setfit: A robust approach for offensive content detection in tamil-english code-mixed conversations using sentence transfer fine-tuning. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 35–42.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, and 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*.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560.

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.