Culturally Aware Content Moderation for Facebook Reels: A Cross-Modal Attention-Based Fusion Model for Bengali Code-Mixed Data

Momtazul Arefin Labib, Samia Rahman, Hasan Murad

Department of Computer Science and Engineering

Chittagong University of Engineering and Technology, Bangladesh

Objectives

- To develop a system that can classify adult, harmful and suicidal facebook reels containing Bengali code-mixed language.
- To develop a multi-modal facebook reel dataset labeled as safe, adult, suicidal and harmful.

Introduction

- Rise of video content and reels in social media
- Challenges of harmful reels
- Lack of culturally-aware frameworks for Bengali harmful content
- Creation of Novel Facebook Reels dataset, and multimodal classification framework

Motivation

- Preserving digital safe environment for all kinds of people
- Protecting young generation form unsafe reels

Methodology

We have proposed a Multimodal Attention-based Gated Fusion architecture for this classification problem. First we have evaluated the models for each modality (shown in Fig 2) and used the best performing model from each modality to use in a cross modal attended gated fusion system using summation method (in Fig 3). The gated fusion output has been sent to an FC layer to finally output the predictions.

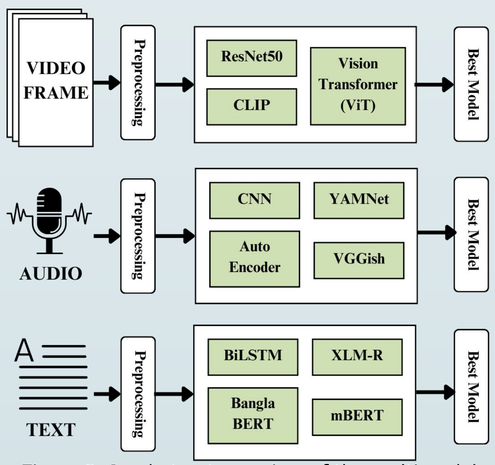

Figure 3: An abstract overview of the multimodal evaluation system of the UNBER dataset

Main Contribution

Developed UNBER, a multi-modal Bengali codemixed unsafe reels dataset, a unique annotation management tool "ReelAn" and a cross-modal attention-based culturally-aware framework enabling fusion techniques.

UNBER : A New Benchmark Dataset

"UNBER" contains 447 Safe, 327 Adult, 221 Harmful, and 122 Suicidal reels in Bengali, English and Banglish code-mixed language, in total 1,111 facebook reels. These reels have been annotated properly with a mean kappa score of 0.821 and for storage efficiency, audios, videos and visual texts has been kept separately after some basic processing

Category	Bengali	English	Banglish
Safe	5040	198	2049
Adult	3067	52	676
Harmful	2541	66	378
Suicidal	1482	55	161

Table 1: Word Distribution of Bengali, English, and Banglish Words

কাপুর টেনে আর লাভ নাই জানালার পর্দা খুলা আসলে এদের কোনো লজ্জা সরম নাই (There is no use in pulling the clothes, the window curtain is open, in fact they have no shame)

মাথা পুরাই নষ্ট মামা (It's crazy, man!!)

হে আল্লাহ, এই দুনিয়ার মানুষের কাছে হারতে হারতে ক্লান্ত আমি, আর কত আঘাত দিবা, এখন অন্তত তোমার কাছে নিয়ে যাও (O Allah, I am tired of losing to the people of this world, how much I will hurt, now at least take me to you)

Figure 1: Example of some unsafe reels found in social media.

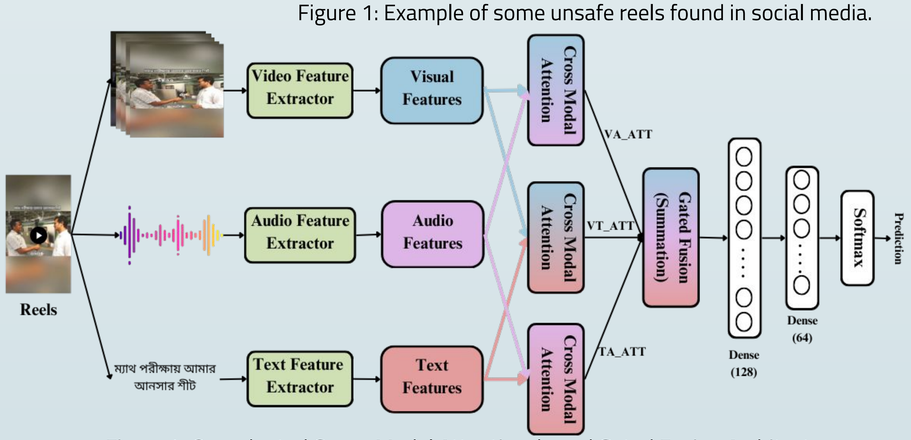

Figure 4: Our adopted Cross-Modal Attention-based Gated Fusion Architecture.

Results

Here, XLM-R, BB, AE, ViT, P, R, F1 representes XLM-Roberta, BanglaBERT, AutoEncoder, Vision Transformer, Precision, Recall, F1-Score.

Table 2 shows that from Text modality, BB outperforms other models. Among the Audio modality and Visual modality, AE and ViT outperformed respectively. The fusion of ViT, BB and AE has performed with two fusion model Early and Gated, where Gated has scored the best F1 score.

	Models	P	R	F1
Text	mBERT	0.55	0.52	0.52
	XLM-R	0.39	0.44	0.41
	BB	**0.58**	**0.58**	**0.55**
Audio	CNN	0.15	0.26	0.12
	AE	**0.41**	**0.41**	**0.41**
	VGGish	0.18	0.26	0.19
	YAMNet	0.10	0.25	0.14
Visual	ResNet50	0.59	0.49	0.51
	ViT	0.59	**0.56**	**0.57**
	CLIP	0.59	0.53	0.56
Fusion of ViT+BB+AE	Early	0.69	0.69	0.69
	Gated	**0.78**	**0.74**	**0.75**

Table 2: Precision (P), Recall (R), and F1-Score (F1) of Different Models

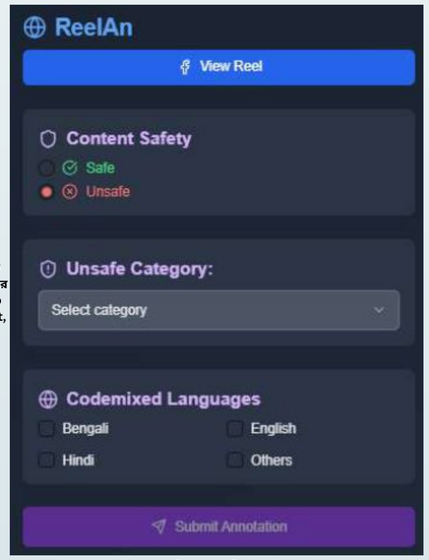

Figure 5: Interface of ReelAn Annotation Tools.

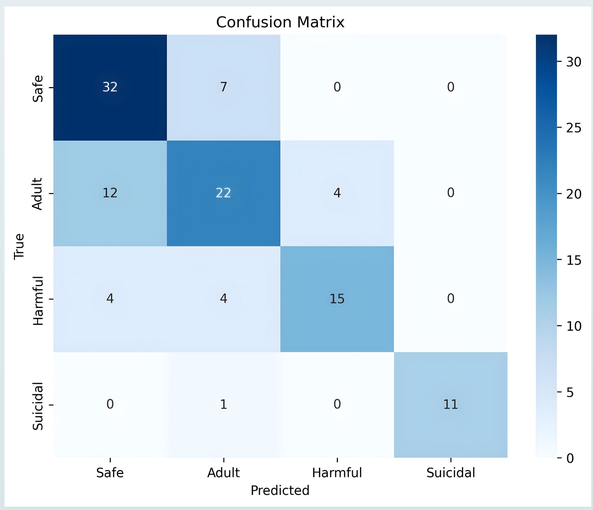

Figure 6: Confusion matrix of the performance of the Gated Fusion (ViT+AE+BB) with Summation model.

Conclusion

- Fusion of ViT+BB+AE has outperformed all other model with F1 score of 0.75
- Error analysis shows the irrelevancy of audio contents and small sized texts.
- Quick frame switching nature and dynamicity of contents impacted a lot.