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
{u1904111, u1904022}@student.cuet.ac.bd, hasanmurad@cuet.ac.bd

Abstract

The advancement of high-speed internet and affordable bandwidth has led to a significant increase in video content and has brought challenges in content moderation due to the spread of unsafe or harmful narratives quickly. The rise of short-form videos like "Reels", which is easy to create and consume, has intensified these challenges even more. In case of Bengali culture-specific content, the existing content moderation system struggles. To tackle these challenges within the culture-specific Bengali codemixed domain, this paper introduces "UNBER" a novel dataset of 1,111 multimodal Bengali codemixed Facebook Reels categorized into four classes: Safe, Adult, Harmful, and Suicidal. Our contribution also involves the development of a unique annotation tool "ReelAn" to enable an efficient annotation process of reels. While many existing content moderation techniques have focused on resource-rich or monolingual languages, approaches for multimodal datasets in Bengali are rare. To fill this gap, we propose a culturally aware cross-modal attention-based fusion framework to enhance the analysis of these fast-paced videos, which achieved a macro F1 score of 0.75. Our contributions aim to significantly advance multimodal content moderation and lay the groundwork for future research in this area.

1 Introduction

In recent years, there has been a rapid development in web users and sufficient bandwidth. Internet connectivity, being very cheap, makes the sharing of information such as text, audio, and video more common and faster. Video is most popular among them. By 2025, it is estimated that 82% of internet traffic will be video content[1]. For both entertainment and information purposes, social media users across all age groups engage with videos. Short videos such as reels have gained massive popularity and currently dominate social media. Their growth has accelerated even more with 5G. Reels are small in duration but rich in content. They also have better delivery and higher engagement compared to text and images. Facebook Reels, a prime example, are short, engaging videos shared on Facebook. Typically lasting a few seconds to a minute, they allow users to enhance content with music, text overlays, filters, and visual effects.

However, a darker side exists—videos, reels that violate community guidelines and spread harmful narratives. The failure to remove toxic content can lead to hostile online environments, echo chambers of hateful users, revenue loss, fines, and legal issues. While human moderators are employed to filter such content, the sheer volume of user-generated posts poses a significant challenge, especially with 5.24 billion social media users worldwide[2]. Additionally, content moderation can take an emotional and psychological toll on moderators. Legal regulations further demand the rapid removal of harmful content, adding to the complexity of the issue.

Another issue that has recently drawn attention from researchers is that, social media content moderation should consider the cultural variations. A content which can be suitable for a culture but inappropriate for another culture. The "one-size-fits-all" approach for content moderation of social networks such as Facebook, Instagram, etc. has been criticized by (Gomes and Sultan, 2024). In their paper, they discovered that a unique community guideline often does not satisfy cultural expression when making decisions. They also find out that marginal communities often adapt to the platform's policies to evade moderation. The findings in this paper are indeed consistent with reality. In case of

[1] https://beverlyboy.com/video-marketing/2025-video-marketing-statistics-you-simply-cant-overlook/ (Accessed: 2025-02-12)

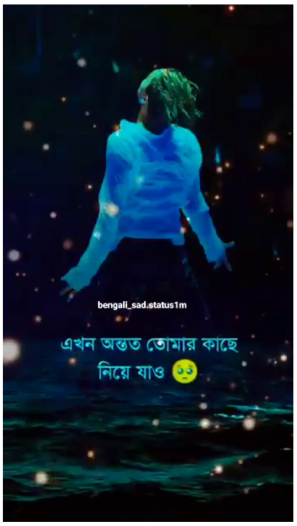
[2] https://backlinko.com/social-media-users (Accessed: 2025-02-12)

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)

class: Adult

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

class: Harmful



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)

class: Suicidal

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

Bengali culture, a lot of unsafe reels are available, which often do not go with the cultural standard of Bangladesh (Some examples have been shown in Figure 1. To eradicate the problem of cultural variation in content moderation, Chan et al., 2023 suggested enhancing content moderation by fine-tuning language models with culturally specific data.

Current research in harmful content detection is predominantly focused on text-based models (Das et al., 2022; Maity et al., 2023). There is, however, limited exploration in image-based methodologies (Kiela et al., 2020; Maity et al., 2022), and even fewer studies on video data (Das et al., 2023; Jha et al., 2024). These studies are mostly in monolingual English or high-resource languages. In the context of Bengali, a language spoken by 237 million native speakers[3], the exploration is particularly scarce in the multimodal domain. Notable works include (Hossain et al., 2022), which focuses on text and images of memes for detecting hateful content, and (Das et al., 2024), which deals with three modalities (text, audio, video) but solely for emotion classification. Additionally, Islam and Rony, 2024 explores toxic speech detection in code-mixed Bengali-English language across text, audio, and video domains by incorporating 431 videos from YouTube. However, this study processes individual utterances in isolation, ignoring the broader context within the video, and has quite a limited data set. The BanVATLLM framework also demands significant GPU resources for its multiple encoders (Whisper, VideoMAE, ChatGPT-3.5), making it impractical for real-time moderation or deployment on low-resource systems.

Our research involves introducing a novel dataset of 1,111 Facebook Reels, short informative videos, which include audio, visual content, and text overlays. We have also formulated and developed efficient, effective frameworks for contextual analysis of these videos. Since there is no suitable tool for video data annotation available, we have developed a tool to ease the process of annotation. It helps annotators focus on every piece of information, resulting in better-quality datasets. We have also proposed a multimodal multiclass classification framework for this dataset and classified the content into four categories: Safe, Adult, Harmful, and Suicidal. Our framework has achieved good performance and obtained the weighted average F1 score of 0.75. The major contributions include:

- Developed **UNBER**, a multi-modal Bengali codemixed unsafe reels dataset containing 1,111 multimodal data points, labeled into

[3]https://www.babbel.com/en/magazine/the-10-most-spoken-languages-in-the-world (Accessed: 2025-02-15)

four classes: *Safe*, *Adult*, *Harmful*, and *Suicidal*.

- Developed a unique annotation tool "ReelAn", efficient for annotating social media reels with multiple annotators.
- Developed a cross-modal attention-based culturally-aware framework enabling fusion techniques to perform on reels which are highly variable, have less relatedness between modalities, less detail, and often fast transitions.

2 Background Study

The widespread emergence of multimodal data has resulted in the development of multimodal deep-learning techniques. However, their development is severely lagging behind compared to the unimodal approaches.

2.1 Unimodal Unsafe Content Detection

Chowdhury et al., 2019 introduced an Arabic social network graph for hate speech detection trained on the dataset by Albadi et al., 2018, consisting of 6000 Arabic tweets. Banik and Rahman, 2019 addressed toxicity detection in the Bengali social media comments dataset. Ghosh et al., 2022 have used a hybrid deep learning approach on a public dataset by Romim et al., 2021 composed of 30,000 samples. Islam et al., 2024 proposed a CNN-BiLSTM model for hate speech classification of 5000 Banglish comments.

For audio-based detection, Rahut et al., 2020 classified abusive and non-abusive Bengali speech using spectrograms and VGG16 with an SVM classifier. Sankaran et al., 2024 explored cross-lingual abuse detection using the Whisper model, utilizing the ADIMA dataset by Gupta et al., 2021, comprised of 11,775 audio samples in 10 Indic languages, including Bengali, Hindi, and 8 more. The MuTox (Costa-jussà et al., 2024) dataset is a multilingual audio-based toxicity dataset consisting of 24,000 audio utterances from 30 languages, including English, Spanish, Arabic, and Bengali.

For video-based detection, Lopes et al., 2009 applied Bag-of-Visual-Features (BoVF) for obscenity detection on a collection of 179 videos. Ochoa et al., 2012 used Sequential Minimal Optimization for training an SVM (SMO) with a normalized polynomial kernel for adult content classification on 287 videos. Karpathy et al., 2014 applied CNNs for large-scale video classification, followed by Yue-Hei Ng et al., 2015, who used LSTM over frame-level CNN activations for improved video classification. CNN-LSTM models have been used for sequence modeling in multi-feature video classification models by Wu et al., 2015 and Wehrmann et al.. Other approaches include CNN-SVM (Aldahoul et al., 2021), CNN-BiLSTM (Yousaf and Nawaz, 2022), and attention-based CNN-BiLSTM (Yousaf and Nawaz, 2024). Transformer-based solutions have also been explored, such as TikGuard (Balat et al., 2024).

2.2 Multimodal Unsafe Content Detection

Multimodal unsafe content detection is less explored than its unimodal counterpart. (Kaushal et al., 2016) used supervised learning to detect child-unsafe content and content uploaders by training classifiers (random forest, K-nearest neighbor, and decision tree) with YouTube metadata (text+video). They applied bigram collocation and naïve Bayes for final classification. (Ngiam et al., 2011) pioneered deep learning in multimodal processing using restricted Boltzmann machines (RBM) on video, image, audio, and text. Some studies have explored multi-modal transformer-based approaches (Kiela et al., 2020). Bengali work in multimodal toxic/harmful content detection is quite unexplored. (Hossain et al., 2022) developed a Bengali text+image meme dataset for evaluation. (Islam and Rony, 2024) introduced the first Bengali dataset incorporating text, audio, and video for toxic content detection in Bengali and code-mixed Bengali-English.

Previous research on harmful content detection has predominantly focused on monolingual languages such as Portuguese (Alcântara et al., 2020), Thai (Maity et al., 2024), English (Rana and Jha, 2022), Bengali (Ghosh et al., 2022), Korean (Kim et al., 2024), Arabic (Chowdhury et al., 2019), Roman-Urdu (Rizwan et al., 2020), and Indonesian (Alfina et al., 2017). (Edstedt et al., 2022) has addressed multilingual harmful content detection, which covers 37 spoken languages, with English, French, Swedish, Spanish, and German being the most common.

2.3 Differences with existing research

In the domain of Bengali culture, datasets for detecting hateful, toxic, abusive, or harmful content in a multimodal setting are scarce. Most existing datasets are text-based, and there is a lack of re-

sources for short-video content analysis. While short-video datasets exist for unsafe content detection, such as TikGuard (Balat et al., 2024), none are available in Bengali.

Our study introduces the first multimodal Bengali unsafe content dataset for Facebook Reels, consisting of 1,111 videos categorized into four classes: Safe, Adult, Harmful, and Suicidal. The dataset incorporates text, audio, and video modalities in low-resource, code-mixed languages, combining Bangla and English. The dataset is annotated very carefully, making it a valuable resource for future research. Given its highly information-dense nature and the inclusion of three modalities, we believe it will significantly contribute to advancing unsafe content detection in Bengali culture-aware contexts.

3 UNBER: A New Benchmark Dataset

We have developed UNBER: a novel multimodal video dataset for Bengali-English Unsafe reel classification, which is firmly based on the Bengali culture. UNBER dataset contains short videos represented with their audio, visual, and text modality. For the text modality, only the texts that are visible in the short videos/reels have been considered. This section discusses about the creation, annotation, and analytics of UNBER.

3.1 Data Accumulation

Short videos or Reels are very much available nowadays due to their availability on most social media. Our primary data source for the reels collection has been Facebook. Our primary observation finds that short videos/reels are more likely to contain unsafe content rather than long videos on Facebook. To accumulate reels for our dataset, we have utilized an efficient extension from ESUIT, named as "Bulk Videos Downloader for Facebook"[4]. This tool helps to download all the collections of a specific profile or page in a very short amount of time. We have significantly focused on code-mixed language conversions in Bengali and English and avoided mixing Bengali and Hindi or any other language. Initially, our collection was 1615 reels. Later, we have retained 1,111 reels and filtered out the rest because of the code-mixing of Hindi or any other language except English with the Bengali language. In our consideration, the maximum allowable time duration of the reels in our dataset has been 300 seconds. Facebook assigns a unique value for each of the reels, called "reel_id". This "reel_id" has been used in our dataset as the key that distinguishes them from other reels.

[4]https://chromewebstore.google.com/detail/esuit-bulk-videos-downloa/bdoijmcmedjehajfdfcipjpjlcckkmce

3.2 Data Annotation

In UNBER, the collection of reels has been manually labelled into four distinct and predefined classes. They are Safe, Adult, Harmful and Suicidal. To ensure the quality of the dataset, it is required to follow a standard definition & cultural consideration for understanding the differences between the classes. We have studied and followed the categorization of unsuitable TikTok content by (Balat et al., 2024). The definition of our classes stands:

1. **Safe:** A reel is considered Safe if it is appropriate for the children and teenagers to view. This type of reel does not express negativity and often provides positive messages or emotions.

2. **Adult:** This type of reel contains content that can be explicitly sexual or implicitly convey obscene messages or emotions.

3. **Harmful:** Reels that contain violence or any kind of dangerous and risky actions that can influence children and teenagers to imitate. Some content in this type of reel can manipulate them negatively.

4. **Suicidal:** Ideation of suicide, discussing or implicitly expressing suicide, extremely sad and depressive reels fall into this category.

3.3 ReelAn: Our Annotation Tool

To make the annotation process simpler for our annotators, we have developed a website-based annotation tool "ReelAn" which has been built with NextJS, a React-based framework for full-stack. MongoDB has been used as the database for "ReelAn". All the collected reels "reel_id" have been uploaded to the database. "ReelAn" followed an efficient algorithm (illustrated in Algorithm 1) to effectively find and choose a reel randomly for the user when s/he enters the tools as an annotator. This algorithm ensures that all the reels in the database have been annotated an equal number of times and have equal importance regardless of how many annotators have been involved at a time,

and reduces the necessity of synchronization of the annotators. For example, if there are n reels, our tool ensures that no reels will be annotated twice unless all of the n reels have been annotated once. One significant corner case for this tool is, if two annotators enter the annotation page at the same time, they may receive the same reel and end up with that particular reel annotated twice. But the randomization at the end of the algorithm reduces the probability of two annotators getting the same reels.

Algorithm 1 Fetching Algorithm for ReelAn

Require: $Reels$ (list of reels with annotation counts)

Ensure: Returns a reel link with the least annotations

1: $Reels \rightarrow (reel\_id, count)$
2: $min\_count \leftarrow \min(r.count \mid r \in Reels)$
3: $candidates \leftarrow r \in Reels \mid r.count = min\_count$
4: $selected\_reel \leftarrow \text{random}(candidates)$
5: **return** $selected\_reel$

In the interface of "ReelAn", there is a button that takes the annotator to the "Facebook" to show the particular reel. After watching, the annotator chooses initially if the reel is Safe or Unsafe. If the "Unsafe" option has been chosen, another division shows up requiring the options for the "Unsafe" category. Another option has been added to manually evaluate the languages contained in that reel. This ensures the purity of our collections, which are in Bengali and English code-mixed and code-switched language only. Figure 2 shows the interface of our annotation tools.

"ReelAn" also have an admin panel, where the progress of the annotation can be tracked and the annotated labels can be downloaded as a JSON file.

3.4 Annotation Process

Annotators have followed predefined class definitions with cultural considerations and provided reasoning for their labels to get expert validation. Twelve independent annotators have annotated the dataset, and an NLP expert verified the labels. The expert resolved whenever there were disagreements. At first, annotators were provided with 100 samples. During their annotations, the conflict was resolved by providing high-level guidance from the NLP expert. After that, when the annotators became trained, they performed annotations on the rest of the 1,111 datasets. But still, some conflict occurred. An inter-annotator agreement has been measured using Cohen's Kappa score, with a mean kappa score of 0.821, indicating moderate agreement.

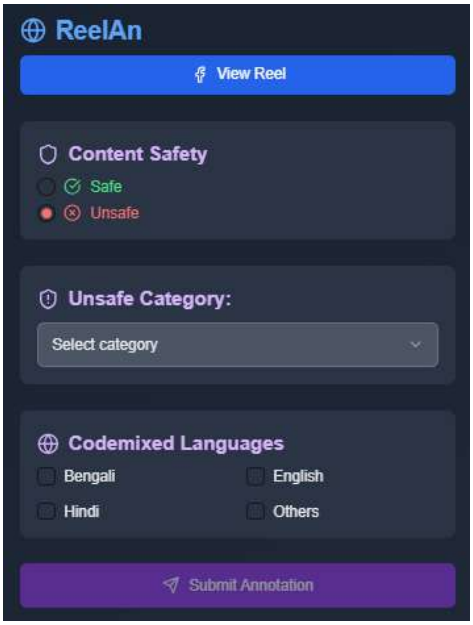


Figure 2: Interface of ReelAn Annotation Tools.

3.5 Dataset Statistics

UNBER contains 1,111 reels/short videos collected from Facebook. For efficient storage, instead of the reel videos, our dataset contains audio, videos, and text with basic processing from the original videos. In the dataset, all the audios are 5 seconds in length, ensuring truncation and padding. The dataset also contains the 5 extracted frames for all the reels. A CSV file contains the annotated visual texts and the label each reel has been assigned. "UNBER" contains 447 Safe, 327 Adult, 221 Harmful, and 122 Suicidal reels in Bengali, English and Banglish code-mixed language. Table 1 shows the distribution of the words of different languages among the classes.

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 Among Categories

A special feature of our dataset is, the short

videos in the dataset highly vary in their content dynamics. Some videos are slow-paced with little difference between the adjacent frames. On the other hand, some videos contain high transitions, fast moving, with high pixel differences between the frames. To prove this variation, we have run a statistical calculation on our dataset. For each subsequent frame in the video, the absolute difference with the previous frame is computed. Given two consecutive grayscale frames F_{t-1} and F_t, the absolute frame difference is calculated as:

$$D_t = |F_t - F_{t-1}|$$

where D_t represents the absolute difference image at time t, F_t and F_{t-1} are the grayscale intensity values of the current and previous frames, respectively. The mean pixel intensity of the difference image is computed as:

$$M_t = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} D_t(i,j) \quad (1)$$

where M_t is the mean intensity difference for frame t. H and W represent the height and width of the frame. $D_t(i,j)$ is the absolute difference value at pixel (i,j).

Finally, to obtain the average frame difference for the entire video, the mean difference values across all frames are averaged:

$$\bar{M} = \frac{1}{N} \sum_{t=1}^{N} M_t \quad (2)$$

where $\bar{M}$ is the overall average frame difference, N is the total number of frames in the video, M_t is the mean frame difference for frame t. The average frame difference $\bar{M}$ has been calculated for all the reels of our dataset, represented in Figure 3. This figure illustrates that a lot of video has average frame differences more than 15, while some contain 0 frame differences too, meaning no change in the content.

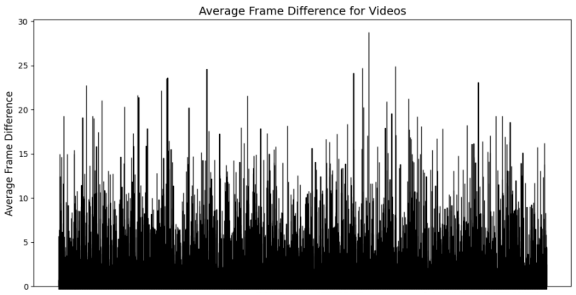


Figure 3: Average Frame Difference for all the reels of UNBER dataset.

4 Methodology

4.1 Problem Formulation

Our problem has been formulated as follows: A reel video will be provided as input, let the reel be denoted as R, our task is a classification problem. The target of this task is to determine whether R can be categorized as any of the four given classes. This categorization helps detect if any unsafe content is present in R. Three types of features have been extracted from each video. They are audio features, visual features and textual features. Let denote audio features as A, visual features as V and textual features as T. Each reel R has been expressed as a sequence of visual feature $V = \{v_1, v_2, ..., v_n\}$, a sequence of sampled audio features $A = \{a_1, a_2, ..., a_m\}$ and a sequence of words $T = \{t_1, t_2, ..., t_q\}$. Our Aim is to develop an efficient classifier $C_{reel}(V, A, T) \rightarrow p$ where p is the assigned category of V. We evaluate several deep learning and transformer-based models as C_{reel} on our dataset (Shown in Figure 4).

4.2 Text Modality

For the text modality of the reels, only the texts that appear on the reels have been extracted manually and further processed and analyzed.

4.2.1 Text Preprocessing

A good preprocessing of the textual part of UNBER has been ensured to maintain the consistency and the quality of the dataset. Stopword removal has been a crucial step in the preprocessing of UNBER, as it is a code-mixed dataset. We have fetched 398 Bengali stopwords from a GitHub source[5] and collected 48 code-switched stop words. In total, 446 stopwords have been used to preprocess the text portion of UNBER. Special characters have been removed using regex. Word tokenization has been applied to tokenize our dataset. All the words have been lowered, and only the words having more than 1 character have been chosen.

4.2.2 Text Feature Extraction

We have applied both word embedding and contextual embedding to extract textual features.

[5]https://raw.githubusercontent.com/stopwords-iso/stopwords-bn/master/stopwords-bn.txt

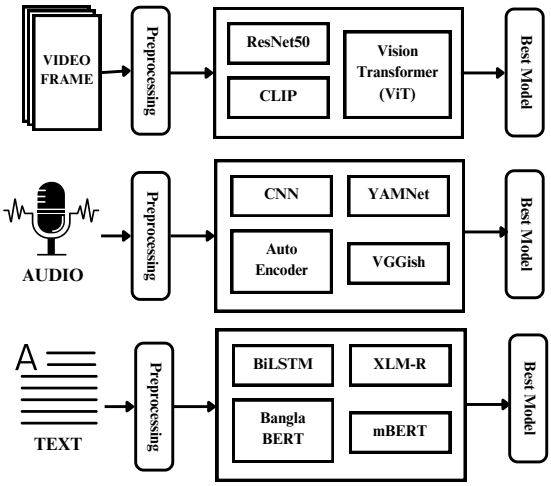


Figure 4: An abstract overview of the multimodal evaluation system of the UNBER dataset.

Our word embedding method consists of a word2vec embedding following a Bi_LSTM network. This deep feature extractor helps to identify semantic similarities between the words. Word2vec embeddings have two variants: skip-gram and continuous bag-of-words (CBOW). Skip-gram has been chosen for our model because of its efficiency and accurate representations. The window size has been chosen as 7, the embedding dimension as 100, the minimum word frequency set as 1, and the number of worker threads has been set to 4. Then, the average word embedding for a given sentence has been computed with the word2vec model. The Bi_LSTM sequential network consists of 2 Bidirectional layers and 2 Dense layers. Input shape was 100×100.

Contextual embeddings have been used for their efficiency in catching context-based features. We have utilized 3 context-based models to extract textual features from UNBER. They are a mBERT-based model "bert-base-multilingual-cased"[6], an XLM-Roberta-based model "xlm-roberta-base"[7] and a BanglaBERT model "csebuetnlp/banglabert"[8]. All these models have been fine-tuned on the text portions of UNBER, adjusting the learning rate, batch size and number of epochs.

[6]https://huggingface.co/google-bert/bert-base-multilingual-cased

[7]https://huggingface.co/FacebookAI/xlm-roberta-base

[8]https://huggingface.co/csebuetnlp/banglabert

4.3 Audio Modality

Audios from the reels have been extracted using the "moviepy" library of Python.

4.3.1 Audio Preprocessing

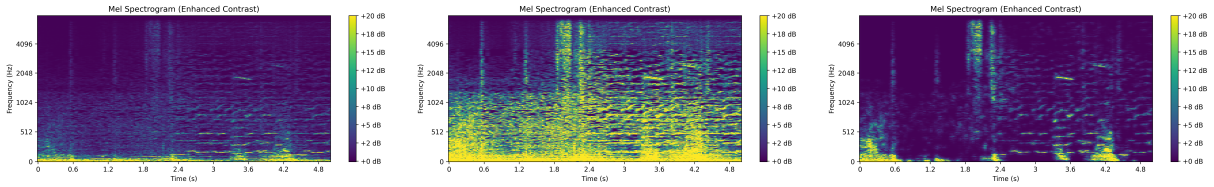
Extracted audios of the "UNBER" dataset have been preprocessed using "librosa", "noisereduce" and "soundfile" libraries of Python. Librosa is used to load audio files at a sampling rate of 22050. Then the audio has been trimmed or padded with silence to the target length of 5 seconds. After that, normalization has been applied to zero mean and unit variance. Noise reduction has been applied using spectral gating to enhance clarity.

4.3.2 Audio Feature Extraction

The preprocessed audio has been used to extract both hand-crafted features and deep features.

Several acoustic features have been extracted from the audio portion of the "UNBER" dataset, which we referred to as hand-crafted features. Mel-Frequency Cepstral Coefficients are one of the most used features in audio analysis. In this work, the coefficient value has been set to 13 to retrieve the features efficiently. MFCCs main advantage is that they can encode the way humans perceive sound, making them highly valuable for analyzing speech and music signals. Another feature of chroma has been used for its ability to analyze musical content in the audio. It denotes the 12 pitch classes energy distribution of the musical octave. The spectral centroid represents the centre of mass of the audio spectrum, where a higher value indicates brighter sounds. Spectral Contrast captures differences between peaks and valleys, which reflect the harmonic structure and timbre variations. The number of frequency bands used in the spectral contrast computation has been 6. The minimum frequency has been set to 200.0 Hz, which specifies the starting point of the frequency range. Frequencies below this value have not been included in the analysis. Spectrograms illustrate how frequency components change over time, providing a clear time-frequency visualization of the audio. Afterwards, all these features have been truncated and padded to a uniform length of 20. Finally, these features have been flattened and concatenated to build a standardized feature vector.

Some advanced deep-learning methods have been used to enhance hand-crafted features and extract deep features from the audio. These methods include CNN, Autoencoder, VGGish and YAMNet.



(a) (b) (c)

Figure 5: Spectrogram of an audio (a) initially, (b) after normalization and (c) after noise removal.

All these deep learning models have been used on all the extracted features: MFCC, Chroma, Spectral Centroid, Spectral Contrast and Spectrogram. A 5-layer Convolutional Neural Network (CNN) has been used with batch normalization and dropouts. We have also utilized an Autoencoder network, which compresses input data into a 32-dimensional representation using a dense encoder and reconstructs it through a decoder. The encoded features extracted from the encoder are then used as input for a classifier, which consists of two dense layers (64 and 32 units with ReLU activation) and a softmax output layer for multiclass classification. We have also used two pretrained deep learning models, YAMNet and VGGish, both developed by Google. YAMNet[9] is popular for being lightweight and efficient. VGGish[10] is based on VGG16, adapted for audio analysis. Both pre-trained models have been used to extract deep features, followed by a 3-layer classifier for the classification task.

4.4 Visual Modality

The visual features play the most crucial role in favour of a perfect classification. This section demonstrates the preprocessing and feature extraction steps performed on UNBER.

4.4.1 Video frame preprocessing

To analyze the visual features of our dataset, we have extracted 5 frames from each video. Though we have allowed reels with time lengths up to 300 seconds, selecting 5 frames is a trade-off between the precise analysis and efficient use of limited memory and processing resources. An efficient algorithm has been applied to ensure the variation of the frames. Initially, 5 distinct frames at regular intervals w have been selected with the formula $W = max(N/5, 1)$, where N is the number of total frames. Let the 5 frames at regular interval W be a_1, a_2, a_3, a_4, a_5. Afterwards, an iterative process checks the similarity between a_{i-1} with a range of frames from a_i to $(a_{i+1}-1)$ to find a frame most dissimilar from the previous one. If no frame is found in the iterative process whose similarity is below the predefined threshold, the $(a_{i+1} - 1)$th frame is finally selected as the ith frame. All the frames have been set to a uniform size (224×224), and the similarity threshold has been set to 0.9 to find a frame that has a good dissimilarity with the previous one.

4.4.2 Video frame feature extraction

To extract deep features from video frames, we have utilized a pretrained deep learning model, "ResNet50" and two transformer-based models: "Vision transformer" and "CLIP". ResNet50[11], or Residual Network with 50 layers, has been used mostly for its strong ability to extract features in images. Vision Transformers are famous for their own feature extractor, which breaks down images into patches and processes the patches further. We have used the "google/vit-base-patch16-224" model[12], utilizing "ViTFeatureExtractor" and "ViTModel" for feature extraction and model loading, respectively. CLIP[13] is a multimodal vision and language model used for its capability to analyze images with texts more efficiently. This model has been used with LSTM to catch the temporal dependencies between the frames.

4.5 Fusion

After performing feature extraction, the best 3 feature extractors have been chosen from the 3 modalities based on their performance. Fusion technique with Cross-Modal Attention has been implemented on the extracted features from these 3 best models.

First, features for audio, text, and video are extracted, producing three sets of feature vectors: $A \in R^{N\times512}$, $V \in R^{N\times256}$, $T \in R^{N\times768}$

[9] https://tfhub.dev/google/yamnet/1
[10] https://tfhub.dev/google/vggish/1
[11] https://huggingface.co/microsoft/resnet-50
[12] https://huggingface.co/google/vit-base-patch16-224
[13] https://huggingface.co/openai/clip-vit-base-patch32

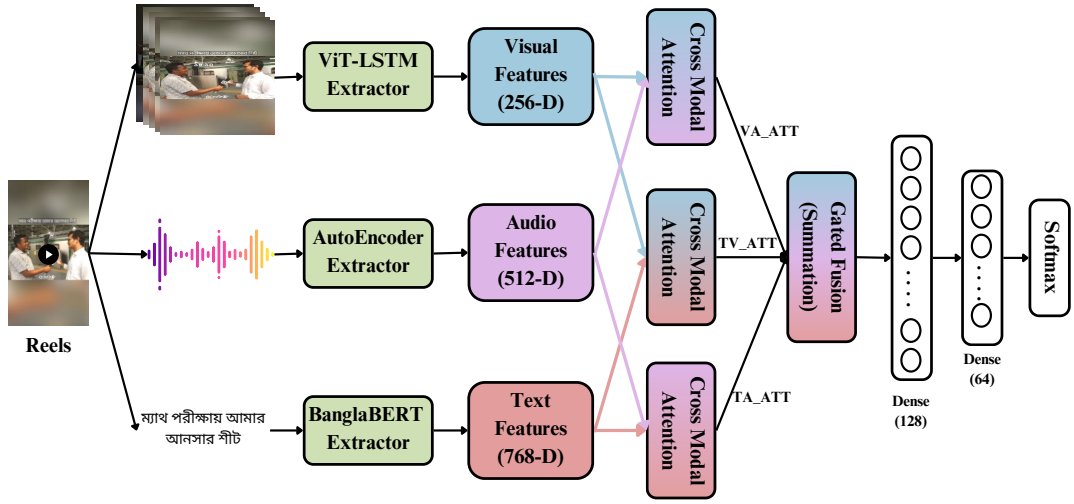


Figure 6: Our adopted Cross-Modal Attention-based Gated Fusion Architecture.

where N is the number of samples. The Cross-Modal attention has been applied to the features to get image-audio (VA_ATT), audio-text(AT_ATT), and text-image (TV_ATT) attended features. Later, A gated fusion (inspired by (Lyu et al., 2023)) has been applied to the cross-modal attended features with a project dimension of 512, where 3 gates have been used with sigmoid activation. The gates are multiplied with respective features and finally summed up to the gated features for the final fusion. For the early fusion, no gates have been used.

Afterwards, A Fully Connected (FC) Network has been used to build a model that can be trained on the fused features and make the predictions. The FC network consists of 2 hidden dense layers followed by batch normalization and dropout layers, with a final output layer to make prediction among the four classes. Figure 6 illustrates our adopted model, consisting with the best-performing unimodal models as the feature extractor.

5 UNBER: Benchmark Evaluation

This section provides a detailed discussion about the experimental setting, comparative results and the error analysis done by us to evaluate our dataset.

5.1 Experimental Setting

Our experimentation for UNBER has been conducted efficiently using an Intel Xeon CPU and Tesla P100-PCIE 16GB GPU. The dataset has been split into train, test, and validation with a ratio of 8:1:1 to ensure proper training and testing. Several configurations have been tested. In almost all cases, the best performance has been achieved using the Adam optimizer, Early stopping with patience 5, and reducing the learning rate on a plateau with patience 3. The per-device batch size has been 16 for the best-performing models. Weight decay has been set to 0.01, metric for best model evaluation set to "F1-score" while ensuring the load of the best model at the end. Models have been tested multiple times to determine the statistical significance of the differences between the performance of the models.

5.2 Results

Table 2 illustrates the performance of the unimodal and multimodal models on the "UNBER" dataset. This comparison clearly depicts that, among the text models, BanglaBERT has performed the best F1 score. AutoEncoder has performed best among the audio models. Vision Transformer (ViT) performed slightly better than CLIP, achieving the best performance among the visual models. Among the 2 fusion models combining the best 3 performed models, Gated Fusion with Summation has achieved an F1 score of 0.75, which outperformed all other models. Figure 7 shows the confusion matrix of the Gated Fusion model. There has been no scope to compare our result with any existing content moderation techniques or baseline methods to assess the relative effectiveness of the approach, because no such existing systems or methods have been found in Bengali Facebook reels content moderation.

	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

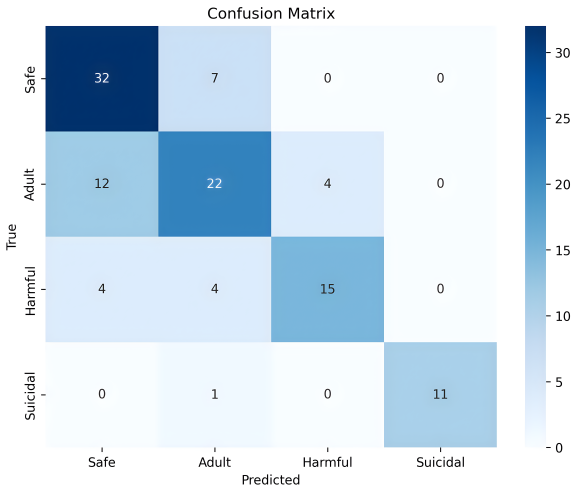


Figure 7: Confusion matrix of the performance of the Gated Fusion (ViT+AE+BB) with Summation model.

5.3 Error Analysis

The result shows that, among the unimodal models, Audio models have performed poor significantly. The reason behind these poor performances relies on the relevance of the audio used in the reel videos. Most of the reel videos used in social media platforms contain background music irrelevant to the original content of the audio. This drastically affects the models to distinguish between the classes. As a result, audio models struggled in the classification task. Text models performed moderately. The reason for the error of these models is that text overlays in the reels generally consist of one or two small sentences and contribute little to the content. Our unimodal visual models struggled due to the quick switching between frames and the dynamic nature of the contents of the reels. Also, our adopted best model, Gated Fusion of Vit, AutoEncoder, and BanglaBert model, outperformed all other models but struggled a little to classify between the safe and the adult contents. This is because the similarity of the adult and safe contents is quite high. In the previous studies, adult content has been determined based on the amount of skin revealed by the actors (Karamizadeh et al., 2023). But our annotation process also observed the use of slang and indirect obscene indications, which are a demand of Bengali culture, making it difficult for the models to distinguish between the adult and the safe contents.

6 Conclusion

Safe use of social media in the Bengali cultural context is a significant demand of a good cultural society. During this generation, where reel/short videos are on top of the trend with high manipulation power, an efficient, culturally-aware moderation process in every language is necessary to pull the reins of the decay of moral values. Our dataset "UNBER" is a crucial contribution to the Bengali Facebook Reel filtering process. This paper describes clearly the creation, annotation, and preparation process of this dataset and also shows an abstract methodology to evaluate the nature and efficiency of the dataset and to create a baseline for this dataset for later use. Also, the proposed framework is adaptable to larger datasets and can suit on different cultural contexts beyond Bengali, making it scalable.

6.1 Future work

In the future, we aim to expand the scope and utility of our dataset and model. Key enhancements will include increasing the size of the data in "UNBER" by adding reels/short videos from other social media such as Instagram, TikTok, and YouTube. Our future plan also includes extending the features of data by collecting comments and other metadata, such as like and dislike counts, for a more comprehensive analysis of the reels. There is a good scope for improving the performance of the model. Instead of adopting the combination of the best 3 models, checking several combinations of unimodal models can be more effective. Also, more cross-modal attended features like Video-Text (VT), Audio-Text(AT), etc can significantly contribute to the fusion model. Multimodal LLM can be used to achieve an excellent performance.

Acknowledgments

We express our heartfelt gratitude to the dataset annotators team, whose dedication, efforts and expertise has been crucial for creating and refining the dataset used in this study. Their commitment to ensuring high-quality annotations significantly contributed to the reliability and validity of our results. The authors gratefully acknowledge Centro Interuniversitario di Ricerca Scienze Umane e Sociali e Intelligenza Artificiale (ELIZA) – University of Naples 'L'Orientale' for its support in covering the registration costs, which enabled their participation.

References

Nuha Albadi, Maram Kurdi, and Shivakant Mishra. 2018. Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere. In *2018 IEEE/ACM International Conference on ASONAM*. IEEE.

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*.

Nouar Aldahoul, Hezerul Abdul Karim, Mohd Haris Lye Abdullah, Abdulaziz Saleh Ba Wazir, Mohammad Faizal Ahmad Fauzi, Myles Joshua Toledo Tan, Sarina Mansor, and Hor Sui Lyn. 2021. An evaluation of traditional and cnn-based feature descriptors for cartoon pornography detection. *IEEE Access*.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekanata. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *2017 ICACSIS*. IEEE.

Mazen Balat, Mahmoud Gabr, Hend Bakr, and Ahmed B Zaky. 2024. Tikguard: A deep learning transformer-based solution for detecting unsuitable tiktok content for kids. In *2024 6th NILES*. IEEE.

Nayan Banik and Md Hasan Hafizur Rahman. 2019. Toxicity detection on bengali social media comments using supervised models. In *2019 2nd ICIET*. IEEE.

Alex J Chan, José Luis Redondo García, Fabrizio Silvestri, Colm O'Donnell, and Konstantina Palla. 2023. Enhancing content moderation with culturally-aware models. *arXiv e-prints*.

Arijit Ghosh Chowdhury, Aniket Didolkar, Ramit Sawhney, and Rajiv Shah. 2019. Arhnet-leveraging community interaction for detection of religious hate speech in arabic. In *Proceedings of the 57th ACL SRW*.

Marta R Costa-jussà, Mariano Coria Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alex Mourachko, Christophe Ropers, and Carleigh Wood. 2024. Mutox: Universal multilingual audio-based toxicity dataset and zero-shot detector. *arXiv preprint arXiv:2401.05060*.

Avishek Das, Moumita Sen Sarma, Mohammed Moshiul Hoque, Nazmul Siddique, and M Ali Akber Dewan. 2024. Avater: A multimodal approach of recognizing emotion using cross-modal attention technique.

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*.

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*.

Johan Edstedt, Amanda Berg, Michael Felsberg, Johan Karlsson, Francisca Benavente, Anette Novak, and Gustav Grund Pihlgren. 2022. Vidharm: A clip based dataset for harmful content detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE.

Tapotosh Ghosh, Ashraf Alam Khan Chowdhury, Md Hasan Al Banna, Md Jaber Al Nahian, M Shamim Kaiser, and Mufti Mahmud. 2022. A hybrid deep learning approach to detect bangla social media hate speech. In *Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021*. Springer.

André Belchior Gomes and Aysel Sultan. 2024. Problematizing content moderation by social media platforms and its impact on digital harm reduction. *Harm Reduction Journal*.

Anirudh Gupta, Harveen Singh Chadha, Priyanshi Shah, Neeraj Chhimwal, Ankur Dhuriya, Rishabh Gaur, and Vivek Raghavan. 2021. Clsril-23: Cross lingual speech representations for indic languages. *arXiv preprint arXiv:2107.07402*.

Eftekhar Hossain, Omar Sharif, and Mohammed Moshiul Hoque. 2022. Mute: A multimodal dataset for detecting hateful memes. In *Proceedings of the 2nd conference of the asia-pacific chapter of the ACL and the 12th international joint conference on NLP: SRW*.

Md Hasibul Islam, Kaniz Farzana, Ibrahim Khalil, Shaneen Ara, Md Ruhul Amin Shazid, and Md Humaion Kabir Mehedi. 2024. Unmasking toxicity: A comprehensive analysis of hate speech detection in banglish. In *2024 6th ICEEICT*. IEEE.

Mohammad Shariful Islam and Mohammad Abu Tareq Rony. 2024. Banvatllm and bantss: A multimodal

framework and a dataset for detecting toxic speech in bangla and bangla-english videos. In *Eighth WiNLP 2024 Phase II*.

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. *arXiv preprint arXiv:2401.09899*.

Sasan Karamizadeh, Saman Shojae Chaeikar, and Alireza Jolfaei. 2023. Adult content image recognition by boltzmann machine limited and deep learning. *Evolutionary Intelligence*.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on CVPR*.

Rishabh Kaushal, Srishty Saha, Payal Bajaj, and Ponnurangam Kumaraguru. 2016. Kidstube: Detection, characterization and analysis of child unsafe content & promoters on youtube. In *2016 14th Annual Conference on Privacy, Security and Trust (PST)*. IEEE.

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*.

Minju Kim, Heuiyeen Yeen, and Myoung-Wan Koo. 2024. Towards context-based violence detection: A korean crime dialogue dataset. In *Findings of the Association for Computational Linguistics: EACL 2024*.

Ana Paula B Lopes, Sandra EF de Avila, Anderson NA Peixoto, Rodrigo S Oliveira, Marcelo de M Coelho, and Arnaldo de A Araújo. 2009. Nude detection in video using bag-of-visual-features. In *2009 XXII Brazilian Symposium on Computer Graphics and Image Processing*. IEEE.

Chenyang Lyu, Wenxi Li, Tianbo Ji, Liting Zhou, and Cathal Gurrin. 2023. Gated multi-modal fusion with cross-modal contrastive learning for video question answering. In *International Conference on Artificial Neural Networks*. Springer.

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 EMNLP*.

Krishanu Maity, Prince Jha, Sriparna Saha, and Pushpak Bhattacharyya. 2022. 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*.

Krishanu Maity, AS 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*.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. 2011. Multimodal deep learning. In *ICML*.

Victor M Torres Ochoa, Sule Yildirim Yayilgan, and Faouzi Alaya Cheikh. 2012. Adult video content detection using machine learning techniques. In *2012 Eighth International Conference on Signal Image Technology and Internet Based Systems*. IEEE.

Shantanu Kumar Rahut, Riffat Sharmin, and Ridma Tabassum. 2020. Bengali abusive speech classification: A transfer learning approach using vgg-16. In *2020 ETCCE*, pages 1–6. IEEE.

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

Hammad Rizwan, Muhammad Haroon Shakeel, and Asim Karim. 2020. Hate-speech and offensive language detection in roman urdu. In *Proceedings of the 2020 conference on EMNLP*.

Nauros Romim, Mosahed Ahmed, Hriteshwar Talukder, and Md Saiful Islam. 2021. Hate speech detection in the bengali language: A dataset and its baseline evaluation. In *Proceedings of IJCACI 2020*. Springer.

Aditya Narayan Sankaran, Reza Farahbaksh, and Noel Crespi. 2024. Towards cross-lingual audio abuse detection in low-resource settings with few-shot learning. *arXiv preprint arXiv:2412.01408*.

Jônatas Wehrmann, Gabriel S Simões, Rodrigo C Barros, and Victor F Cavalcante. Adult content detection in videos with convolutional and recurrent neural networks. *Neurocomputing*.

Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. 2015. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification. In *Proceedings of the 23rd ACM international conference on Multimedia*.

Kanwal Yousaf and Tabassam Nawaz. 2022. A deep learning-based approach for inappropriate content detection and classification of youtube videos. *IEEE Access*.

Kanwal Yousaf and Tabassam Nawaz. 2024. An attention mechanism-based cnn-bilstm classification model for detection of inappropriate content in cartoon videos. *Multimedia Tools and Applications*.

Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *Proceedings of the IEEE conference on CVPR*.