MTikGuard System: A Transformer-Based Multimodal System for Child-Safe Content Moderation on TikTok

Dat Thanh Nguyen[1,4,5], **Nguyen Hung Lam**[3,4,5],
Anh Thi-Hoang Nguyen[1,4,5], **Trong-Hop Do**[2,4,5]

[1]Faculty of Information Science and Engineering, University of Information Technology
[2]Faculty of Software Engineering, University of Information Technology
[3]Faculty of Computer Science, University of Information Technology
[4]University of Information Technology, Ho Chi Minh City, Vietnam
[5]Vietnam National University, Ho Chi Minh City, Vietnam

Emails: {22520224, 22520968}@gm.uit.edu.vn, {anhnth, hopdt}@uit.edu.vn

Corresponding author: Trong-Hop Do (hopdt@uit.edu.vn)

Abstract

With the rapid rise of short-form videos, TikTok has become one of the most influential platforms among children and teenagers, but also a source of harmful content that can affect their perception and behavior. Such content, often subtle or deceptive, challenges traditional moderation methods due to the massive volume and real-time nature of uploads. This paper presents **MTikGuard**, a real-time multimodal harmful content detection system for TikTok, with three key contributions: (1) an *extended* TikHarm dataset expanded to 4,723 labeled videos by adding diverse real-world samples, (2) a multimodal classification framework integrating visual, audio, and textual features to achieve **state-of-the-art** performance with 89.37% accuracy and 89.45% F1-score, and (3) a scalable streaming architecture built on Apache Kafka and Apache Spark for real-time deployment. The results demonstrate the effectiveness of combining dataset expansion, advanced multimodal fusion, and robust deployment for practical large-scale social media content moderation. The dataset is available at `https://github.com/ntdat-8324/MTikGuard-System.git`.

Disclaimer: This paper contains images and content from social networks that may be considered sensitive, harmful, or related to adult content and suicide.

1 Introduction

Harmful video content refers to videos containing visuals, audio, or messages that may negatively impact viewers' cognition, emotions, and behaviors - particularly those of children. Such content may involve depictions of violence, explicit language, antisocial behavior, hate speech, or dangerous challenges disguised as "fun" activities that children can easily imitate. In the era of short-form video as a dominant mode of entertainment, the spread of harmful content occurs faster than ever before, often beyond the supervision of parents or schools. Children, with limited analytical and self-protective capabilities, are especially vulnerable to these negative role models, often failing to recognize their long-term consequences for personality development and behavior. Consequently, the early detection and timely warning of harmful video content are essential to maintaining a safe online environment for young audiences.

Online platforms, particularly social media, enable harmful videos to proliferate rapidly, making control and prevention increasingly challenging. To address this issue, we build upon and extend the publicly available TikHarm dataset to improve model training effectiveness and develop a system that can curb the spread of harmful trends targeting children.

In this study, we expand the TikHarm dataset by collecting additional videos following the original selection criteria, thereby improving the model's learning capacity and robustness in real-world deployment. Leveraging this extended dataset, we introduce **MTikGuard**, *a real-time harmful content detection system for TikTok*. MTikGuard processes multiple modalities from videos, including visual frames and textual information (extracted via OCR and speech-to-text), enabling richer contextual understanding (see Appendix A.1 for the formal task definition). With a multimodal transformer-based architecture, MTikGuard outperforms single-modal baselines, particularly in detecting harmful trends presented in subtle or disguised forms. Our main contributions are summarized as follows:

- We expand the TikHarm dataset by collecting and annotating an additional 775 videos, resulting in a total of 4,723 labeled samples.

This extension improves model robustness and generalization.

- We propose a transformer-based multimodal architecture for harmful video detection that integrates visual frames, OCR-extracted text, and speech-to-text transcripts. By jointly modeling these modalities, the system captures nuanced cross-modal cues, enabling more accurate identification of harmful content, even when disguised or indirectly presented.

- We introduce the MTikGuard system, a real-time video analysis and detection pipeline for harmful content. Beyond its robust multimodal model, MTikGuard integrates advanced technologies such as Apache Spark, Apache Kafka, Docker, and Apache Airflow, enabling scalable, reliable deployment for detecting harmful content from online sources in production environments.

2 Related Work

Many studies have focused on detecting inappropriate content using only video data. (Singh et al., 2019) introduced the KidsGUARD system, which utilizes an LSTM-based autoencoder to learn video representations from features extracted by the VGG16 CNN network, aiming to detect unsafe content for children. The model achieved a recall of 81% and a precision of 80% on a dataset of 109,835 video clips. (Yousaf and Nawaz, 2022) proposed a method using EfficientNet-B7 for feature extraction and a BiLSTM network for video representation to classify inappropriate content in animated YouTube videos, achieving 95.66% accuracy and an F1-score of 92.67%. (Murthy et al., 2024) conducted a study applying computer vision techniques to detect e-cigarette-related content in TikTok videos, building an object detection model based on YOLOv7 that achieved a recall of 77%, precision of 86.30%, and F1-score of 81.40%. Transformer-based models such as TimeSformer (Bertasius et al., 2021), VideoMAE, and ViViT have also been effectively applied in video classification, enhancing accuracy and efficiency in detecting inappropriate content. Notably, TikGuard (Balat et al., 2024) employs advanced Transformer models like TimeSformer, VideoMAE, and ViViT to classify TikTok content, achieving 86.7% accuracy on the TikHarm dataset, demonstrating the potential of these models in detecting harmful content for children.

Recognizing the limitations of relying solely on video data, recent studies have shifted towards multimodal approaches, combining information from video, audio, and text to improve the detection of inappropriate content. (Phukan et al., 2025) proposed SNIFR, a new framework that leverages Transformers to integrate audio and visual information for detecting harmful content targeting children, outperforming single-modality methods. (Das et al., 2023) introduced HateMM, a multimodal dataset comprising approximately 43 hours of video from BitChute labeled as hate or non-hate, along with frame-level rationale annotations for label decisions. Their findings showed that incorporating all modalities significantly improved overall hate speech detection performance (accuracy = 79.80%, macro F1-score = 79%), with a gain of approximately 5.7% in macro F1-score compared to the best single-modality model.

3 Dataset

3.1 TikHarm Dataset

TikHarm is a dataset developed to support the task of detecting harmful content on the TikTok platform, with a particular focus on protecting children from such content. The TikHarm dataset is available at Link . The dataset consists of 3,948 videos collected from TikTok through unofficial APIs, focusing on popular hashtags and keywords commonly appearing in child-oriented content. Videos were selected based on engagement metrics (views, likes, comments) while ensuring diversity in topics and posting times.

After the collection process, all videos were manually annotated into four content categories: **Safe**, **Adult Content**, **Harmful Content**, and **Suicide**. The annotation was carried out by a team of trained experts, following unified guidelines to minimize subjective bias. The inter-annotator agreement, measured by the Fleiss' Kappa coefficient, reached 81.25%, indicating high consistency and reliability in the classification process.

The dataset is split into three subsets: *training*, *development*, and *testing*. Table 1 presents the number of samples, average duration, and total video duration for each split.

The class distribution of the entire dataset is detailed in Table 2. It can be observed that videos in the **Safe** class tend to have a significantly longer

	Duration (s)		
	Samples	**Min**	**Max**	**Avg**
Train set	2,762	3.88	600	38.71
Dev. set	396	1.95	600	38.77
Test set	790	5.04	600	38.57

Table 1: Distribution of Video Samples and Duration by Data Split in the TikHarm Dataset.

average duration, reflecting the characteristics of educational or wholesome entertainment content. Conversely, the **Suicide** class shows a much shorter average duration, partly due to TikTok's strict moderation policies. With balanced class distribution and reliable annotations, TikHarm serves as a robust foundation for training and evaluating deep learning models in harmful content detection tasks. Notably, it is the first dataset in Vietnam dedicated to TikTok content safety for children, supporting research in video processing and real-time content moderation systems.

	Duration (s)		
	Samples	**Min**	**Max**	**Avg**
Safe	997	16.96	65.36	18.10
Adult Content	977	9.84	36.25	18.10
Harmful Content	990	9.88	35.92	18.10
Suicide	984	4.63	16.96	18.10

Table 2: Distribution of Video Samples and Duration by Label in the TikHarm Dataset.

3.2 Extended Dataset

One of the **main contributions** of this study is the construction and release of an Extended Dataset (available at https://www.kaggle.com/datasets/kusnguyen/extra-dataset) to enhance the coverage and representation of trending harmful content, thereby improving the detection of more subtle and sophisticated forms of harmful material on TikTok. This dataset was collected and annotated using a rigorous procedure derived from the original TikHarm dataset, ensuring consistent quality and labeling criteria, while incorporating new samples that reflect current content trends on the platform.

Dataset Collection. The supplementary data collection began by visiting TikTok's hashtag pages to identify trending hashtags. These hashtags were categorized based on their potential to yield videos falling into the four labels of the dataset. From this filtered list, we randomly collected 1,000 videos
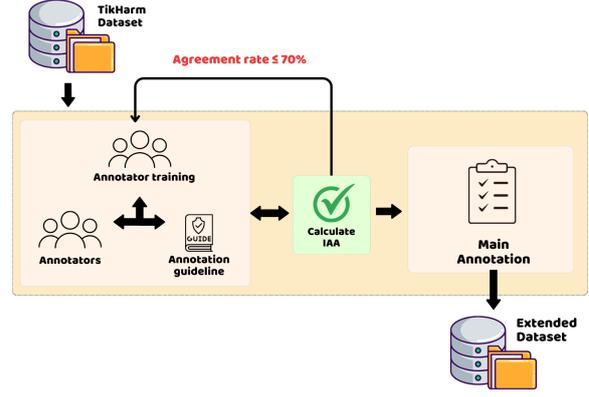

Figure 1: The dataset extending process.

using Selenium in combination with the TikTok-Content-Scraper tool. The selection was guided by engagement metrics (views, likes, comments) and ensured diversity in topics and posting times, consistent with the original TikHarm collection methodology.

Dataset Annotation. Figure 1 illustrates the overall dataset extending process. Before annotation, we refined the labeling guidelines by clearly defining each label and providing detailed illustrative examples. This ensured the extended dataset maintained high quality and alignment with the original TikHarm annotations. Three annotators, thoroughly trained on the updated guidelines, first labeled a test set of 100 samples from the original dataset. Their agreement with the ground truth achieved Cohen's Kappa scores of 0.82, 0.78, and 0.80, reflecting substantial to almost perfect agreement. Inter-annotator analysis revealed very high consensus, with two annotators achieving a Cohen's Kappa of 0.97. The overall Fleiss' Kappa across all three annotators reached 0.86, indicating almost perfect agreement. These results validated the annotators' consistent application of the labeling criteria, enabling each annotator to proceed independently with the remaining extended data.

After annotation and removal of low-quality samples, we obtained 775 qualified videos. Unlike the original TikHarm dataset, which was intentionally balanced across labels, the extended dataset has an imbalanced label distribution (Table 3). This imbalance was intentionally preserved to better reflect the real-world distribution of harmful content on TikTok, where certain categories appear more frequently than others. Such a distribution allows models to learn from "in-the-wild" data, improving their generalization in real-world scenarios. Further

details of the annotation guidelines are presented in Appendix A.2.

	Duration (s)			
	Samples	Min	Max	Avg
Safe	251	3.04	60	25.97
Adult Content	164	4.69	60	18.78
Harmful Content	203	5.06	60	22.61
Suicide	157	6.00	60	19.83

Table 3: Distribution of Video Samples and Duration by Label in the Extended Dataset.

3.3 Final Combined Dataset

The Extended Dataset (Section 3.2) was merged with TikHarm to form the Final Combined Dataset, which integrates both the balanced distribution of the original dataset and the realistic distribution of the extended one. This combination provides a unique benchmark that not only retains comparability with prior studies (by keeping the original TikHarm test set intact) but also challenges models with a more representative and diverse range of harmful content.

The extended samples were split into training and development sets at an 8.5:1.5 ratio, while the TikHarm test set was kept unchanged to ensure fair evaluation against existing methods. The detailed distribution of samples across the three datasets is presented in Table 4.

Dataset	Train	Dev.	Test	Total
TikHarm	2,762	396	790	3,948
Extended Dataset	656	119	-	775
Final Combined Dataset	3,418	515	790	4,723

Table 4: Sample distribution across TikHarm, Extended Dataset, and the Final Combined Dataset.

By uniting the curated balance of TikHarm with the real-world skew of the Extended Dataset, the Final Combined Dataset represents one of the most comprehensive and practical resources for harmful content detection on TikTok to date, making it a key contribution of this research.

4 Method

4.1 Data Processing

Data preprocessing is a crucial step to ensure high-quality input for the multimodal classification system, especially in the noisy and diverse context of short-form TikTok videos. We design a processing pipeline that extracts and refines linguistic information from both the audio and the visual frames of each video.

Text Extraction from Audio. The audio signal is first separated from the video using the `librosa` library, with each sample limited to a maximum duration of 60 seconds to optimize processing efficiency. After normalization, low-energy segments are removed based on the Root Mean Square (RMS) criterion. Speech recognition is performed in parallel for Vietnamese and English using the *Whisper large-v3* model (Radford et al., 2022). The hypothesis with the most complete contextual coverage is retained. To reduce hallucinated or generic outputs, we apply a filter to remove short or repetitive sentences, as well as phrases without meaningful content (e.g., "thank you", "hello", "bye"). In addition, a specialized spam filter, implemented via regular expressions, is used to remove advertising text, engagement calls, or meaningless repetitions common in social-media content.

Text Extraction from Frames. In parallel with the audio processing, text appearing in the video frames is extracted using *EASYOCR*[1], which supports Vietnamese and English. To balance recognition accuracy and computational cost, two representative frames are selected at 30% and 70% of the total video duration. Each frame is resized so that its longer side does not exceed 640 pixels. The OCR results are then processed through the same spam and language filters as the audio transcripts. Only Vietnamese or English text that is non-duplicated and passes the noise filter is retained.

Synthesis and Standardization of Text Input. The retained transcripts from both modalities (audio and OCR) are concatenated into a single string following the template: `Audio: ... | OCR: ...`. A lightweight language detector (Joulin et al., 2016) ensures that only Vietnamese or English content is preserved, while all other languages and non-linguistic text are removed. This process produces a compact, semantically rich textual input that reduces noise and improves the reliability of downstream classification.

4.2 Multimodal Harmful Content Detection Architecture

The proposed architecture (Figure 2) leverages the strengths of two pretrained encoders: *TimeSFormer*

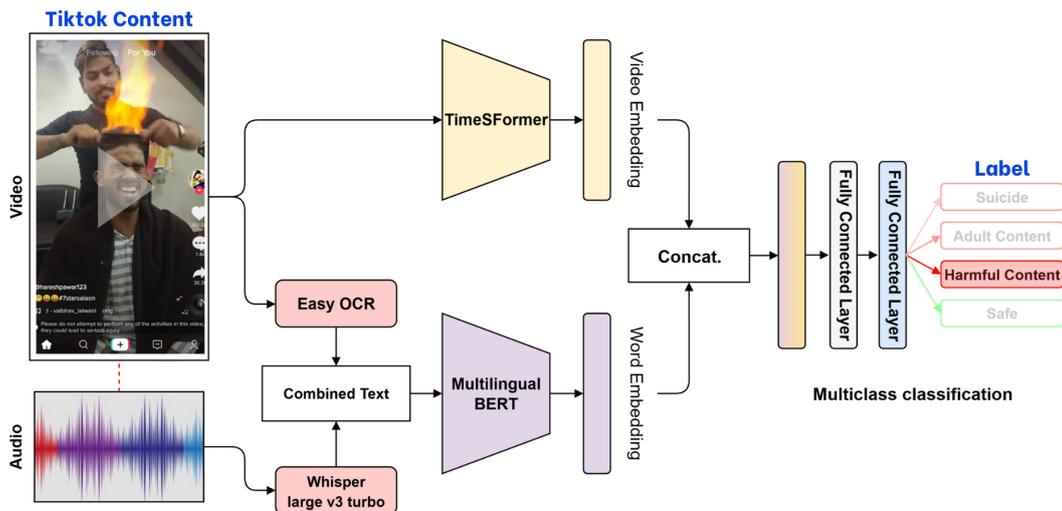[1]https://github.com/JaidedAI/EasyOCR

Figure 2: Proposed multimodal harmful content detection architecture.

for video representation and a transformer-based language model for text representation.

Video Encoder. We adopt *TimeSFormer*, an advanced transformer-based model that applies the Divided Space–Time Attention mechanism to separately model spatial and temporal dependencies. This design has been shown to improve classification efficiency and achieve state-of-the-art performance on large-scale benchmarks such as Kinetics-400 and Kinetics-600.

Text Encoder. Since the content of the videos primarily includes Vietnamese and English, we use *Multilingual DistilBERT* (Sanh et al., 2020) to effectively capture multilingual context. We also experiment with *ViSoBERT* (Nguyen et al., 2023), a language model optimized for Vietnamese social-media text. Each encoder output is passed through a separate projection block consisting of a linear layer, a ReLU activation, and dropout to reduce overfitting.

Feature Fusion and Classification. The projected features from the video and text encoders are concatenated and passed through a fusion network composed of multiple linear layers, Layer Normalization, ReLU activations, and dropout. The resulting joint multimodal representation is then fed into a final linear layer to perform four-way classification into *Safe*, *Adult Content*, *Harmful Content*, and *Suicide*. This late-fusion strategy preserves the strengths of each individual modality while

enabling the system to exploit cross-modal cues, improving its generalization ability in multimodal harmful content detection. In addition to the late-fusion strategy, we also design an architecture with attention-based fusion. Specifically, similar to prior studies on multimodal learning, after concatenating the two modalities, the combined features are passed through Multihead Self-Attention layers, which allow the model to capture global relationships between textual and visual features, enrich the joint representation, and focus on the most relevant information.

5 Experiments

5.1 Experimental Configurations

We conducted a series of offline experiments to evaluate the effectiveness of the proposed multimodal architecture for harmful video content detection. The evaluation was performed on both the original *TikHarm* dataset and an *expanded dataset*, where the latter was constructed by augmenting *TikHarm* with additional crawled samples in a train–validation ratio of 8.5:1.5.

For the video encoder, we adopted the pretrained *TimeSFormer* model, while for the text encoder, we experimented with *Multilingual DistilBERT* and *ViSoBERT* to determine the most suitable language model for integration into the actual system.

To ensure fairness across experiments, we fixed the hyperparameters for all runs: training was conducted for 6 epochs with a batch size of 8 for both

training and evaluation. The initial learning rate was set to 1e-4, combined with a warmup ratio and weight decay, and gradients were accumulated every two steps. Losses on the training and validation sets were computed every 100 steps, and performance was evaluated using accuracy, macro-precision, macro-recall, and macro-F1. The macro formulation was chosen to mitigate class imbalance by giving equal weight to each class, including minority labels, thereby reducing bias toward dominant classes - a common challenge in harmful content classification tasks.

We retained up to three checkpoints per training run and selected the best model based on the highest macro-F1 score on the validation set. The final reported results on the test set were obtained using the checkpoint with the best macro-F1 score. The implementation was carried out in *PyTorch* and executed on the Kaggle platform with an Intel(R) Xeon(R) CPU @ 2.00GHz and an NVIDIA Tesla P100 GPU (16 GB, CUDA 11.4).

5.2 Main Results

As shown in Table 5, on the original TikHarm dataset, the results demonstrate the clear advantage of multimodal integration with attention-based fusion. On the validation set, ViSoBERT with attention-based fusion achieved the highest performance, surpassing its concat-based late-fusion variant. This indicates that when linguistic and visual features are deliberately integrated at the attention layer, the model can better exploit cross-modal interactions to reinforce classification decisions. On the TikHarm test set, Multilingual BERT achieved the highest macro-F1 score of 89.45 (Acc 89.37), followed by its attention-based variant. The inversion of rankings between the validation and test sets suggests that the cross-lingual generalization capability of Multilingual BERT is advantageous when evaluating on unseen data. In contrast, TimeSFormer-only, relying solely on visual features, performed substantially worse (F1 86.62 on validation and 86.68 on test), reinforcing the critical role of textual information in decoding the semantics of short-form videos.

On the Final Combined dataset (our extended version), validation performance dropped slightly to around 80 F1, with Multilingual BERT achieving the highest score (F1 80.77; Acc 80.78), followed by ViSoBERT (F1 80.41; Acc 80.58). This decline is expected, as the dataset was deliberately expanded to validate model robustness un-

der real-world conditions with label imbalance and ambiguous cases. On the test set, however, models maintained strong performance close to 88% F1. Multilingual BERT with attention-based fusion achieved the best result (F1 88.44; Acc 88.48), slightly outperforming its concat late-fusion counterpart (F1 88.31; Acc 88.35) and both ViSoBERT configurations. This finding suggests that when data becomes more diverse and noisy, attention-based fusion enables the model to better capture cross-modal interactions, thereby improving the detection and classification of harmful video content in real-world scenarios.

Notably, since both the original and extended datasets naturally include noisy audio, blurred frames, and code-switched language, the strong performance across all settings indicates that MTik-Guard remains robust under such imperfect conditions. This highlights the system's ability to generalize well even in noisy, real-world TikTok scenarios, with qualitative examples provided in Appendix A.4.

5.3 Ablation Study

To better understand the contribution of each modality, we conducted an ablation study where the model was trained and evaluated using only one modality at a time: **OCR**, **ASR**, or **Video**. The results on both TikHarm and the Final Combined Dataset are presented in Table 6.

The results indicate that the **Video** modality plays the most critical role, achieving nearly 89% accuracy and F1-score on the TikHarm dataset, and maintaining high performance on the Final Combined Dataset. **ASR** contributes moderately, with F1 around 47–50%, showing that speech transcripts provide useful but less stable information due to noise, informal speech, or lack of context. **OCR** performs weakest, below 37% accuracy and around 30% F1, largely because textual content in TikTok videos is often sparse, noisy, or unrelated (e.g., ads or captions). Overall, although the Video modality dominates, combining it with ASR and OCR in a multimodal fusion setup significantly improves performance, particularly for subtle or disguised harmful content that is difficult to detect with visuals alone.

5.4 Discussion

The confusion matrix (Figure 3) reveals that the system achieved an overall accuracy of **89.37%** on the test set. The *Safe* category was most accurately pre-

Dataset	Language Model	Validation set		Test set	
		Acc.	**F1**	**Acc.**	**F1**
TikHarm	TimeSFormer-only	86.66	86.62	86.71	86.68
	Multilingual BERT	88.64	88.56	**89.37**	**89.45**
	Multilingual BERT [♦]	88.13	88.14	<u>88.35</u>	<u>88.38</u>
	ViSoBERT	<u>89.14</u>	<u>89.19</u>	87.22	87.18
	ViSoBERT [♦]	**90.15**	**90.17**	87.97	87.87
Final Combined	Multilingual BERT	**80.78**	**80.77**	<u>88.35</u>	<u>88.31</u>
	Multilingual BERT [♦]	<u>80.58</u>	80.32	**88.48**	**88.44**
	ViSoBERT	<u>80.58</u>	<u>80.41</u>	87.47	87.39
	ViSoBERT [♦]	80.39	80.22	88.10	88.03

Table 5: Performance Comparison of Multimodal Models on TikHarm and Final Combined Datasets. The table shows classification results for various language models integrated with TimeSFormer, evaluated on the Validation and Test sets. The notation [♦] denotes models employing attention-based fusion to integrate visual and textual features. Bold values represent the best result, while underlined values signify the second-best performance within each dataset group. The TimeSFormer-only row refers to the baseline architecture utilizing only the visual encoder without any language model, with results adopted directly from (Balat et al., 2024).

Dataset	Modality			Validation set		Test set	
	Video	**OCR**	**ASR**	**Acc.**	**F1**	**Acc.**	**F1**
TikHarm	×			**88.64**	**88.64**	<u>88.99</u>	<u>88.92</u>
		×		36.11	31.33	35.44	30.27
			×	46.72	47.05	49.11	48.88
	×	×	×	**88.64**	<u>88.56</u>	**89.37**	**89.45**
Final Combined	×			<u>80.58</u>	<u>80.44</u>	87.97	87.90
		×		33.40	27.77	36.2	30.52
			×	44.85	44.72	50.51	50.26
	×	×	×	**80.78**	**80.77**	**88.35**	**88.31**

Table 6: Ablation Study on Modality Effectiveness Using the Best-Performing Architecture. This table shows the contribution of each modality: Video, OCR and ASR, using the best model configuration on TikHarm Dataset (last row) as the baseline. The × symbol indicates the modality was retained; removing it (empty cell) demonstrates its individual impact on performance (Acc. and F1).

dicted (183 correct classifications) with minimal mislabeling, indicating high reliability in detecting non-harmful content. *Suicide* also performed strongly (183 correct predictions) but showed some confusion with *Adult Content* and *Harmful Content*. While *Adult Content* achieved 172 correct identifications, the *Harmful Content* category (168 correct predictions) exhibited the **highest misclassification rate into *Suicide***, clearly suggesting overlapping expression patterns in sensitive themes.

We adopted a multi-class approach due to the observed content overlap and the implicit hierarchy among harmful labels (e.g., Suicide is more severe than Harmful). Therefore, the authors of the TikHarm dataset and we prioritized the maximum risk in the labeling process, assigning a single, definitive label to each piece of content. This choice simplifies the decision-making process into a clear, automated action (BLOCK/ALLOW), avoiding the complexity of multi-label output. To manage cases where the model has low confidence (is "confused"), we will utilize Label Smoothing during training and set a confidence threshold to flag ambiguous cases for human review, thus effectively managing the overlap without altering the core label structure.

Besides the intrinsic issue of label overlap, data imbalance is also a challenging factor, particularly affecting the minority class, *Suicide*. We observed higher misclassification between *Suicide* and *Harmful/Adult content*. Nevertheless, we retained this imbalance to reflect real-world TikTok distributions, where suicide-related videos are rarer and often subtle. This improves generalization in deployment, though future work may adopt reweighting strategies or focal loss to further boost minority-
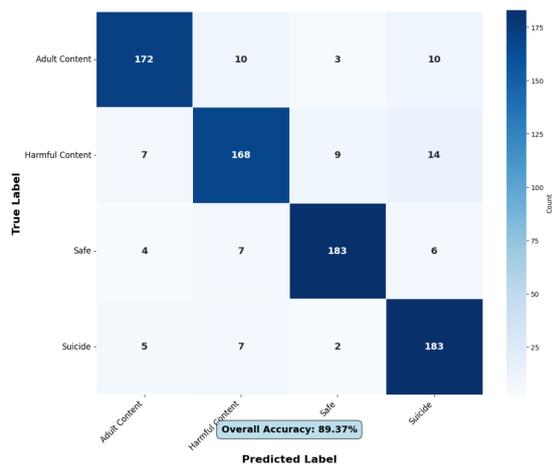
Figure 3: Confusion matrix of the best-performing model on the test set.

class recall.

TikTok videos often contain noisy audio, blurred frames, and code-switched language, which naturally introduce errors in OCR and ASR. Despite these challenges, MTikGuard demonstrates strong robustness thanks to multimodal fusion, where errors in one modality can be compensated by signals from another. For instance, background noise may reduce ASR quality, but OCR-extracted harmful text can still support correct classification, and vice versa. Ablation results (Table 6) confirm this effect: unimodal OCR or ASR alone performs poorly, while multimodal fusion consistently improves accuracy by about 2–3%. Moreover, Multilingual DistilBERT effectively handles code-switching (e.g., Vietnamese speech with English captions). Together, these findings highlight the system's reliability in noisy, real-world TikTok scenarios.

Overall, the model demonstrated robust discriminative ability across all categories despite the diverse and unpredictable nature of TikTok short videos.

6 MTikGuard System

Figure 4 illustrates the overall architecture of **MTikGuard**, a modular and extensible system designed for real-time multimodal content moderation on TikTok. The design ensures adaptability to various deployment environments and scalability to handle high-throughput streaming data. The system is structured into three primary functional layers, presented from bottom to top as follows:

The foundation layer focuses on dataset preparation and model training. The system leverages the TikHarm dataset, augmented with an extended dataset to enhance coverage and robustness. For each video, visual features are extracted using the TimeSFormer video transformer, while textual features are derived from transcribed audio via Whisper and scene text via EasyOCR. Text representations are encoded using either Multilingual DistilBERT or ViSoBERT, as described in previous sections. The resulting multimodal architecture is trained and evaluated using macro-F1 as the primary selection metric. Once the optimal model is obtained, it is serialized and integrated into the real-time censorship pipeline.

The middle layer implements the automated streaming pipeline, orchestrated with Apache Airflow[2] and containerized using Docker[3] for portability and reproducibility. The pipeline includes:

- A Selenium-based crawler[4] together with the TikTok Content Scraper module retrieves trending TikTok videos based on hashtags, extracts metadata and video IDs, and publishes them to an Apache Kafka[5] topic .

- Kafka consumers download the videos, extract text from audio using Whispe and from frames using EasyOCR, and then classify the content using the trained multimodal model. The classification results, along with metadata, are stored in MongoDB Atlas [6] for further querying and analysis.

- Airflow coordinates the execution of producers and consumers, monitors pipeline health, and manages periodic reporting, ensuring stability and scalability.

The top layer emphasizes the modular nature of MTikGuard, allowing flexible deployment in both cloud-based and on-premise environments. Each component - such as Kafka consumers, inference services, or data storage - can be scaled independently. The architecture also facilitates integration with additional modalities or extension to other short-video platforms in future work.

7 Conclusion

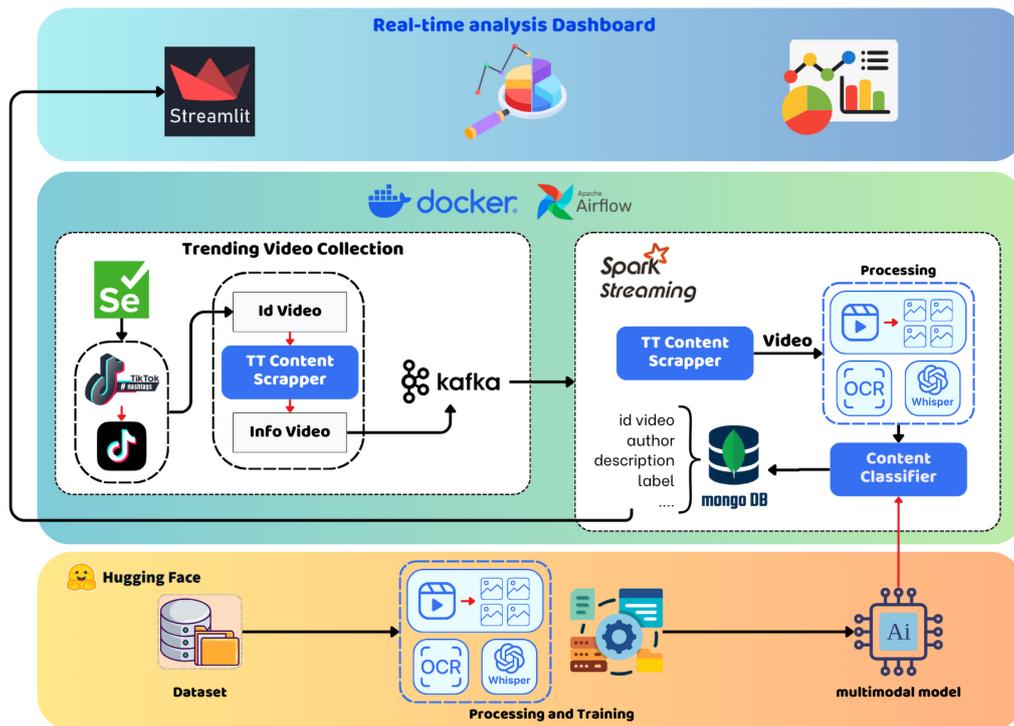We introduced **MTikGuard**, a real-time multimodal detection system for harmful TikTok con-

---

Figure 4: Overview of the MTikGuard system.

tent that jointly analyzes visual, audio, and textual modalities. By extending the TikHarm dataset to 4,723 annotated videos and leveraging transformer-based encoders (TimeSFormer and DistilBERT) within a scalable streaming pipeline, MTikGuard achieves **state-of-the-art performance** with 89.37% accuracy and 89.45% macro-F1. These results highlight the effectiveness of combining dataset expansion, multimodal fusion, and practical deployment infrastructure in safeguarding young users from harmful online content. In addition, our pipeline is inherently multilingual: Whisper and EasyOCR support dozens of languages, while Multilingual DistilBERT enables robust cross-lingual text encoding. Thus, extending to other regions would require minimal adjustments, primarily re-training with a small annotated subset to adapt to local linguistic nuances.

While the current system adopts a late-fusion strategy, it does not yet explore more advanced mechanisms to dynamically weight each modality's contribution. In future work, we plan to enhance audio processing to better handle non-speech cues and incorporate attention-driven multimodal fusion to improve robustness and interpretability in complex real-world scenarios.

References

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 Novel Intelligent and Leading Emerging Sciences Conference (NILES)*, pages 337–340.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? *Preprint*, arXiv:2102.05095.

Mithun Das, Rohit Raj, Punyajoy Saha, Binny Mathew, Manish Gupta, and Animesh Mukherjee. 2023. Hatemm: A multi-modal dataset for hate video classification. *Preprint*, arXiv:2305.03915.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *Preprint*, arXiv:1607.01759.

Dhiraj Murthy, Rachel Ouellette, Tanvi Anand, Srijith Radhakrishnan, Nikhil Mohan, Juhan Lee, and Grace Kong. 2024. Using computer vision to detect e-cigarette content in tiktok videos. *Nicotine and Tobacco Research*, 26:S36–S42.

Quoc-Nam Nguyen, Thang Chau Phan, Duc-Vu Nguyen, and Kiet Van Nguyen. 2023. Visobert: A pre-trained language model for vietnamese social media text processing. *Preprint*, arXiv:2310.11166.

Orchid Chetia Phukan, Mohd Mujtaba Akhtar, Girish, Swarup Ranjan Behera, Abu Osama Siddiqui, Sarthak Jain, Priyabrata Mallick, Jaya Sai Kiran Patibandla, Pailla Balakrishna Reddy, Arun Balaji Buduru, and Rajesh Sharma. 2025. Snifr : Boosting fine-grained child harmful content detection through audio-visual alignment with cascaded cross-transformer. *Preprint*, arXiv:2506.03378.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Shubham Singh, Rishabh Kaushal, Arun Balaji Buduru, and Ponnurangam Kumaraguru. 2019. Kidsguard: fine grained approach for child unsafe video representation and detection. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 2104–2111, New York, NY, USA. Association for Computing Machinery.

Kanwal Yousaf and Tabassam Nawaz. 2022. A deep learning-based approach for inappropriate content detection and classification of youtube videos. *IEEE Access*, 10:16283–16298.

A Appendix

A.1 Task Definition

The harmful content detection task in TikTok videos is formulated as a supervised **multimodal classification** problem. Given a video V, the system processes three modalities: (1) *visual frames* F_v, (2) *audio transcripts* T_a (from speech-to-text), and (3) *scene text* T_o (from OCR on frames). After noise removal and normalization, each modality is encoded into feature vectors: x_v for visual data and x_t for text (combining T_a and T_o). These are fused using a late fusion strategy:

$$x_m = \text{Fusion}(x_v, x_t) \qquad (1)$$

The classifier f_θ maps x_m to one of four labels:

$$y \in \{\text{Safe, Adult Content, Harmful Content, Suicide}\} \quad (2)$$

The goal is to minimize cross-entropy loss on the labeled dataset $\mathcal{D}$ while maximizing the **macro-F1 score** to ensure balanced performance across classes.
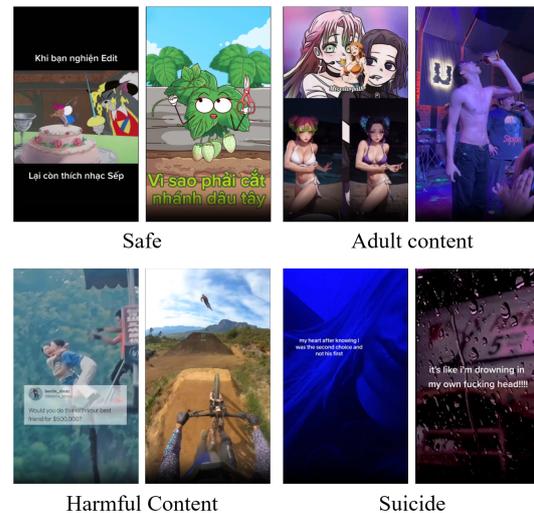

Safe   Adult content

Harmful Content   Suicide

Figure 5: An example of the Harmful Content Detection task.

A.2 Annotation Guidelines

A.2.1 Annotation Tool

To support this complex annotation task, we developed a dedicated web-based annotation tool (available at `https://github.com/Kusssssssss/tiktok-labeling-tool.git`). The system displays a list of videos and associated metadata (user name and video ID), a video player for observing content directly, and a label selection panel. The
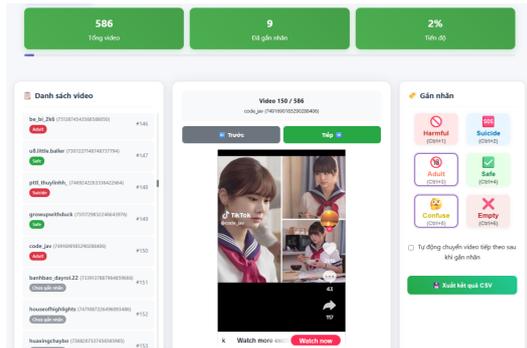
Figure 6: The web application interface for annotators.

interface also provides progress tracking, together with two auxiliary options, *confuse* and *empty*, to handle ambiguous cases and ensure labeling quality and consistency (Figure 6).

A.2.2 Annotation Guidelines

Each annotator must strictly follow the label definitions derived from the TikHarm dataset. Only one label can be assigned to each video; multiple labels are considered invalid. Samples that are unclear are first labeled based on the annotator's best judgment and later cross-validated with others.

- **Safe:** Content that is appropriate and safe for children, without violence, sexuality, strong language, or dangerous behaviour. *(e.g., educational programs, family-friendly vlogs, cartoons, DIY activities).*

- **Adult Content:** Content containing sexual elements, nudity, or language not suitable for children, including sexually suggestive material. *(e.g., provocative dancing, lingerie try-ons, sexual jokes or innuendo).*

- **Harmful Content:** Content depicting violence, dangerous acts, or harmful behaviour that may negatively influence children. *(e.g., fights, bullying, dangerous challenges involving weapons or chemicals).*

- **Suicide:** Content referring to suicidal behaviour, intentions, or methods that may negatively impact viewers, especially minors. *(e.g., self-harm attempts, suicide tutorials, confessions of suicidal thoughts).*

A.3 Error Analysis Details

We observed two main error types. The first is misclassification, where the predicted label differs from the ground truth (Figure 7). For example,
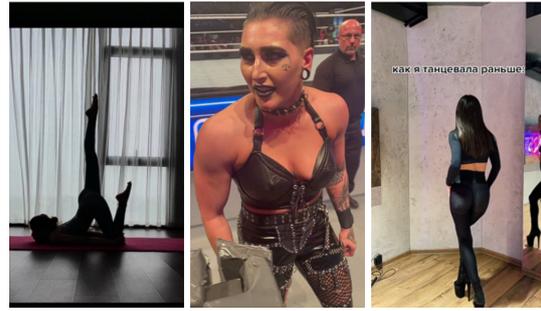

Figure 7: Examples of misclassification cases where predicted labels differ from ground truth. From left to right, Safe as Harmful Content, Harmful Content as Adult Content and Adult Content as Harmful Content.

Figure 8: Examples of overlapping-label cases where videos could plausibly belong to multiple categories. From left to right, Harmful Content also Adult Content, Suicide also Adult Content and Adult Content also Harmful Content.

Safe videos predicted as *Harmful Content* due to poses resembling unsafe contexts, *Harmful Content* misclassified as *Adult Content* because of clothing or makeup cues, and *Adult Content* predicted as *Harmful Content* due to aggressive thematic elements. In some cases, these predictions may still be justifiable under stricter moderation perspectives. The second type is overlapping-label cases, where videos plausibly belong to more than one harmful category (Figure 8), such as content mixing aggressive and sexual elements, suicide-related context with adult cues, or mature visuals with harmful implications. These findings suggest that future work should explore multi-label classification and improved multimodal fusion to address ambiguous cases.

A.4 Noisy Cases

As shown in Figure 9, the case (a) combines multiple types of noise: teencode in the opening frames, numeric overlays in later segments, loud rock-style background audio, and flickering visuals. The case (b) suffers from strong lighting distortions
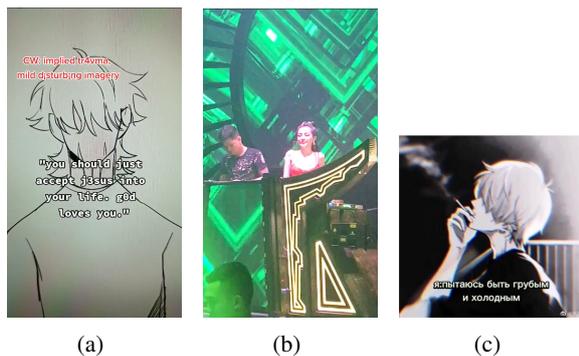
Figure 9: Examples of noisy cases. From left to right: (a) stylized and numeric text with blurred, flickering frames, (b) nightclub environment with light and audio noise, (c) code-switched content with unrelated audio.

and audio that lacks linguistic content, while (c) mixes cross-lingual subtitles with English audio unrelated to the annotated label. Despite these imperfections, MTikGuard successfully classified all cases into their correct categories - (a) *Suicide*, (b) *Adult Content*, (c) *Suicide*, confirming that the multimodal fusion design provides robustness even under noisy and ambiguous conditions typical of TikTok videos.