NoHateZone: A Multi-Modal Framework for censoring hateful videos

Mert Ülgüner, Noah Truttmann, Amene Gafsi

Group 40

Problem definition

- Hate is spread in a multi-modal fashion and detecting hate speech in text is not enough.
- Hate speech on social media is often embedded in short videos in memes format.
- Current models don't identify or censor where the hate is occurring in videos.
- Need for a multimodal framework to detect and censor only hateful segments of videos on social media.

Key Related Works

Towards a Robust Framework for Multimodal Hate Detection: A study on Video vs. Image-based Content (Khousik et. al., 2025)

Multi-modal framework capturing hate in text-image and HateMM videos.

- **Contribution**: Demonstrated that evaluating videos and understanding text-image relationships are not separate problems regarding HateMM videos.

A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of Youtube Videos (Yousaf, Nawaz, 2022)

Introduced an architecture blurring inappropriate frames in videos

- **Contribution:** Proposed a multi-modal framework which additionally censors inappropriate speech simultaneously alongside frames.

HateMM: A Multi-Modal Dataset for Hate Video Classification (Das et. Al., 2023)

Introduced a video dataset and simple architecture to capture hate in videos

- **Contribution:** Introduced **attention-based fusion architecture** helping to capture cross-modal relationships as opposed to a simple concatenation of features

A comprehensive framework for multi-modal hate speech detection in social media using deep learning (Prabhu, Seethalakshmi, 2025)

Novel architecture for hate detection in HateMM videos

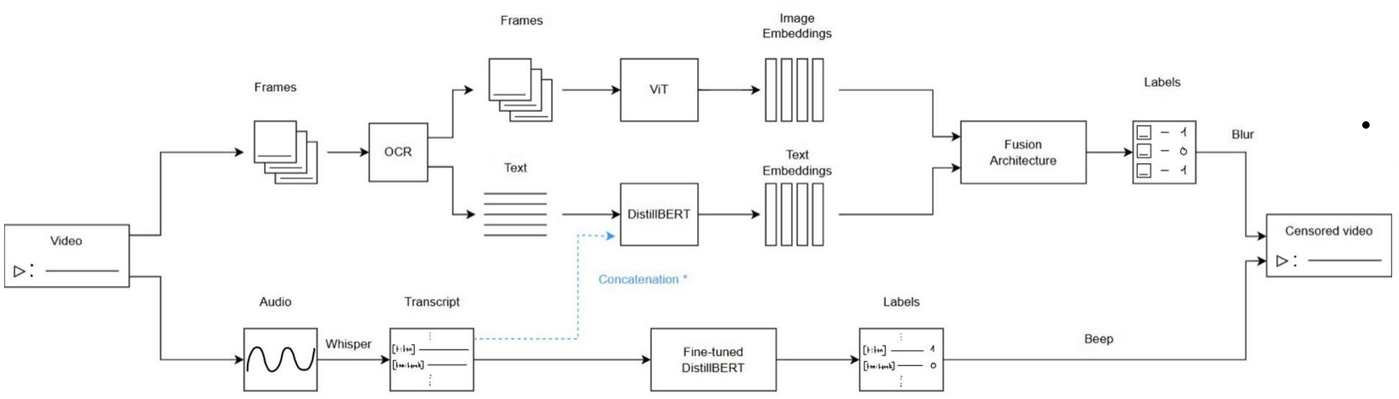
- **Contribution:** Used a **ViT** instead of a combination of CNN, RNN and LSTMs leveraging pre-trained weights and patch-wise attention scores for better contextual understanding

Datasets

- **HateSpeechDataset**: A curated dataset containing over 417,000 text samples, each labeled as either hateful or non-hateful.
- **MMHS150K**: dataset including 150,000 tweets, memes and text extracted from the memes.
- **HateMM (HMM)**: dataset including 411 hateful and 652 non-hateful videos with binary labels including the specific time intervals indicating the hateful parts. [2]
  - **Reannotation for modalities**: Each group member annotated the source modality of hate (e.g. Audio, Text, Image) for all videos present in the dataset
  - **Manual adjustment for time-intervals**: Re-annotated the time-intervals capturing the hate part in 159 hateful videos containing 2247 hateful frames

Method

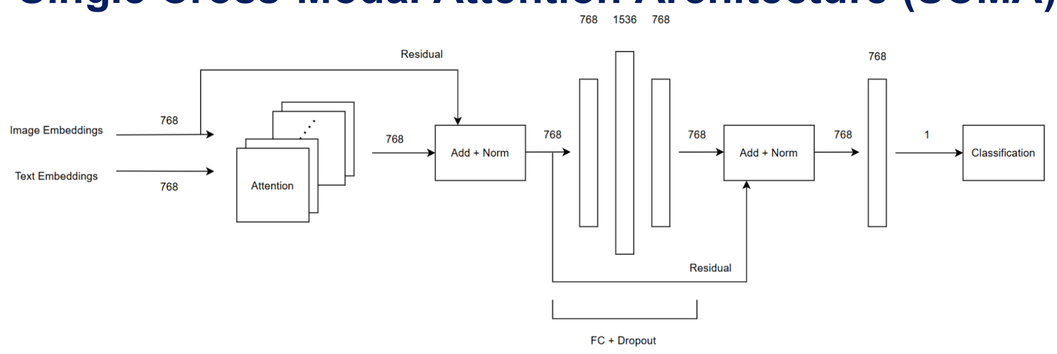
Main Framework



*Concatenation is performed based on the choice of the Fusion Architecture (Single Cross-Modal Attention)

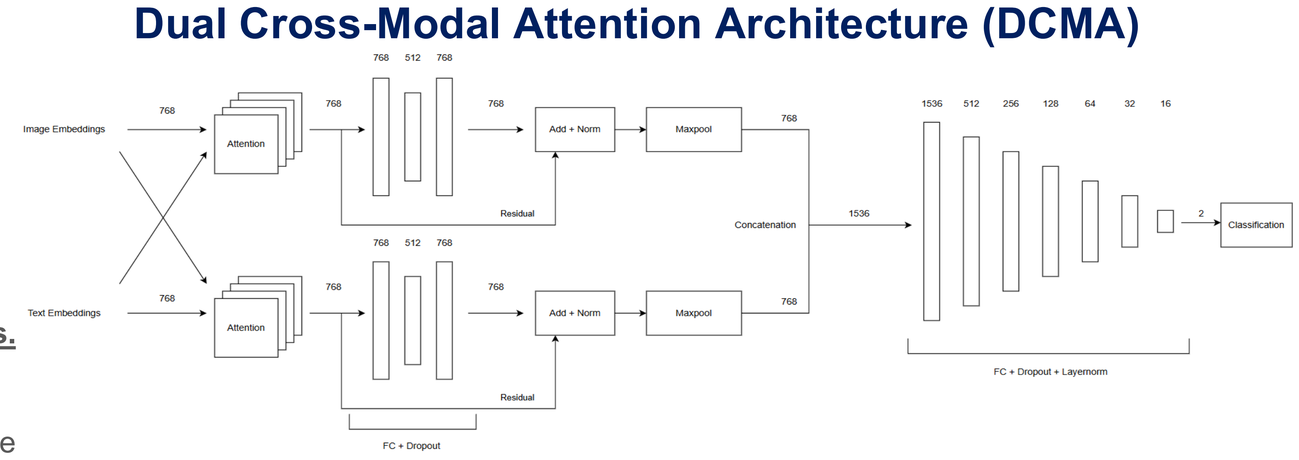
- **Audio censoring mechanism**
  - **Whisper** (*OpenAI/whisper-large-v3-turbo*) model is used to transcribe the audio into text and seperate into sentence-level chunks with timestamps
  - **DistillBERT** fine-tuned on HateSpeechDataset
- **Frame censoring mechanism**
  - Text on frames extracted using OCR (*ucaslcl/GOT-OCR2_0*)
  - Frame embeddings produced by Vision Transformer (ViT) specifically (*google/vit-base-patch16-224*) and Text embeddings produced by DistilBERT
  - Fused by architectures for classification: **Single-Cross Modal Attention (SCMA)** and **Dual-Cross Modal Attention (DCMA)**
  - The fusion architectures are pretrained on the MMHS150K dataset, which includes both tweet text and OCR-extracted text from images. These two modalities are concatenated using a `[SEP]` token.
  - Once pre-trained, the fusion architectures are fine-tuned on the HMM dataset. Each video is split into 1-second frames.
  - Stratified cross-validation is used to identify optimal hyperparameters.

Single Cross-Modal Attention Architecture (SCMA)



- **One Cross-Modal Attention block**: Image attended to concatenated text (**image** embeddings as **queries** and **text** embeddings as both **key** and **values)**
- Conditions the semantic meaning of audio transcript chunks and OCR texts on video frame embeddings [1]
- **Input dataset**: Audio transcription chunks with timestamps concatenated with image embeddings (single image is chosen randomly per audio chunk interval)

Dual Cross-Modal Attention Architecture (DCMA)



- **Two Cross-Modal Attention blocks**: Image attended to OCR text & OCR text attended to image
- Each Cross-Modal Attention block is followed by a Fully Connected (FC) block to capture modality-specific patterns.
- The resulting outputs are concatenated and passed through another Fully Connected (FC) block, followed by a binary classification head.

Validation

- **Fine-tuned DistilBERT (DBertXhate)**: State-of-the-art results have been achieved on hate detection in sentences, exceeding the F1 and AUC scores of existing models on the HateXplain dataset. [4]

Model	Dataset	F1	AUC	ACC
CNN-GRU		0.627	0.606	0.793
BiRNN		0.595	0.575	0.767
BiRNN-Attn		0.621	0.614	0.795
BiRNN-HateXplain	HateXplain	0.629	0.629	0.825
BERT		0.690	0.674	0.843
BERT-HateXplain		0.698	0.687	0.851
DBertXhate		0.741	0.746	0.690
DBertXhate	HateSpeechDataset	0.729	0.939	0.896
DBertXhate	HMM Audio	0.645	0.762	0.669

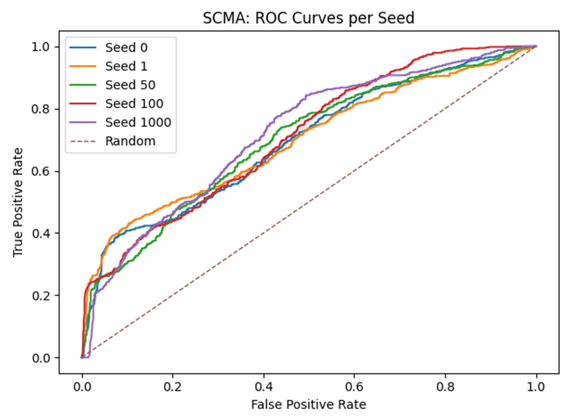
- **Pre-trained Fusion Architectures**: Comparable AUC and improved accuracy have been achieved, although the F1-score remains lower than that of the architectures reported in [3].

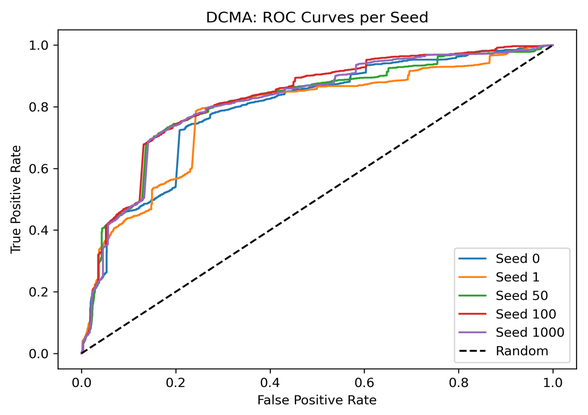
Model	Dataset	F1	AUC	ACC
FCM		0.704	0.734	0.684
SCM		0.702	0.732	0.685
TKM	MMH150K	0.701	0.731	0.682
SCMA		0.490	0.715	0.643
DCMA		0.519	0.732	0.701

- **Fine-tuned Fusion Architectures**: Since consecutive frames can be highly similar, or even identical, the train-test split is performed at the video level rather than the frame level. Some videos were found to be duplicated in the HMM dataset, these duplicates have been removed to ensure the test set remains clean and unbiased.

Model	Input	Target	F1	AUC	ACC	Precision	Recall
Random	–		0.301	0.508	0.503	0.219	0.503
SCMA	*F&A*	*F*	0.477 ± 0.013	0.703 ± 0.008	0.661 ± 0.008	0.508 ± 0.017	0.549 ± 0.015
DCMA	*F*		**0.695** ± 0.007	**0.811** ± 0.001	**0.760** ± 0.009	**0.636** ± 0.017	**0.766** ± 0.024

F: HMM Frames F&A : HMM Frames + Audio





Sources of variability: randomness in mini-batch sampling, dropout choices are random and seed-dependent, moment-estimators in ADAM accumulate differently

Limitations

- Audio hate detection limited by the performance of Whisper which is known to hallucinate transcriptions and miss the beginning and end of conversations.
- Frame classification constrained by OCR quality causing label mismatch
- Non-verbal sounds (e.g., aggressive tones, offensive sound effects) can be hateful but are not detected by current models focused on speech/audio.
- HMM annotations are corrupted due to time interval mismatch

Conclusion

- Strong results have been achieved, demonstrating the effectiveness of our approach. The framework NoHateZone is fully functional and can be extended to support automated moderation of online videos or movies.
- Future work may consider incorporating audio features and temporal information for spatio-temporal videos

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.
[2] M. Das, R. Raj, P. Saha, B. Mathew, M. Gupta, and A. Mukherjee, “Hatemm: A multi-modal dataset for hate video classification,” 2023.
[3] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, “Exploring hate speech detection in multimodal publications,” arXiv preprint arXiv:1910.03814 [cs.CV], Oct. 2019.
[4] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, “HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection,” in Proc. AAAI Conf. Artif. Intell., vol. 35, pp. 14867–14875, Feb. 2021.