NoHateZone: A Multi-Modal Deep Learning Framework for Censoring Hateful Videos

Amene Gafsi **Mert Ülgüner** **Noah Truttmann**

Group 40

Abstract

We introduce NoHateZone, a deep learning framework for multi-modal hate speech detection. Unlike traditional systems that classify entire videos or posts as hateful or not, our approach identifies and censors specific audio segments and video frames containing hate speech. By leveraging re-annotated datasets across text, image (memes), audio, and video modalities, the system applies blurring and beeping to hateful content. We propose a framework consisting of SoTA Vision Transformer (ViT), BERT type models, and fine-tuned fusion architectures.

Keywords: Censoring, Multi-Modal, Hateful Videos, Transformers, Attention, Fine-Tuning

1. Introduction

Social media platforms are powerful tools for connection, expression, and global discourse. However, the same platforms fostering creativity and discussion can also amplify hate speech. Hate speech, defined as *"any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor."* [1], is particularly prevalent in multi-modal formats such as videos, memes, and audio clips.

Therefore, there is a need for an automated censoring mechanism which is capable of censoring the necessary parts of the hateful videos online which restricts the spread of hateful content while preserving user utility.

2. Related Work

Early efforts in hate speech detection are uni-modal and focus on hate detection only in textual data. It is a limited approach because it fails to capture hate conveyed in audio or visual channels.

To address this, researchers have shifted towards multi-modal hate speech detection. Das et al. introduced a dataset comprising hateful and non-hateful videos. Their work demonstrates that fusing modalities significantly boosts detection performance over uni-modal baselines. In their work, the fusion strategy they used was simple concatenation which limits the understanding of cross-modal relationships within the data. In our work, we use an attention-based fusion architecture for improved understanding of these patterns [2].

Building on these efforts, Koushik et al. compared fusion strategies across both video and image-text (memes) content. Their findings show that while their proposed framework detects hate well on videos, it struggles with memes. Their work highlights the need for advanced fusion mechanisms [3]. In our work, to overcome the under-performance on image-text memes, we pre-train our fusion architectures with memes data and then fine-tune on video data.

In the context of multi-modal hate detection, Prabhu and Seethalakshmi introduced a framework that combines CNNs and RNNs to process text, image, audio, and video modalities together. The architecture incorporates attention mechanisms to prioritize cross-modal signals for improved context capturing [4]. However, the framework is limited by the capabilities of CNNs, and RNNs. In our paper, we propose using a Vision Transformer (ViT) to leverage pre-trained weights and path-wise attention scores to boost contextual understanding in our fusion architectures.

The mentioned papers mainly focus on the classification side of the hate detection problem. In our paper, we also aim to develop an automated censoring framework. To the best of our knowledge, the latest work suggesting to use deep learning techniques for censoring content is by authors Yousaf and Nawaz, who developed a multi-modal architecture consisting of certain ViTs and CNNs focusing on censoring only images, disregarding other modalities [5]. Our framework has the capability to additionally censor audio, incorporate textual information on the frames, and understand cross-modal patterns.

EE-559: *Deep Learning*, 2025.

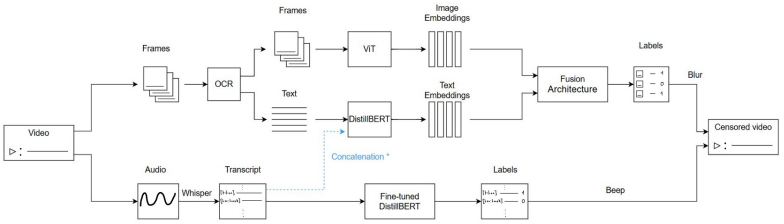

Figure 1: Main architecture

3. Method

3.1. Datasets

HateSpeechDataset: a dataset containing over 417,000 text samples, each labeled as either hateful or non-hateful [6].

HateMM (HMM): dataset including 411 hateful and 652 non-hateful videos coming from BitChute with binary labels (hate/non-hate) including the specific time intervals indicating the hateful parts [2]. As the dataset included inaccurate time intervals, we re-annotated the time intervals in 159 hateful videos containing 2247 hateful frames.

In order to adapt the data for our fusion architecture, we annotated for different modalities: Audio, Text (embedded in the frame), and Image for all videos present in the dataset with binary labels.

MMHS150K: dataset including 150,000 tweets, memes and text extracted from the memes via OCR [7].

3.2. Main Framework

To censor hateful segments from videos, we follow the process illustrated in Figure 1. The input video is first separated into its audio and visual components. For the audio, we extract the raw waveform and pass it through the *whisper-large-v3-turbo* model [8] to obtain transcriptions along with their corresponding timestamps. These transcriptions are segmented into sentences while preserving their begin and end times. Each sentence is then classified using a fine-tuned DistilBERT model [9]. If a sentence is predicted as hateful, a beep sound is overlaid on the corresponding segment to censor it.

For the visual stream, we divide the video into 1-second frames and associate each with its corresponding timestamp. Each frame is passed through *GOT-OCR2_0* [10] to extract any embedded text. The extracted text is embedded using DistilBERT, and the visual content of the frame is embedded using a Vision Transformer *(ViT)* [11]. These embeddings are then fed into a multi-modal fusion architecture to classify each frame. Frames predicted as hateful are blurred using Gaussian blur, and the final output video is reconstructed with censored audio and visual segments.

3.3. Audio-Based Sentence Classification

To detect hate in transcribed speech, we fine-tuned a DistilBERT model on the HateSpeechDataset [6]. This model serves as a binary classifier to label each transcribed sentence as hateful or non-hateful. This provides the basis for censoring decisions at the audio segment level.

3.4. Frames-Based Multi-Modal Fusion Architectures

We designed two fusion architectures for segment-level hate detection and compared their performance in this work. Each model was initially pre-trained on the *MMHS150K* dataset, which includes tweets combining text and image/meme modalities, and then fine-tuned on the *HMM* video frames to adapt the models for detecting hate content in video-based multi-modal settings.

3.4.1. Pre-Training

We follow the same training procedure for both proposed fusion frameworks, using stratified cross-validation on the *MMHS150K* dataset, with 5,000-sample validation and 10,000-sample test sets [7]. The validation set selects the best model based on F1-score from a hyperparameter grid including learning rate, weight decay, dropout rate, number of attention heads, and hidden layer dimensions. Tweet text and frame OCR text are concatenated using a [SEP] token.

3.4.2. Fine-tuning

We use the trained weights and fine-tune using the *HMM* dataset with a 80/10/10 split, repeated for 5 different seeds. Here, the validation set is used to maximize the F1-score by choosing the optimal probability threshold for binary classification (hate/non-hate). This threshold helps account for class imbalance in the test set, since we don't apply weighted sampling when computing test metrics. The threshold value can also be adjusted depending on how conservative or permissive we want the model to be in predicting hate speech.

SCMA Fusion: Our Single Cross-Modal Attention based Fusion Architecture (SCMA) uses a single cross-modal attention block in order to condition the semantic meaning of audio transcript chunks and OCR texts on video frame embeddings. The training data is constructed by concatenating the audio chunks with images and their OCR texts on the timestamps. For each audio chunk time interval, a single image is selected randomly.

In the cross-modal attention block, the image embeddings attend to text embeddings, i.e. the image embeddings are the queries, and text embeddings are both key and values. Then, the block is followed by Fully Connected (FC) layers resembling the structure of a Transformer [12] and finalized by the addition of a classification head.

DCMA Fusion: Our Dual Cross-Modal Attention based Fusion (DCMA) uses two cross-modal attention blocks to enable bidirectional semantic conditioning between modalities. The first attention block allows image embeddings to attend to OCR text representations, capturing how visual features relate to the textual context. The second attention block reverses this process, allowing OCR text embeddings to attend to image features, enriching textual understanding with visual cues.

Each attention block is followed by FC layers to learn modality-specific patterns. The resulting outputs are concatenated and passed through another FC layers, followed by a binary classification head.

4. Results

As can be seen in Table 1, we managed to exceed the F1 and AUC scores of existing models on the *HateXplain* dataset [13]. This implies that DistillBERT has word-level understanding capability even though it was trained on sentence-level samples. The robust results achieved on the test set of HateSpeechDataset demonstrate that it generalizes to the data well and understands the patterns. Ultimately, the fine-tuned model is tested on audio chunks gathered from *HMM* videos where the audio transcripts are separated according to the re-annotated time intervals. The results imply that the model can classify well enough to implement beeping.

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

Table 1: Comparing model performances on text

In Table 2, we observe that our fusion architecture, despite achieving high AUC and accuracy, yields a relatively low F1-score on the *MMHS150K* dataset. This may be due to the CNN used in the study being trained directly on this dataset which allows it to extract more context-related features, unlike our ViT and DistilBERT, which are used without task-specific fine-tuning. Fine-tuning both ViT and DistilBERT on *MMHS150K* may help improve performance.

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

Table 2: Performance of models on tweet-image/memes

Table 3 presents fine-tuning results on the *HMM* dataset. Both fusion models perform significantly better than a random baseline, capturing meaningful cross-modal relationships. DCMA outperforms SCMA, likely because text-to-image attention captures richer interactions.

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

*HMM Frames is F and HMM Frames + Audio is $F\&A$.

Table 3: Performance comparison of different models

Overall, both models show stable convergence across different seeds (2).

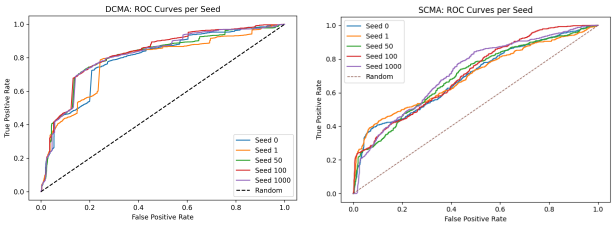

Figure 2: ROC Comparison for SCMA and DCMA

5. Limitations and Conclusion

A key limitation of our work is the lack of high-quality data. Our re-annotation of 159 videos covers a small sample, limiting generalizability. Additionally, the OCR model struggles with stylized fonts and complex visual backgrounds, leading to label mismatches across modalities. Similar issues occur when transcribing the audio as Whisper may hallucinate or assign inaccurate timestamps. Lack of temporal modeling between consecutive frames is another limitation as we treat videos as collections of independent memes and our framework ignores audio features, which can be a crucial source of information for enhancing the model's hate-detection performance. Therefore, future work should explore incorporating Mel-Frequency Cepstral Coefficients (MFCC) [14] as audio features and extend the current architecture to model temporal dependencies across frames. As the results suggest that increasing the dimensionality of the feature space by adding different modalities increases hate-detection capability of the framework, DCMA fusion is used in the main inference pipeline.

In conclusion, our framework represents a meaningful step toward sophisticated automated content moderation systems that balance freedom of expression with effective hate speech mitigation. The code used in this work is available.[1]

[1] https://github.com/Amene-Gafsi/NoHateZone

References

[1] U. Nations, "What is hate speech?."

[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] G. Koushik, D. Kanojia, and H. Treharne, "Towards a robust framework for multimodal hate detection: A study on video vs. image-based content," 02 2025.

[4] R. Prabhu and V. Seethalakshmi, "A comprehensive framework for multi-modal hate speech detection in social media using deep learning," *Scientific Reports*, vol. 15, 2025.

[5] K. Yousaf and T. Nawaz, "A deep learning-based approach for inappropriate content detection and classification of youtube videos," *IEEE Access*, vol. 10, pp. 16283–16298, Jan. 2022.

[6] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate Speech Dataset from a White Supremacy Forum," Sept. 2018.

[7] R. Gómez, J. Gibert, L. Gómez, and D. Karatzas, "Exploring hate speech detection in multimodal publications," in *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1459–1467, 2020.

[8] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, P. Welinder, M. Brundage, J. Pachocki, and W. Zaremba, "Robust speech recognition via large-scale weak supervision," 2022.

[9] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," 2020.

[10] H. Wei, C. Liu, J. Chen, J. Wang, L. Kong, Y. Xu, Z. Ge, L. Zhao, J. Sun, Y. Peng, C. Han, and X. Zhang, "General ocr theory: Towards ocr-2.0 via a unified end-to-end model," 2024.

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2021.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[13] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, "Hatexplain: A benchmark dataset for explainable hate speech detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, pp. 14867–14875, Feb. 2021.

[14] Z. K. Abdul and A. K. Al-Talabani, "Mel frequency cepstral coefficient and its applications: A review," *IEEE*, 2022.