HateMM: A Multi-Modal Dataset for Hate Video Classification

Mithun Das[1], **Rohit Raj**[1], **Punyajoy Saha**[1], **Binny Mathew**[1], **Manish Gupta**[2], **Animesh Mukherjee** [1]

[1] Indian Institute of Technology (IIT), Kharagpur
[2] Microsoft, India
{mithundas,rrohit2901,punyajoys,binnymathew}@iitkgp.ac.in,
gmanish@microsoft.com, animeshm@cse.iitkgp.ac.in

Abstract

Hate speech has become one of the most significant issues in modern society, having implications in both the online and the offline world. Due to this, hate speech research has recently gained a lot of traction. However, most of the work has primarily focused on text media with relatively little work on images and even lesser on videos. Thus, early stage automated video moderation techniques are needed to handle the videos that are being uploaded to keep the platform safe and healthy. With a view to detect and remove hateful content from the video sharing platforms, our work focuses on hate video detection using multi-modalities. To this end, we curate ~ 43 hours of videos from BitChute and manually annotate them as hate or non-hate, along with the frame spans which could explain the labelling decision. To collect the relevant videos we harnessed search keywords from hate lexicons. We observe various cues in images and audio of hateful videos. Further, we build deep learning multi-modal models to classify the hate videos and observe that using all the modalities of the videos improves the overall hate speech detection performance (accuracy=0.798, macro F1-score=0.790) by $\sim 5.7\%$ compared to the best uni-modal model in terms of macro F1 score. In summary, our work takes the first step toward understanding and modeling hateful videos on video hosting platforms such as BitChute.

Introduction

Disclaimer: The article contains material that many will find offensive or hateful; however this cannot be avoided owing to the nature of the work.

Social media platforms allow users to publish content themselves. With 82% of consumer Internet traffic expected to be video (Wilson 2022) in 2023, video hosting platforms like YouTube, Dailymotion etc. have emerged as a major source of information. On YouTube itself, people watch more than a billion hours of video every day[1]. The viral nature of such videos is a double-edged sword; on one hand it can help very quick news propagation, on the other hand it can spread hate or misinformation quickly as well. These videos cover a wide-range of topics and while most of the content on YouTube is harmless, there are videos which violate the community guidelines (O'Connor 2021). This issue is more severe for some of the alternative video hosting platforms like BitChute[2], Odysee[3] etc. While platforms like YouTube, Facebook, Twitter have strong moderation policies in place, these Alt-Tech platforms[4] allow users to post any content with little to no moderation. The non-removal of such content could be detrimental for the users and the website as a whole. It could lead to a hostile environment with echo-chambers of hateful users. It could also lead to a loss of revenue as well as attract fines (Troianovski and Schechner 2017) and lawsuits.

Some platforms employ several human moderators to find the harmful content and remove them from their site. However, given the amount of content posted daily, it is a very daunting challenge. For example, Facebook employs around 15K moderators to review content flagged by its AI and users (Koetsier 2021) and makes around 300K content moderation mistakes every day. Further, the moderators themselves are at the risk of emotional and psychological trauma (Newton 2019). This issue is further exacerbated by laws which require the platforms to remove hateful content within a fixed period of time. Failure to abide by these regulations could lead to fines (Troianovski and Schechner 2017). While platforms like YouTube have machine learning algorithms in place to detect hateful content, smaller platforms might not have the revenue/technology to develop datasets/models for hate speech detection in videos. Thus, there is a need to develop open efficient models which could detect hate speech in videos. However, the current research on hate speech is mostly focused on text based models (Badjatiya et al. 2017; Cheng et al. 2020; Juuti et al. 2020; Kennedy et al. 2020; Parikh et al. 2021; Das, Banerjee, and Mukherjee 2022) with very few image-based ones (Yang et al. 2019; Das, Wahi, and Li 2020; Gomez et al. 2020; Kiela et al. 2020). Detecting hateful actions in videos needs leveraging a combination of multi-frame video processing and speech processing signals, and thus image-based hate detection methods cannot be directly adapted.

Research objectives and contributions: In this paper, we take a step toward an end-to-end solution for this novel problem setting. We release HateMM, a collection of videos

[1] https://www.youtube.com/intl/en-GB/about/press/

[2] https://www.bitchute.com/

[3] https://odysee.com

[4] https://en.wikipedia.org/wiki/Alt-tech

annotated for hate speech. The dataset contains ∼ 43 hours of videos composed of a total of ∼ 144K frames. We make the HATEMM dataset public[5] to promote further research in multi-modal hate speech detection. We rely on BitChute for our data collection as it has low content moderation. Launched in 2017, BitChute serves as a video hosting and sharing platform similar to YouTube and is quite popular among far-right users.

Overall, we make the following **contributions**.

- We curate one of the largest known datasets of hateful videos consisting of 1083 videos spanning ∼ 43 hours and ∼ 144K frames. Each individual video was annotated as hateful or not, along with the frame spans which justify the labelling decision. The average time taken by the annotators to label a single video was approximately twice the video duration.
- We develop detection models using three different modalities (text, audio and video) individually as well as jointly[6]. Our best fusion model (BERT ⊙ ViT ⊙ MFCC) which combines all the modalities attains a macro F1-Score of 0.790. The precision and the recall for the hate class are 0.742 and 0.758 respectively. Among the individual modalities, transformer encodings of text and video-based features seem to be more effective for detecting hateful videos.
- We further perform some preliminary analysis of the importance of each of the modalities. We observe that the text based model is successful when the transcript is relatively clean. The audio based model is most effective when there is shouting and expression of aggression in the video. Finally, the vision based model works the best if there is evidence of visual hateful activity with presence of victim in the video.
- As a last step, we analyze the performance based on the frame spans and observe that the text-based and vision-based models can leverage this information the best. Besides, the vision-based model performs the best in case the hate target in the video are 'Blacks' or 'Jews', while the text-based model does very well on the 'Other' target communities.

Related Work

With the huge availability of multi-modal data, multi-modal deep learning has been harnessed to improve the accuracy for various tasks like visual question answering (Singh et al. 2019), fake news/rumour detection (Khattar et al. 2019), etc. Recently, multi-modal hate speech detection has become popular where text posts are combined with extra contexts like user and network information (Cheng et al. 2020; Founta et al. 2019) or images (Yang et al. 2019; Das, Wahi, and Li 2020; Gomez et al. 2020; Kiela et al. 2020) to improve detection accuracy. Such multi-modal schemes typically use unimodal methods like CNNs, LSTMs or BERT to encode text and deep CNNs like ResNet or InceptionV3 to encode images, and then perform multi-modal fusion using simple concatenation, gated summation, bilinear transformation, or attention-based methods. Multi-modal bitransformers like ViLBERT and Visual BERT have also been applied (Kiela et al. 2020).

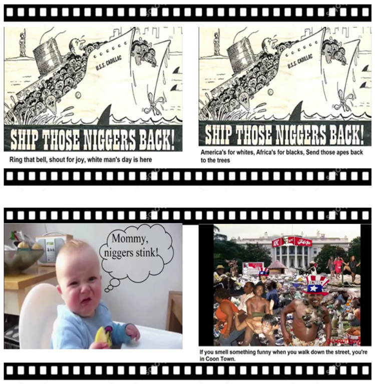
U.S.S. CADILLAC
SHIP THOSE NIGGERS BACK!
Ring that bell, shout for joy, white man's day is here
U.S.S. CADILLAC
SHIP THOSE NIGGERS BACK!
America's for whites, Africa's for blacks, Send those apes back to the trees
Mommy, niggers stink!
If you smell something funny when you walk down the street, you're in Coon Town.

Figure 1: Examples of hate videos.

There is almost no work on the detection of offensive/hate videos barring the following three – for Portuguese (Alcântara, Moreira, and Feijo 2020) and English (Wu and Bhandary 2020; Rana and Jha 2022). Nevertheless, the first two works (Alcântara, Moreira, and Feijo 2020; Wu and Bhandary 2020) only consider textual features for their classification purpose by extracting the transcript. Further, the size of the annotated dataset is less than 500 videos. The work done by Rana et al. (2022) considered both textual and audio features, though the dataset is not publicly available, the data curation and annotation steps are not fully described and the dataset statistics are not precisely revealed. Unlike our dataset, they choose videos where the speech is clear; in contrast, we did not have any such constraint since hateful content can be as well expressed in only visual form without having any associated speech. To the best of our knowledge, we are the first to experiment with multi-modal hate video detection, where we leverage all three data modalities – text, audio, and video. Our annotation is far richer and larger compared to the state-of-the-art in order to appropriately leverage all the modes. We believe that our dataset and the benchmark models trained on it will help the moderators identify genuine hateful cases while reducing false alarms.

HATEMM Dataset

The BitChute Platform

BitChute is a social video-hosting platform with low content moderation launched as an alternative to YouTube (Trujillo et al. 2020). The website launched in 2017 is gaining popularity and is becoming a "haven" for far-right users. BitChute has high prevalence of hateful content and hosts several content producers who were banned from traditional and moderated platforms (Labarbera 2020).

[5] https://doi.org/10.5281/zenodo.7799469

[6] The source code of the baseline models is available at https://github.com/hate-alert/HateMM

Data Collection

To sample the videos for annotation we used lexicons from (Mathew et al. 2020a) that studied Gab and other alt-right platforms. These lexicons consist of derogatory keywords/slurs targeting different protected communities.

Each of the keywords is used to search on BitChute; the links returned are added to a database. In total, we collected $\sim 8K$ links. Next we download the videos using BitChute-dl software[7]. While downloading we did not find the videos for 25% of the links. Further few videos were corrupted as well. Finally we end up with $\sim 6K$ videos.

Annotation Guidelines

The labeling scheme stated below constitute the main guidelines for the annotators, while a codebook ensured common understanding of the label descriptions. We construct our codebook (which consists the annotation guidelines) for identifying hateful content on the YouTube policy of hate speech[8]. We consider a video as hateful if –

> *"It promotes discrimination or disparages or humiliates an individual or group of people on the basis of the race, ethnicity, or ethnic origin, nationality, religion, disability, age, veteran status, sexual orientation, gender identity etc."*

In addition, we also ask the annotators to mark the parts (i.e., frame spans) of a hate video which they felt are hateful (as rationales) and the communities the video targets. We believe that the rationales can later serve as an explainability signal and targets can be used to measure if the detection algorithms are getting biased toward some targets in the lines of what has been presented in (Mathew et al. 2020b).

Annotation Process

Training the annotators The annotation process was led by two PhD students as expert annotators and performed by four under-graduate students who were novice annotators. All the undergraduate students are computer science majors, All of them participated voluntarily for the task with complete consent and were rewarded through an online gift card at the end of the task. Both the expert annotators had experience in working with harmful content in social media. In order to train the annotators we needed a pilot gold tagged dataset. To this end, the expert annotators initially annotated 30 videos. The initial set consisted of 20 hate videos and 10 non-hate videos. We gave these 30 videos to the undergraduate annotators who annotated these based on the annotation codebook. Once they finished their annotations we discussed the incorrect cases with them to improve their annotation skills.

Annotations in batch mode Subsequent to the above training, we released a set of 30 videos per week in a batch mode. Being aware that while annotating hate videos, annotators can have "negative psychological effects" (Ybarra et al. 2006), we advised them to take at least 10 minutes break after the annotation of each video. We further imposed an additional constraint that no more than 10 videos should be annotated per day. Finally, we also had regular meetings with them to ensure the annotations did not have any adverse effect on their mental health.

	Hate	Non Hate	Total
Count	431 (39.8%)	652 (60.2%)	1083
Total len (hrs)	18.39	24.87	43.26
μ **video len**	2.56 ± 1.69	2.28 ± 4.77	2.40 ± 3.86
μ **rationale len**	1.71 ± 1.27	-	-
μ **#frames**	154	137	144
μ **#words**	228	209	217

Table 1: Basic statistics of the HATEMM dataset. Frames were sampled per second. Video and rationale length are in minutes. len: length. hrs: hours. μ : Mean.

Annotation tool Off-the-shelf tools like Toloka[9] and ANVIL (Kipp 2014) allow annotations for tasks like object detection, but they do not support annotations of any kind of spans in videos. PAVS[10] only allows span selection but fortunately it is open-source. Consequently, we modified PAVS to support span annotations, hate or not video annotation, and target community labeling. We shall make our annotation tool (see a snapshot of the tool in Figure 2) public to facilitate further research.

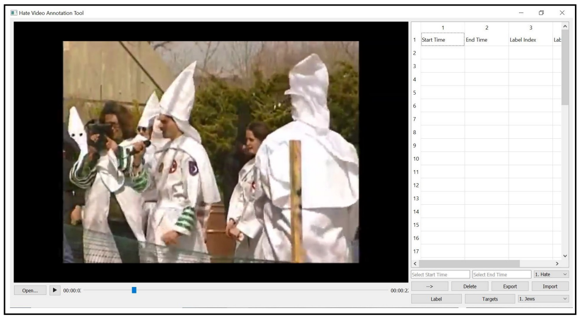

Figure 2: Snapshot of the (hate) video annotation tool.

Each video was annotated by two independent annotators. They were instructed to watch the complete video and based on the guidelines provided, select the appropriate class (hate or non hate). Average time required by the annotators to annotate a video was approximately *twice* the video duration. The Cohen's kappa for the inter-annotator agreement was κ=0.625. On completion of each batch of annotation, if there was a mismatch between the two annotators, one of the expert annotators annotated the same video to break the tie. This yielded a final dataset of 431 hate and 652 non-hate videos and constitutes our set of a total of 1083 labelled instances.

[7] https://pypi.org/project/bitchute-dl/

[8] https://support.google.com/youtube/answer/2801939?hl=en

[9] https://toloka.ai/ml/computer-vision

[10] https://github.com/kevalvc/Python-Annotator-for-VideoS

Dataset Statistics

Our final dataset contains 1083 videos spanning over $\sim$ 43 hours of content. On average the videos are $\sim$ 2.40 mins in length with hate videos being slightly longer at an average of $\sim$ 2.56 mins. Overall we are able to curate a roughly balanced dataset with 39.8% of the samples labelled as hate. The class balance is better than many of the textual hate speech datasets. For each video, we also get the audio transcribed using Vosk offline speech recognition[11] tool. There are $\sim$ 217 words in the transcripts on average. Further, the number of words in the transcripts of hate videos are slightly higher ($\sim$ 228) compared to transcripts of non-hate videos. The other important statistics of the dataset are noted in Table 1. Snapshots of some examples hate videos are shown in Figure 1.

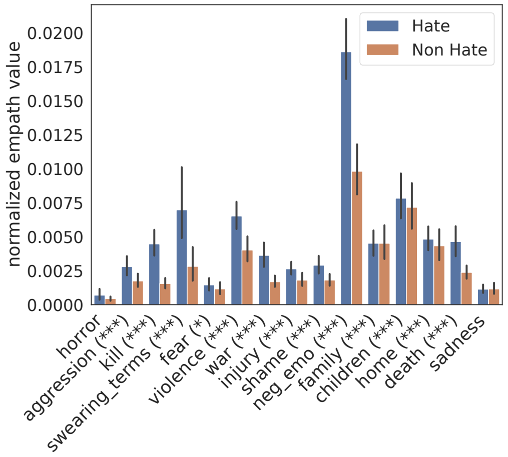

Figure 3: Lexical analysis of video transcripts using Empath. We report the mean values for several categories of Empath. Hate video transcripts scored significantly high in categories like 'aggression', 'swearing terms', 'violence' and 'negative emotion'. For each category, we use the Mann-Whitney U test and show the significance levels ***($p < 0.0001$), **($p < 0.001$), *($p < 0.01$).

Dataset Analysis

Empath analysis In order to understand the dataset better, we identify important lexical categories present in the video transcripts using Empath (Fast, Chen, and Bernstein 2016), which has 189 such pre-built categories. First, we select 70 categories ignoring the irrelevant topics to hate speech, e.g., technology and entertainment. We report the top 15 significantly different categories in Figure 3. Hate video transcripts scored significantly high in categories like 'aggression', 'swearing terms', 'violence' and 'negative emotion'.

OOV words To understand the extent of noise in the dataset (and the transcript quality), we calculate the percentage of Out-of-Vocabulary (OOV) words present in both the 'hate' and the 'non-hate' classes. For this purpose, we use the PyEnchant dictionary[12], which is Python's spellchecking dictionary, to identify the words that are not present in the standard English library. Figure 4 shows the % OOV words per video (transcript) for both the 'hate' and the 'non-hate' classes. The plot shows that for both the 'hate' and the 'non-hate' classes, the mean percentage of OOV words is almost 22%.

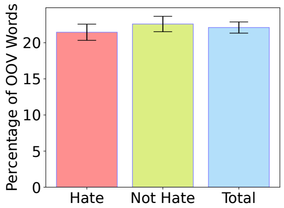

Figure 4: Percentage of OOV words present in the dataset.

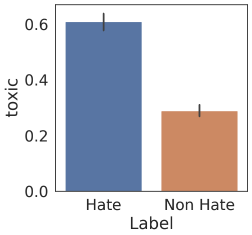

Figure 5: Toxicity comparison based on Perspective API. The results are significant at $p < 0.0001$ based on Mann-Whitney U test.

Toxicity score One easy solution to detect toxic videos could be to use Google's Perspective API[13] on the transcript; hence we measured the toxicity of the transcripts. As shown in Figure 5, hate video transcripts have almost twice the toxicity ($\sim$ 0.61) compared to the non-hate video transcripts ($\sim$ 0.28). However, relying on transcripts has its own drawbacks. As discussed above, the transcripts, in general, are quite noisy, and this indicates why transcripts alone might not be sufficient for hate video classification.

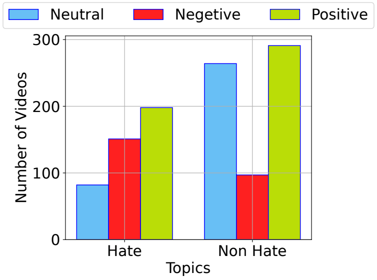

Figure 6: Number of posts having different sentiments in hate and non-hate category.

[11] https://alphacephei.com/vosk/

[12] https://github.com/pyenchant/pyenchant

[13] https://www.perspectiveapi.com/

Hate	Non Hate
k k k k	joe
k k k	joe rogan
k k	george zimmerman
uncle sam	joe biden
jack	bush
joe	chris
klan	mug
robert hours	mike
max	eric
k k k k k	johnson

Table 2: Most frequent PERSON entities in Hate and Non Hate classes

Sentiments We also measure the sentiment associated with the videos that we have annotated. Sentiment analysis is used to identify the associated feelings/emotions within a text. Sentiment analysis includes three types of polarity: negative, neutral, and positive. In this study, the word-based method was used and the polarity of each transcript was determined by the score from -1 to 1 according to the word used. A negative score means a negative sentiment, and a positive score means a positive sentiment. Sentiment analysis was carried out using TextBlob API[14]. Figure 6 represents the number of videos having neutral, negative and positive sentiment for both video categories. We observe that though overall videos with positive sentiment are more, 'hate' video transcripts have more negative sentiment compared to 'non-hate' videos.

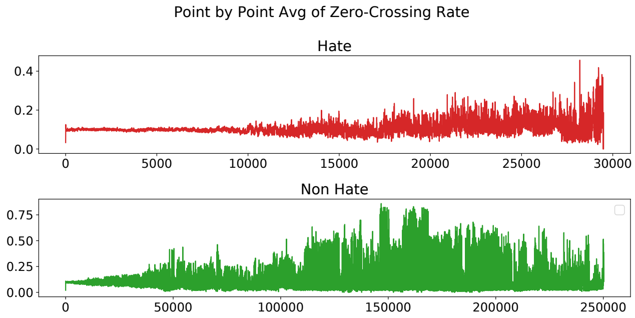

Figure 7: Zero Crossing rate for the hate and non-hate videos.

NER analysis We also analyzed the named entities associated with the transcript to find out if the distribution of different NER tags is different for hate and non hate classes. To this purpose, we use the spacy library [15], which provides a set of entity tags. We observed that for the PERSON tag, the normalized number of named entities are more in the hate class than the non hate class. We further inspected the most frequent entities associated with the entity type PERSON which are noted in Table 2. For the hate class, phrases like 'k k k' are very common; this possibly corresponds to the KKK[16], an American white supremacist terrorist and hate group whose primary targets are African Americans, Jews, etc.

Audio analysis We also analyze the audio signal associated with each video. Specifically, we calculate the Zero-Crossing Rate (ZCR), Spectral Bandwidth, Root Mean Square (RMS) Energy and report the mean for all the audios.

In Figure 7 we plot the time series of the ZCR averaged over all audio files in the two respective classes. We observe that the plots are distinctly different for the hate and the non hate classes. It is well known that ZCR can be interpreted as a measure of the noisiness of a signal, and higher values indicate more noisiness of the audio signal. This indicates that while the noise level is roughly uniformly spread over the whole time series for the non hate videos, for the hate videos this is predominantly flat and only appear to go up toward the end of the time series indicating that the hate videos are possibly crafted to have better quality audio signal. The same results are also observed for the Spectral Bandwidth (data not produced for brevity).

In Figure 8 we present the RMS Energy plot for the hate and the non hate videos averaged over all the audio files in the respective classes. RMS energy is helpful in estimating the average loudness of an audio track. We observe that hate videos are louder only in the initial part of the time series unlike for non hate videos. A manual inspection showed that in many hate video there are instances of shouting which possibly manifests in the form of high loudness.

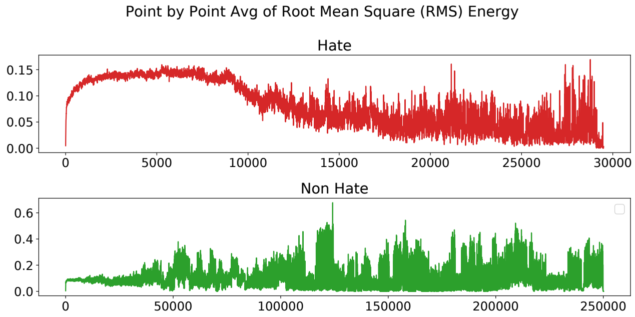

Figure 8: Root Mean Square (RMS) Energy of Hate and non-hate videos.

Video analysis We further attempt to analyze what kind of objects are mainly present in the videos. To this purpose, we use the ImageAI [17] object detection package. For each hate and non-hate video, we randomly select 20 frames and extract all the objects associated with the frames. We assume if an object has been seen in any frame of a given video, it would mean that the object is present in that video. We observe that 45% of the time, the object "person" appears in hateful videos, whereas 59% of the time, the object "person"

[14] https://textblob.readthedocs.io/en/dev/

[15] https://spacy.io/

[16] https://en.wikipedia.org/wiki/Ku_Klux_Klan

[17] https://imageai.readthedocs.io/

appears in non hateful videos. In the hateful videos we observe use of important religious persons like a Jewish rabbi at the background with a lot of hateful text embedded on them. Similarly the black persons detected are often associated with dirtiness and food mongering. Further, for the hate videos, we observe 'stop sign' and certain play items like 'teddy bear', 'kite' and 'sports ball'. A manual inspection shows that these play items are mostly 'cartoon-ish' figures used to mock a target community.

Overall, in this section we observe that all three modalities – text, audio and video have certain latent indicators that should be helpful in differentiating the hate from the non-hate class of videos. This observation, as we shall see, is corroborated by the superior performance of the joint model as observed in section .

Methodology

This section discusses the pre-processing steps and models we implemented for hate video detection.

Problem Formulation

We formulate the hate video detection problem in this paper as follows. Given a video $\boldsymbol{V}$, the task can be represented as a binary classification problem. Each video is to be classified as hate ($y = 1$) or non-hate ($y = 0$). A video $\boldsymbol{V}$ can be expressed as a sequence of frames, i.e., $\boldsymbol{F} = \{f1, f2, .., fn\}$, the associated audio $\boldsymbol{A}$ and the extracted video transcript $\boldsymbol{T} = \{w_1, w_2, ..., w_m\}$, consisting of a sequence of words. We aim to learn such a hate video classifier $Z : Z(\boldsymbol{F}; \boldsymbol{A}; \boldsymbol{T}) \rightarrow y$, where $y \in \{0, 1\}$ is the ground-truth label of a video.

Pre-processing

We remove numbers and special characters from the transcripts and perform text normalization wherever required. For vision-based models, we first sample the video at one frame-per-second and sample 100 such frames for each video. For videos with less than 100 frames, we add an image with white background as a padding. For the videos having more than 100 frames, we uniformly sample 100 frames from the total number of available frames.

Text-Based Models

fastText: We obtain 300 dimensional fastText (Grave et al. 2018) embedding of all the video transcripts, pass it through two dense layers of 128 nodes, and finally pass it to the output node for the final prediction. We name this model as **T1**.

LASER: We obtain 1024 dimensional LASER (Artetxe and Schwenk 2019) embedding of all the video transcripts, pass it through two dense layers of 128 nodes, and finally provide it to the output node for the final prediction. We name this model as **T2**.

BERT: We use BERT (Devlin et al. 2018) since it is known to be highly effective for many text classification tasks, including text-based hate speech detection. For each transcript, we get the CLS embedding and pass it through two dense layers of 128 nodes, and finally provide it to the output node for the final prediction. We call this model as **T3**.

HateXPlain: We also experiment with another BERT model (Mathew et al. 2020b). The model is already finetuned on pre-trained BERT using English hate speech data. Since this model has been already finetuned on hate speech data, we expect that it should yield better performance. We denote this model as **T4**.

Audio-Based Models

MFCC: One of the popular methods for representing audio is the Mel Frequency Cepstral Coefficient (MFCC) (Xu et al. 2004) which has been found to be effective for complex tasks like lung sound classification (Jung et al. 2021) and speaker identification (Kalia et al. 2020). We obtain a representation of the audio of our dataset using the MFCC features. To generate the MFCC features, we use the open-source package - Librosa[18] and construct a 40 dimensional vector to represent the audio. These vectors are passed through three fully connected layers to generate the final label. We refer to this model as **A1**.

AudioVGG19: We also use waveforms of audios (extracted from the videos) and generate 1000 dimensional feature vectors by using a pre-trained VGG-19 model (Simonyan and Zisserman 2015; Grinstein et al. 2018). Similar to **A1**, these vectors are passed through three fully connected layers to generate the final label. We call this model **A2**.

Vision-Based Models

In order to handle the spatial and temporal information in the videos, we consider several vision-based classification models such as 3D-CNN, InceptionV3, Vision Transformer, etc.

3D-CNN: The 3D-CNN (Ji et al. 2013) model contains two Conv3D and BatchNorm3D layers. After these layers we add ReLU, dropout and maxpool layers to generate the final representation. This is further passed through three fully connected layers to generate the final label (please see experimental setup for further details on the layer sizes). We name this model as **V1**.

InceptionV3: We also construct feature vectors by using pre-trained InceptionV3 model (Szegedy et al. 2015). We extract a 1000 dimensional feature vector for all the 100 frames and then pass it through an LSTM (Hochreiter and Schmidhuber 1997) network, which is finally fed to the output node for classification. We use LSTM to capture the sequential nature of the video frames. We name this model as **V2**.

Vision Transformer: In this approach, the image is divided into a sequence of patches and then fed to the a transformer model. Like BERT, the extra learnable [class] token is also prepended with the sequence of patches for the classification task. As our focus is to detect hateful videos, so we cannot use Vision Transformer directly. Like the **InceptionV3** model, we take 100 frames for each video and pass it through the pre-trained Vision Transformer(ViT) (Dosovitskiy et al. 2020) model to get a 768 dimensional feature vector for each frame and finally pass it through the LSTM network to obtain the prediction. We refer to this model as **V3**.

[18] https://librosa.org/doc/latest/index.html

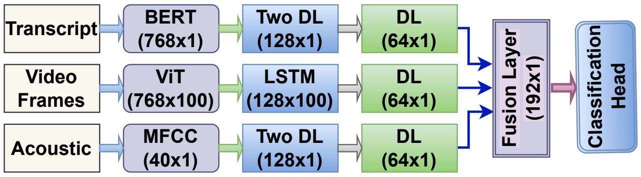

Figure 9: A schematic of the multi-modal model. DL: Dense Layer.

Multi-Modal Hate Video Detection

The models discussed in the previous subsections are incapable of leveraging the relationship among the features extracted through different modalities (i.e., video, text transcript, and audio). To capture the benefits of all the modalities, we attempt to meaningfully combine the text, audio, and vision-based models. In particular we build the following models – **M1** (BERT $\odot$ ViT $\odot$ MFCC), **M2** (BERT $\odot$ ViT $\odot$ AudioVGG19), **M3** (HateXPlain $\odot$ ViT $\odot$ MFCC) and **M4** (HateXPlain $\odot$ ViT $\odot$ AudioVGG19). $\odot$ refers to the combination operation of the three modalities through a trainable neural network (aka fusion layer). Figure 9 illustrates the overall modeling pipeline.

Experiments and Results

Experimental Setup

We evaluate our models using k-fold stratified cross-validation, which is beneficial in assessing models having less labeled data. For all the experiments, we set k to 5 here, and for each fold, we use 70% data for training, 10% for validation, and the rest 20% for testing. We use the same test sets across all the models to ensure a fair comparison. For all the unimodal neural network models, the internal layer has two fully connected layers of 128 nodes, reduced to a feature vector of length 2. For the uni-modal LSTM based models, all the frame embeddings are passed to an LSTM network with hidden size of 128, which is finally reduced to a feature vector of length 2. For the 3D-CNN, each frame has been resized to 100 × 125. Both the Conv3D layers have 256 nodes, and the number of channels are 32 in the first layer, and 42 in the second layer. The kernel size for the first layer is $(5, 5, 5)$, and the second layer is $(3, 3, 3)$. Further, we use a stride of $(2, 2, 2)$ with zero padding for both the layers, resulting in a feature vector of length 2. We pass the final feature vector through a log-softmax layer with negative log-likelihood loss. This gives the probability of whether the video is hateful or not. For the fusion models, the text-based and audio based features are passed to two dense layers of size 128, which are finally fed to another dense layer of 64 nodes; the extracted vision-based features are passed to an LSTM network with a hidden size of 128, which is further passed to another dense layer of 64 nodes. Finally we concatenate all the nodes and reduce to a feature vector of length 2 as shown in Figure 9. All the models are run for 20 epochs with Adam optimizer, batch_size = 10, learning_rate = $1e-4$. We store the results at the best validation score in terms of macro-F1 score. All models are coded in Python, using the Pytorch library.

Evaluation Metric

To remain consistent with the existing literature, we evaluate our models in terms of the standard metrics – accuracy, F1 score, precision, and recall. Together, these metrics should be able to thoroughly assess the classification performance of the models in distinguishing between the two classes – hate vs non-hate. The best result is marked in **bold**, and the second best is underlined.

Results

Performance across different models Table 3[Left side] shows the performance of each model. We observe among all the text-based models, the transformer-based models perform the best, especially the **HateXPlain** model, which is earlier fine-tuned on a hate speech dataset. Among the audio-based models, we see **AudioVGG19** performs better than **MFCC**, though, in terms of macro-F1 score, the difference between these two models is marginal. For the vision-based models, we see the features extracted from **ViT** are very helpful in detecting hate videos among all other vision-based models. Further, we find that all the multi-modal models outperform all the unimodal models (in terms of accuracy, F1 score), and **BERT $\odot$ ViT $\odot$ MFCC** performs the best among all the models.

Performance based on video length We divide all the test datasets across all the folds in terms of video length. Empirically, we have the lower bucket with video length ≤ 105 secs and the rest in the higher bucket to have almost the same number of videos across the two buckets. In Table 3[Right Side] we report the macro F1 for both buckets. We observe that among the text-based models, except **T2**, all others perform better in the higher bucket. All the audio-based models, perform better in the higher bucket. On the other hand, except **V2**, all the vision-based models perform better in the lower bucket compared to their respective higher buckets. This is indicative that the vision-based model captures context well when the video duration is less. When the text, audio, and vision-based models are integrated together, as expected the performance improved in both buckets.

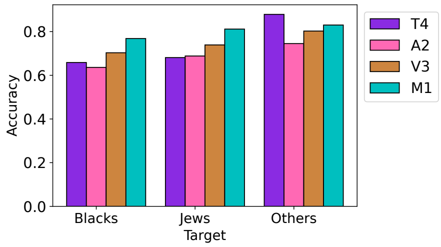

Figure 10: Target-wise performance. Only the best performing model for each modality and the best ensemble model are shown.

Target-wise performance We also compute the target wise performance of the best uni-modal and fusion models

Model	**Architecture**	**Acc**	**M-F1**	**F1 (H)**	**P (H)**	**R (H)**	**VL <= 105 sec (542 videos)**	**VL >105 sec (541 videos)**
T1	**fastText**	0.687	0.673	0.609	0.611	0.614	0.609	0.700
T2	**LASER**	0.730	0.720	0.668	0.655	0.686	0.675	0.655
T3	**BERT**	0.735	0.722	0.664	0.675	0.667	0.672	0.708
T4	**HXP**	0.757	0.733	0.653	**0.753**	0.577	0.698	0.727
A1	**MFCC**	0.675	0.665	0.622	0.593	0.679	0.603	0.687
A2	**AVGG19**	0.690	0.669	0.589	0.629	0.559	0.583	0.669
V1	**3D-CNN**	0.674	0.653	0.571	0.619	0.547	0.637	0.587
V2	**InceptionV3**	0.720	0.706	0.643	0.653	0.637	0.672	0.707
V3	**ViT**	0.748	0.733	0.672	0.695	0.656	0.718	0.703
M1	**BERT ⊙ ViT ⊙ MFCC**	**0.798**	**0.790**	**0.749**	0.742	**0.758**	**0.772**	**0.759**
M2	**BERT ⊙ ViT ⊙ AVGG19**	0.755	0.765	0.718	0.723	0.719	0.743	0.733
M3	**HXP ⊙ ViT ⊙ MFCC**	0.777	0.767	0.720	0.718	0.726	0.744	0.741
M4	**HXP⊙ ViT ⊙ AVGG19**	0.767	0.756	0.707	0.714	0.712	0.733	0.731

Table 3: [Left side] Model performance on the task of classification of hate videos. [Right Side] Macro F1 Score with respect to video length (VL) in secs. H: hate class, Acc: accuracy, M-F1: macro-F1, P: precision, R: recall, HXP: HateXplain.

Video name	**Description**	**Mode**	**Explanation**
Terrorist Jew Hates Hollywood Traitor Kikes [ID=g11ysqwzlKj6]	In this video, a person is seated and abusing Jews saying derogatory words like kikes.	Text	The video was unrelated to hate speech, and the transcript was clean; the audio-based model also failed due to the absence of high aggressiveness in the voice.
Grinded Nig Freezer Full Of Ni**er Heads [ID=lHk2E8MNU5HB]	Here some group was yelling some type of song containing slur words like "ni**er" with nazi flag visible in the background	Audio + Video	The transcript was erroneous. Other modalities found useful signals based on nazi flag (video) and yelling (audio).
When Youre In Coon Town [ID=OngXc0A4DXxo]	A song is yelled as a part of the audio about "What is a c**n town?" In the video, irrelevant images are shown.	Audio + Text	The audio models were able to capture it due to the presence of yelling. The transcript has derogatory words as signals. The video was fairly unrelated.
A Filthy Jew Straight From Hell Short Film [ID=uSH9Z7tEj9vp]	In this video a person dressed in Nazi attire is abusing Jews people.	Video + Text	The video contained some hateful symbolism toward Jews and derogatory keywords were identified in the transcript, so both text and vision models succeeded.

Table 4: Examples of a few hate videos along with their description. We also mention the modality/ies which could predict the hate correctly in the *Mode* column. In addition, we also provide a possible explanation for this prediction.

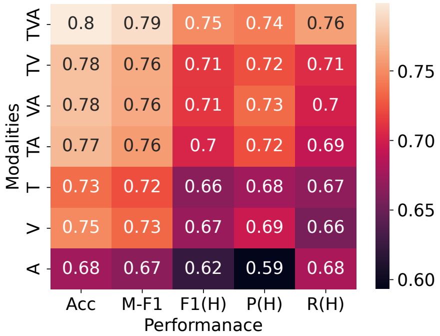

Figure 11: Heatmap of the performance of the different modalities. T: Textual, V: Vision, A: Audio.

and show the results in Figure 10. We observe that among the uni-modal models the vision based model **V3** performs better for the videos targeting 'Blacks' and 'Jews', whereas the text-based model **T4** performs the best for the other targets (taken all together). Further, we notice that integrating these models, **M1**, gives consistently good performance across all the target communities.

Effectiveness of the Modalities

To understand how the different modalities contribute to the prediction task, we perform ablation studies to demonstrate the effectiveness of all the modalities. We select our best multi-modal model **M1**, which utilizes all the features of a video for predicting the labels of the videos. We remove the modalities one at a time and train our models. We illustrate our result in Figure 11 using a heatmap. We observe jointly training using all the modalities brings the highest performance. With the removal of at least one modality, performance drops by around 2-3%. Further, with the removal of two modalities, we observe that the performance drops drastically. Overall, we find that the audio-based feature is successful when there was shouting or aggression in the voice present in the video because the MFCC features can capture these sound effects present in the audio. For example, one of the videos which showed a KKK[16] member shouting derogatory words is identified as hate by the audio but not by the other two modalities. The text-based model's performance depends on the accuracy of the automatic speech recognition (ASR). This model is successful most of the time when the ASR can correctly detect the hateful words in the transcript. Finally, the vision-based model is successful when it contains victims present in the video itself. There were some unrelated images in a few videos, like some game-play while the audio was derogatory. In such cases, the vision-based model fails due to the lack of useful signals. This also justifies the need for fusion models. We show a few examples of such videos in Table 4.

Conclusions and Future Work

This paper takes a step toward identifying hateful content in videos by leveraging signals across all three modes. To achieve this, we crawled videos from the BitChute platform and manually annotated them as hate and non-hate. Analyzing the annotated dataset HATEMM revealed interesting aspects about the hate videos. We utilized all the modalities of the video to detect whether it is hateful or not. We showed that models which take multiple modalities into account performed better compared to the uni-modal variants. We also performed a preliminary analysis to understand how different modalities contribute to the prediction. We found that text-based model performs well when the transcript is clean, the audio-based model is successful when there is shouting or aggression in the video, and the vision-based model is able to capture the hateful content when hateful activities or the target of the abuse are present in the video.

In future we plan to use other vision and speech transformers such as ViViT (Arnab et al. 2021), Wav2Vec (Baevski et al. 2020), etc. which can possibly further boost the classification performance. One of the hardships here however is that we need much larger-sized videos to be annotated in order to train and fine-tune such data-hungry models. Besides, we plan to annotate videos of longer length. We also envisage to build models which not only would detect a video as hateful but also identify the sections of the video which made it hateful. Thus, instead of looking into the full video, a moderator can watch the portions of the videos, which has been marked as hateful by the model and, subsequently, decide for moderation actions.

Ethical Statement

Ethical Considerations

Our database constitutes videos with labeled annotations and does not include any personally identifiable information about any user or the Bitchute channel where the videos have been uploaded. We only analyzed publicly available data. We followed standard ethical guidelines (Rivers and Lewis 2014), not making any attempts to track users across sites or deanonymize them. Since the video used in our analysis contain hateful elements, care should be taken to not use it for negative purposes like spreading further hatred or maligning an individual or a community.

Biases

Any biases noticed in the dataset are unintentional, and our intention is not bring any individual or a target community to harm. We believe it can be subjective to determine if a video is hateful or not; thus, biases in our gold-labeled data or label distribution are inevitable. Nonetheless, we are confident that the label given to the data is most accurate due to the significant inter-annotator agreement we have achieved.

Intended Use

We share our data to encourage more research on hate video classification. We only release the dataset for research purposes and do not grant a license for commercial or malicious use.

References

Alcântara, C.; Moreira, V.; and Feijo, D. 2020. Offensive video detection: dataset and baseline results. In *LREC*, 4309–4319.

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. *arXiv preprint arXiv:2103.15691*.

Artetxe, M.; and Schwenk, H. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7: 597–610.

Badjatiya, P.; Gupta, S.; Gupta, M.; and Varma, V. 2017. Deep learning for hate speech detection in tweets. In *WWW*, 759–760.

Baevski, A.; Zhou, Y.; Mohamed, A.; and Auli, M. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. *Advances in Neural Information Processing Systems*, 33.

Cheng, L.; Shu, K.; Wu, S.; Silva, Y. N.; Hall, D. L.; and Liu, H. 2020. Unsupervised Cyberbullying Detection via Time-Informed Gaussian Mixture Model. In *CIKM*, 185–194.

Das, A.; Wahi, J. S.; and Li, S. 2020. Detecting Hate Speech in Multi-modal Memes. *arXiv preprint arXiv:2012.14891*.

Das, M.; Banerjee, S.; and Mukherjee, A. 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*, 32–42.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Fast, E.; Chen, B.; and Bernstein, M. S. 2016. Empath: Understanding topic signals in large-scale text. In *Proc. of the 2016 CHI Conference on Human Factors in Computing Systems*, 4647–4657.

Founta, A. M.; Chatzakou, D.; Kourtellis, N.; Blackburn, J.; Vakali, A.; and Leontiadis, I. 2019. A unified deep learning architecture for abuse detection. In *WebSci*, 105–114.

Gomez, R.; Gibert, J.; Gomez, L.; and Karatzas, D. 2020. Exploring hate speech detection in multimodal publications. In *WACV*, 1470–1478.

Grave, E.; Bojanowski, P.; Gupta, P.; Joulin, A.; and Mikolov, T. 2018. Learning Word Vectors for 157 Languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Grinstein, E.; Duong, N. Q. K.; Ozerov, A.; and Pérez, P. 2018. Audio Style Transfer. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 586–590.

Hochreiter, S.; and Schmidhuber, J. 1997. Long Short-Term Memory. *Neural Comput.*, 9(8): 1735–1780.

Ji, S.; Xu, W.; Yang, M.; and Yu, K. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1): 221–231.

Jung, S.-Y.; Liao, C.-H.; Wu, Y.-S.; Yuan, S.-M.; and Sun, C.-T. 2021. Efficiently Classifying Lung Sounds through Depthwise Separable CNN Models with Fused STFT and MFCC Features. *Diagnostics*, 11(4): 732.

Juuti, M.; Gröndahl, T.; Flanagan, A.; and Asokan, N. 2020. A little goes a long way: Improving toxic language classification despite data scarcity. In *EMNLP: Findings*, 2991–3009.

Kalia, A.; Sharma, S.; Pandey, S. K.; Jadoun, V. K.; and Das, M. 2020. Comparative analysis of speaker recognition system based on voice activity detection technique, MFCC and PLP features. In *Intelligent Computing Techniques for Smart Energy Systems*, 781–787. Springer.

Kennedy, B.; Jin, X.; Davani, A. M.; Dehghani, M.; and Ren, X. 2020. Contextualizing Hate Speech Classifiers with Post-hoc Explanation. In *ACL*, 5435–5442.

Khattar, D.; Goud, J. S.; Gupta, M.; and Varma, V. 2019. MVAE: Multimodal variational autoencoder for fake news detection. In *WWW*, 2915–2921.

Kiela, D.; Firooz, H.; Mohan, A.; Goswami, V.; Singh, A.; Ringshia, P.; and Testuggine, D. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *NIPS*, 33.

Kipp, M. 2014. *ANVIL: The video annotation research tool*, 420–436. ISBN 9780199571932.

Koetsier, J. 2021. Report: Facebook makes 300,000 content moderation mistakes every day. https://www.forbes.com/sites/johnkoetsier/2020/06/09/300000-facebook-content-moderation-mistakes-daily-report-says/. Accessed: 2023-04-01.

Labarbera, G. 2020. Deplatforming and the rise of the alt-tech video hosting platform BitChute in 2020.: Masters of Media. http://mastersofmedia.hum.uva.nl/blog/2020/09/27/deplatforming-and-bitchute/. Accessed: 2023-04-01.

Mathew, B.; Illendula, A.; Saha, P.; Sarkar, S.; Goyal, P.; and Mukherjee, A. 2020a. Hate begets hate: A temporal study of hate speech. *HCI*, 4(CSCW2): 1–24.

Mathew, B.; Saha, P.; Yimam, S. M.; Biemann, C.; Goyal, P.; and Mukherjee, A. 2020b. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *arXiv preprint arXiv:2012.10289*.

Newton, C. 2019. The terror queue. https://www.theverge.com/2019/12/16/21021005/google-youtube-moderators-ptsd-accenture-violent-disturbing-content-interviews-video. Accessed: 2023-04-01.

O'Connor, J. 2021. Building greater transparency and accountability with the Violative View Rate. https://blog.youtube/inside-youtube/building-greater-transparency-and-accountability/. Accessed: 2023-04-01.

Parikh, P.; Abburi, H.; Chhaya, N.; Gupta, M.; and Varma, V. 2021. Categorizing Sexism and Misogyny through Neural Approaches. *TWEB*.

Rana, A.; and Jha, S. 2022. Emotion Based Hate Speech Detection using Multimodal Learning. *arXiv preprint arXiv:2202.06218*.

Rivers, C.; and Lewis, B. 2014. Ethical research standards in a world of big data. *F1000Research*, 3.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards VQA Models That Can Read. In *CVPR*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2015. Rethinking the Inception Architecture for Computer Vision. *CoRR*, abs/1512.00567.

Troianovski, A.; and Schechner, S. 2017. Germany to social networks: Delete hate speech faster or face fines. https://www.wsj.com/articles/germany-to-social-networks-delete-hate-speech-faster-or-face-fines-1498757679. Accessed: 2023-04-01.

Trujillo, M.; Gruppi, M.; Buntain, C.; and Horne, B. D. 2020. What is BitChute? Characterizing the. In *Proc. of the 31st ACM Conference on Hypertext and Social Media*, 139–140.

Wilson, A. 2022. 2023 video marketing statistics you simply can't overlook. https://beverlyboy.com/video-marketing/2023-video-marketing-statistics-you-simply-cant-overlook/. Accessed: 2023-04-01.

Wu, C. S.; and Bhandary, U. 2020. Detection of Hate Speech in Videos Using Machine Learning. In *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, 585–590. IEEE.

Xu, M.; Duan, L.-Y.; Cai, J.; Chia, L.-T.; Xu, C.; and Tian, Q. 2004. HMM-based audio keyword generation. In *Pacific-Rim Conference on Multimedia*, 566–574. Springer.

Yang, F.; Peng, X.; Ghosh, G.; Shilon, R.; Ma, H.; Moore, E.; and Predovic, G. 2019. Exploring deep multimodal fusion of text and photo for hate speech classification. In *Proc. of the Third Workshop on Abusive Language Online*, 11–18.

Ybarra, M. L.; Mitchell, K. J.; Wolak, J.; and Finkelhor, D. 2006. Examining characteristics and associated distress related to Internet harassment: findings from the Second Youth Internet Safety Survey. *Pediatrics*, 118(4): e1169–e1177.