A. Dataset Details

A.1. Annotation Label Descriptions

We present descriptions of all the annotation labels of the 4 taxonomies in Table 7. We shared these descriptions and sample videos to the annotators for reference.

A.2. Dataset

User Privacy: We have taken due measures for safeguarding the privacy of people present in our dataset. We will only be releasing the public URLs for the videos which would allow the users to delete their videos if they do not wish to be included in the dataset at any point of time. We have masked the user identifiers and the date of upload of these videos also to protect privacy. Additionally, these videos are already publicly available on the platform and we did not augment the videos with any information which is not in the public domain.

User Consent: To further our efforts towards user privacy, we have created an opt-out form where users can reach out to the team and have their videos removed from the dataset. The dataset would be available for only research purposes and commercial usage of the dataset will be strictly prohibited.

Geographical and Linguistic Diversity: Our dataset contains videos from 11 Indic languages which alludes to the high linguistic diversity. However, we understand that the dataset is dominated towards certain population and does not represent the global demographics. Similar observations have been made by Piergiovanni et al.[2] for large scale video datasets like HVU, HACS and Kinetics where majority of the videos belong to North America with less representations for Asia, Africa, Europe and Latin America. We encourage follow-up works to be cognizant of the demographics of our dataset. As part of future work, we intend to substantially increase our dataset and increase the coverage to more languages and countries.

Age and Gender Bias: We extract faces from our videos using an off-the-shelf facial analysis library[3]. We note a healthy male-to-female ratio of 3:2 highlighting gender diversity of our dataset. The average age computes to 27 years (7 years as standard deviation) and ranges from infants to over 80 years demonstrating the age diversity of 3MASSIV. In Figure 9, we plot the age distribution of our dataset. We understand that this analysis is not entirely accurate due to the mis-classification of the facial analysis models but it helps to understand the overall age and gender representation of the dataset.

Offensive Videos: Since the videos uploaded on the platform are user-generated, there are fair chances of uploading offensive videos also. The content moderation pipeline of the platform already filters out majority of the offensive videos. To further ensure a clean dataset, our expert annotation team flagged all the videos containing offensive content like hate-speech, pornography, explicit or suggested nudity, violence (guns/knives), gore, not-suitable-for-work (NSFW), abuses etc. We removed all of these videos from our dataset.

De-duplication: On social media, popular and viral videos are usually re-uploaded with minor or no changes resulting in lot of similar videos. We used de-duplication strategies to remove these visually similar duplicate videos from the dataset. The de-duplication was performed by extracting the visual features from the videos using 3D ResNet [23]) and removing the video samples having more than 0.8 cosine similarity scores. We employed this de-duplication to promote more diversity in the dataset. After this filtering, we ended up with a corpus of 50K videos that were then annotated by domain experts.

A.3. Creator Profiling

For modelling the creators of the videos, we select recent posts uploaded by them on the platform. These posts were selected such that they do not have an overlap with the labelled $50k$ videos. We use the global average of the creator representations for the creators with insufficient historical posts.

A.4. In-depth Inter-Annotator Agreement

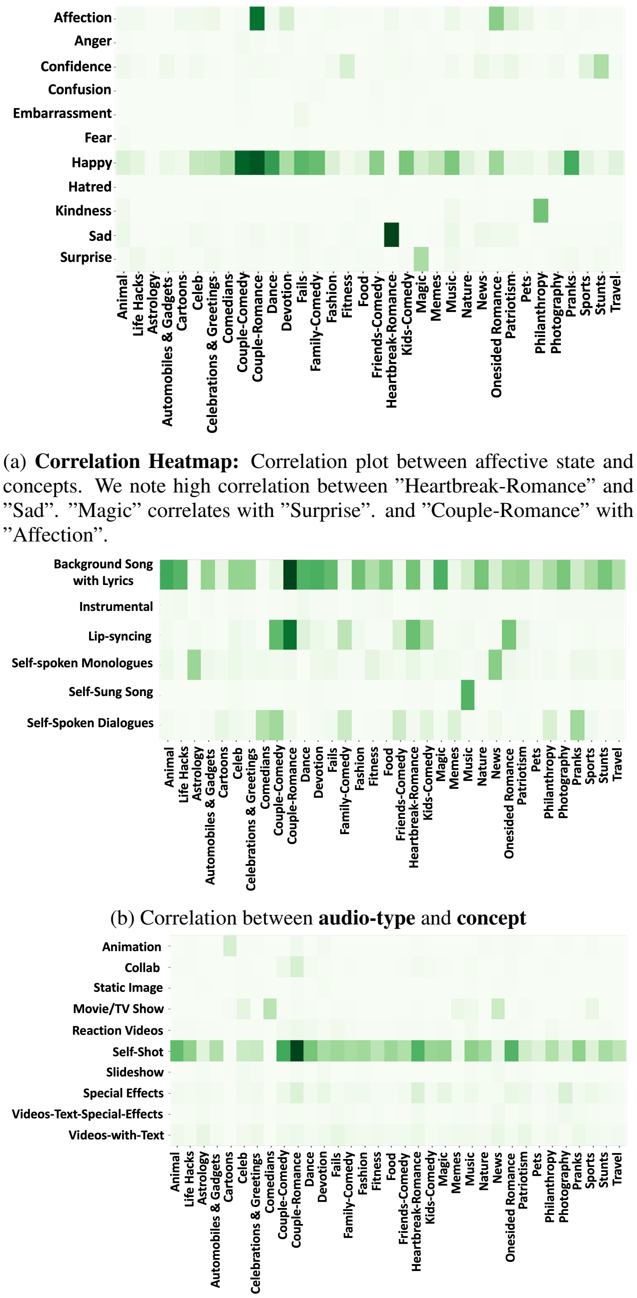
We summarize the per-label and per-taxonomy inter-annotator agreement values in Table 6. In general, inter-annotator agreement for all labels in theme taxonomy is high. Compared to other aspects, affective states show the most contrast in inter-rater confidence scores. This phenomenon has been observed in past works as well that have shown that infrequent emotions can have relatively high inter-rater correlation and frequent emotions can have have relatively low inter-rater correlation [14]. Similarly, for the audio type and video type, it is evident from Table 6 that labels that are infrequent have lower agreement values.

A.5. More Data Analysis

Figure 5b and Figure 5c presents the correlations between audio type and theme labels and video type and theme labels respectively. We observe that *comedy* and *romance* show a high co-occurrence with *lip-syncing* videos, while a large majority of *music videos* have *self-sung songs*. Similarly, it can be observed that *news* category is mostly composed of *self-spoken monologues* as the creators tend to use a narrative style for such content. *Pranks* and *comic* scenes are often presented as *self-spoken dialogues* as they generally involve at least two or more people. Most videos in our dataset are *self-shot*. Videos featuring *comedian*s and *trending news* are generally directly sourced from *movie/TV show clips* which finds evidence in Figure 5c. In Figure 6, we present the distribution of the length of videos in sec-

[2] https://arxiv.org/pdf/2007.05515.pdf

[3] https://github.com/deepinsight/insightface



(a) **Correlation Heatmap:** Correlation plot between affective state and concepts. We note high correlation between ”Heartbreak-Romance” and ”Sad”. ”Magic” correlates with ”Surprise”. and ”Couple-Romance” with ”Affection”.

(b) Correlation between **audio-type** and **concept**

(c) Correlation between **video-type** and **concept**

Figure 5. Correlation heatmap between different taxonomies of 3MASSIV.

onds. Our videos range from 5 seconds to 116 seconds with an average duration of 20 seconds.

A.6. Hashtag Analysis

For the 50K annotated posts, we analyzed the co-relation between the most common hashtags and `concept` taxonomy after removing common/noisy tags like *”trending”, ”hot”, ”viral”* using TF-IDF. We note that 40% of our videos did not have hashtags reiterating the role of a curated taxonomy for semantic understanding. Few tags which aligned with our taxonomy eg. ”comedy” and ”fun” were sparse and moreover, they were not specific enough to distinguish between the challenging yet popular categories like Kids-Humour vs Family-Humour. For affective states, audio and media type, we did not find relevant hashtags. Since social media platforms spread multiple geographies, we note that lot of hashtags are also code-mixed in nature. Our expertly curated taxonomy and annotations helped us in mitigating these problems with hashtags.

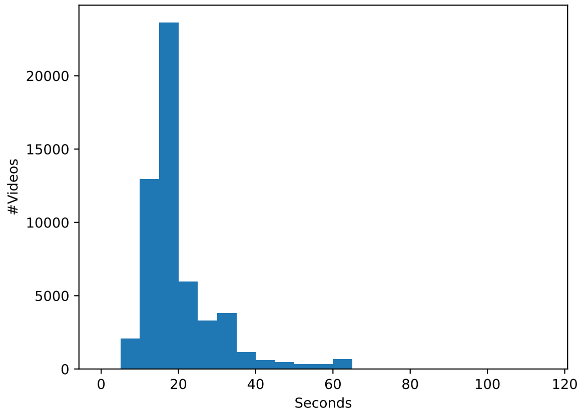


Figure 6. Distribution of video durations of 3MASSIV

B. Baseline Experiments: Training Details and Hyperparameters

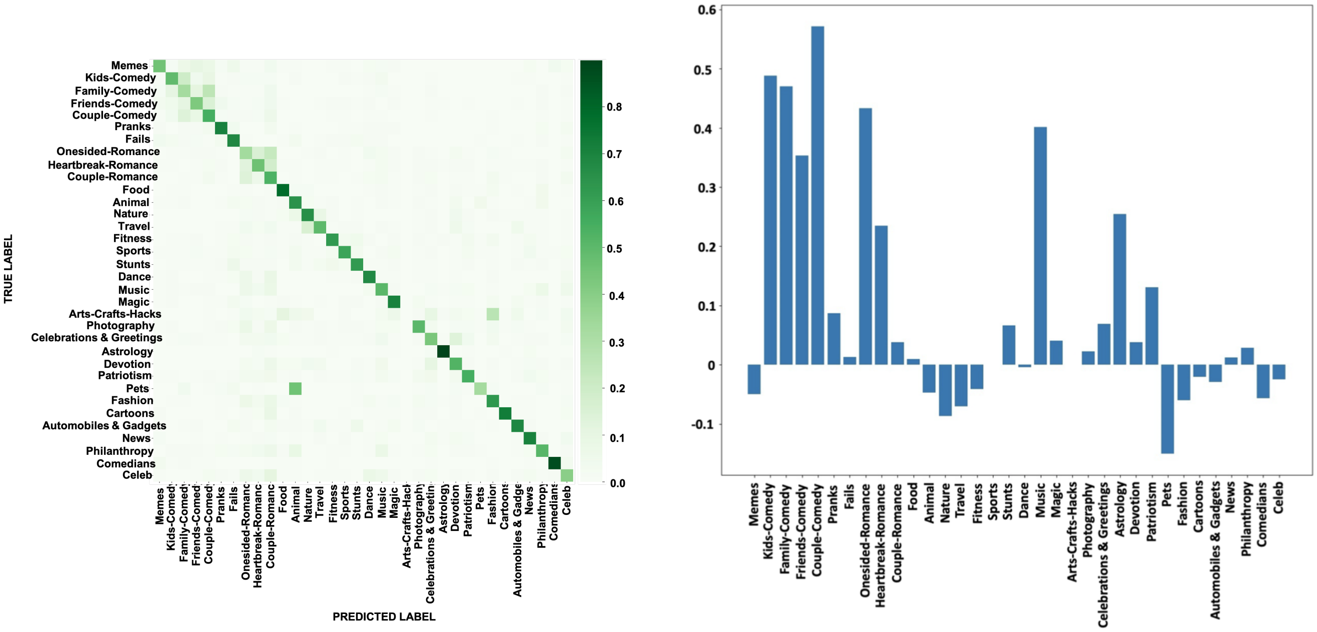
B.1. Concept Classification

B.1.1 Feature Extraction, Backbone Architecure and Hyperparameters

Visual Representations: We use ResNext [74] pretrained on Kinetics dataset as the backbone for extracting spatio-temporal features. We sample videos at 25 frame per second and save the frames with 240 as the size of the shortest side while maintaining the aspect ratio. We use 16 centrally cropped frames of dimension 112 x 112 x 3 as clips for extracting 2048-dimensional ResNext features from the last convolution layer. The features across clips of the videos are averaged to generate a 2048-dimensional representation for the video.

Audio Representations: We extract the audio channel as mono-channel from our videos using ffmpeg[4]. We sample the audios at 16kHz and use VGG [18] and CLSRIL-23 [22] models for audio feature extraction. In case of CLSRILS-23, the audio features are averaged across the clip to get a

[4] https://www.ffmpeg.org/



(a) **Confusion Matrix for Audio-Visual Concept Classification:** This is the confusion matrix for the audio-visual concept classification experiments as discussed in Section 4.1

(b) **Multimodal Analysis for Concept Classification:** We analyze the concept labels for which audio and the visual modalities help or degrade the accuracy of the concept classification.

Figure 7. **Further Analysis Audio-Visual Concept Classification:** We present an in-depth analysis regarding the audio-visual concept classification as discussed in Section 4.1.

512-dimensional vector. The features extracted from VGG are 128-dimensional vector as we tap them out before the classification layer.

Creator Representations: We use our trained audio-visual model to generate the 34 dimensional probability distribution across concepts for the videos posted by the creators. For each creators, we average the probability distribution to represent each creator by a 34 dimensional vector.

Architecture: For each modality, we pass them through 2 fully connected layers ($512 \rightarrow 512$ for creator and audio and $1024 \rightarrow 512$ for video) followed by softmax function for normalization. We concatenate these normalized outputs and pass them through a 2 layered late fusion network ($1536 \rightarrow 512 \rightarrow num\_classes$). We use batch normalization and ReLU activation function for the linear layers.

The network is trained for 500 epochs using cross entropy loss with 0.5 as dropout, 256 as batch size and 0.005 as learning rate. We use AdamW optimizer and decay the learning rate by 0.3 every 15 epochs. We use PyTorch[5] for all our experiments and train our models on A100 GPUs. We use the validation set for hyper-parameter tuning and early stopping and report the results on the test set.

B.1.2 More Analysis

In Figure 7a, we plot the confusion matrix of the audio-visual model. We notice misclassifications among the concept labels like *memes, kids, family, friends and couple comedy*, demonstrating the challenges in semantic understanding of such content. Similar confusion can also be observed for *romance* labels also. To analyze further, in Figure 7b, we study the impact on the accuracy of concept categories after using the audio modality also. We note that the majority of the categories benefit by adding audio indication rich information encoded in the audio channel also. However, some classes which are intuitively less aligned with audio and have generic background audios like *gadgets, animals, fashion, nature, travel* get impacted negatively.

B.2. Temporal Trends

In Figure 8, we collect top performing posts (50k) from each week (Aug, 2021 - Nov, 2021) on the basis of number of views. We use our trained audio-visual model to predict the probability distribution for these weekly posts to understand the correlation between real-life events and content uploaded on platform. We observe higher number of posts related to sports due to an upcoming major sports league. Similarly, we see some increase in posts related to festivals because of the recent festive season. These trends show how the temporal distribution of videos uploaded on social media platforms is dynamic and an interesting research direc-

[5] https://pytorch.org/

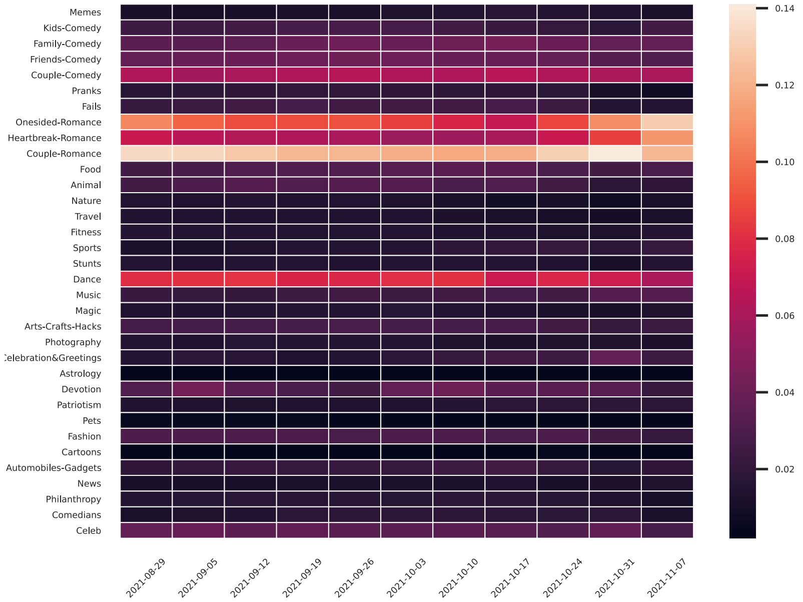


Figure 8. **Temporal Trends:** We plot the predictions of our audio-visual model on top performing post for 11 weeks (Aug, 2021 - Nov, 2021) to analyze the correlation between content uploaded and real-life events. *Sports* and *Festivals* show positive correlation with a sports league and festive season.

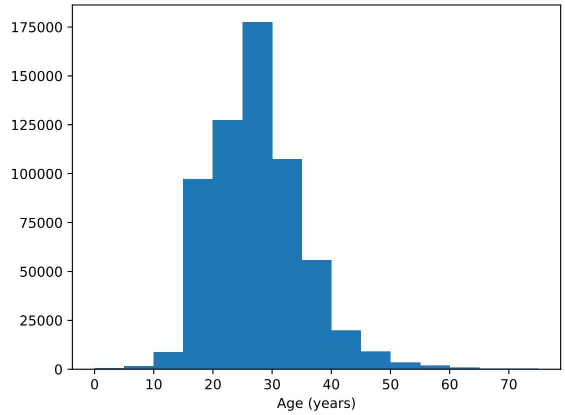


Figure 9. Distribution of age (years) of using the faces appearing in 3MASSIV. Average age of 3MASSIV computes to 27 years.

tion.

B.3. Broader Impact, Limitations and Potential Negative Impact

Majority of the existing datasets source videos from social media for video understanding, but are focused towards specific tasks like action recognition, object detection/segmentation etc. Our dataset 3MASSIV provides a unique opportunity for modeling the dynamics behind creations and consumption of these social media videos on short video platform which is a unique, novel and popular video format. 3MASSIV can be a stepping stone towards improving our understanding of human behaviour and preferences on video-first social media platforms. Our annotation of affective states can be instrumental in detecting signs of stress, cyber-bullying, cruelty on social media video platforms and help in providing a safe user experience. We annotate our dataset for 11 Indic languages is an attempt to develop an inclusive dataset, reducing the language bias and provide more representation to an under-represented population. However, this also forms the limitation of our dataset for truly global understanding, which we intend to tackle as follow-up work. Our dataset shows healthy distribution across age and gender, but also shows the natural imbalance in the age distribution owing to the adoption pattern of social media platforms across different age groups. Our dataset currently captures social media trends during a 9 month duration, which may not capture the entire range of existing trends occurring on social media. Since our dataset contains a variety of popular videos, often created by popular personalities, it could enable malicious parties to track and monitor people. However, we took substantial measures to preserve the identity of people by masking the dates and user identifiers for these videos. Additionally, we only use videos which are publicly available to prevent breach of user privacy and trust.

Taxonomy	Labels	Inter-Annotator Agreement	Avg. Inter-Annotator Agreement
Theme	Music	0.63	0.77
	Animal	0.92	
	Dance	0.84	
	Devotion	0.82	
	Magic	0.93	
	News	0.91	
	Celeb	0.81	
	Fails	0.81	
	Memes	0.85	
	Life Hacks	0.86	
	Fashion	0.81	
	Food	0.9	
	Nature	0.83	
	Philanthropy	0.78	
	Patriotism	0.89	
	Stunts	0.84	
	Festival	0.79	
	Sports	0.83	
	Photography	0.9	
	Automobiles & Gadgets	0.85	
	Fitness	0.87	
	Travel	0.73	
	Astrology	0.96	
	Cartoons	0.85	
	Pets	0.92	
	Pranks	0.83	
	Comedians	0.68	
	Couple Romance	0.65	
	Heartbreak Romance	0.62	
	Onesided Romance	0.68	
	Kids Comedy	0.8	
	Couple Comedy	0.82	
	Friends Comedy	0.91	
	Family Comedy	0.89	
Affective State	Happy	0.35	0.40
	Affection	0.33	
	Sad	0.61	
	Confidence	0.42	
	Surprise	0.18	
	Kindness	0.72	
	Anger	0.17	
	Confusion	0.07	
	Embarrassment	0.02	
	Fear	0.14	
	Hatred	0.08	
Audio Type	Background song with lyrics	0.63	0.59
	Lip-syncing	0.51	
	Self-spoken dialogues	0.65	
	Self-spoken monologues	0.61	
	No sound	0.38	
	Self Sung Songs	0.84	
	Instrumental Music Recording	0.31	
Video Type	Self-shot video	0.71	0.62
	Video with Special Effects	0.52	
	Video with Text	0.52	
	Splitscreen	0.73	
	Movie / TV Show Clips	0.67	
	Animation & Digital Art	0.72	
	Slideshow	0.44	
	Static Image	0.41	

Table 6. **Inter Annotator Agreement:** We summarize the per-label and per-category annotator agreement for 3MASSIV.

Taxonomy	Labels	Description
Concept	Music	Singing, beat-boxing, playing an instrument or other musical performance
	Dance	People performing solo/group dances
	Devotion	Videos related to divinity, spirituality and religion
	Magic	Magicians performing tricks and illusions
	News	Videos containing news, reports or coverage of events
	Celeb	Videos of celebrities from entertainment industry
	Fails	Some unplanned event creating a humorous situations
	Memes	Viral audio/video which are slightly edited to suit different contexts
	Life Hacks	Simple tricks for making daily activities easier
	Fashion	Videos showing/making aware of fashion tricks and trends
	Food	Videos focusing on preparation/consumption/review of food or beverage
	Nature	Videos capturing natural scenes like rivers, trees, mountains
	Philanthropy	Selfless and kind actions like helping poor
	Patriotism	Generate the feeling of love towards country
	Stunts	Showcasing challenging and thrilling skills and activities
	Festivals	Celebrating and sharing greetings during various festivals
	Sports	Clips contain any professional sports
	Animal	Videos containing wild animals
	Photography	Display tricks related to photography
	Automobiles & Gadgets	Video featuring automobiles or gadgets
	Fitness	Videos focusing on improving mental and physical health
	Travel	Videos capturing travel destinations and journeys
	Astrology	Videos containing astrology
	Cartoons	Videos with animated characters
	Pets	Pet animals like dogs, cats etc
	Pranks	Mischievous tricks that generate humorous situations
	Comedians	Popular comedian performing jokes or stand up act
	Couple Romance	Videos showing romance between couples
	Heartbreak Romance	People expressing feelings after breakup/betrayal
	One sided Romance	Videos showing that only one person is in love
	Kids Comedy	Humorous acts performed majorly by kids
	Couple Comedy	Funny content about couple relationship
	Friends Comedy	Funny content about friends
	Family Comedy	Humor generated around family and its members
Affective State	Happy	Videos eliciting happiness in viewers
	Affection	Videos depicting love and fondness
	Sad	Videos expressing sadness, grief, pain, suffering
	Confidence	People showing high confidence and self-esteem
	Surprise	Videos having an element of surprise
	Kindness	Video having an element of kindness
	Anger	Videos with people expressing anger and annoyance
	Confusion	Videos showing confusion among people
	Embarrassment	Videos depicting a feeling of embarrassment
	Fear	Videos having an element of Fear
	Hatred	Videos showing hatred and disapproval
Audio Type	Background song with lyrics	Background song having both lyrics and instrumentals
	Lip-syncing	People lip-syncing to a pre-recorded song or dialogue
	Self-spoken dialogues	Two or more people conversing with each other
	Self-spoken monologues	Single person speaking
	No sound	Feeble noise or complete silence in the video
	Self-sung songs	Audio contains song sung by the individual themselves
	Instrumental	Instrumental music in the background. No lyrics.
Video Type	Self-shot video	Original videos created by users
	Video with Special Effects	Videos with special artifacts like masks, blur effect etc.
	Video with Text	Text superimposed on the video
	Splitscreen	Two or more screens placed side by side
	Movie / TV Show Clips	One or multiple shots compiled together from an existing movie or TV-show
	Animation & Digital Art	Visuals generated using digital technology
	Slideshow	Sequence of images/video-snippets combined together as a video
	Static Image	Static image throughout the video

Table 7. **Taxonomy Descriptions:** We summarize the instructions shared with the annotation team for annotation 3MASSIV videos.