

BookTubing Across Regions: Examining Differences Based on Nonverbal and Verbal Cues

Chinchu Thomas
Multimodal Perception Lab
IIIT Bangalore, India
chinchu.thomas@iiitb.org

Dinesh Babu Jayagopi
Multimodal Perception Lab
IIIT Bangalore, India
jdinesh@iiitb.ac.in

Daniel Gatica-Perez
Idiap Research Institute and EPFL
Switzerland
gatica@idiap.ch

ABSTRACT

BookTubers are a rapidly growing community in YouTube who shares content related to books. Previous literature has addressed problems related to automatically analyzing opinions and mood of video logs (vlogs) as a generic category in YouTube. Unfortunately, the population studied is not diverse. In this work, we study and compare some aspects of the geographic/cultural context of BookTube videos, comparing non-western (Indian) and Western populations. The role played by nonverbal and verbal cues in each of these contexts are analyzed automatically using audio, visual, and text features. The analysis shows that cultural context and popularity can be inferred to some degree using multimodal fusion of these features. The best obtained results are an average precision-recall score of 0.98 with Random Forest in a binary India vs. Western video classification task, and 0.75 in inferring binary popularity levels of BookTube videos.

CCS CONCEPTS

• **Human-centered computing** → **Social media**; • **Social and professional topics** → **Cultural characteristics**; • **Computing methodologies** → **Supervised learning by classification**.

KEYWORDS

YouTube; BookTube; social video; diversity

ACM Reference Format:

Chinchu Thomas, Dinesh Babu Jayagopi, and Daniel Gatica-Perez. 2019. BookTubing Across Regions: Examining Differences Based on Nonverbal and Verbal Cues. In *ACM International Conference on TVX '19, June 5–7, 2019, Salford (Manchester), United Kingdom*

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
TVX '19, June 5–7, 2019, Salford (Manchester), United Kingdom

© 2019 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6017-3/19/06...\$15.00

<https://doi.org/10.1145/3317697.3323357>

Interactive Experiences for TV and Online Video (TVX '19), June 5–7, 2019, Salford (Manchester), United Kingdom. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3317697.3323357>

1 INTRODUCTION

The automatic analysis of online conversational video, particularly video logs (vlogs) in sites like YouTube [20], has become popular in the last decade for several reasons. First, from the perspective of social and behavioral science, the nature of human behavior generated by YouTubers is more diverse than was typically studied in the past, as people vlogging are no longer university students taking part in lab experiments. Second, from a multimedia analysis perspective, the recordings are in the wild, with no viewpoint, illumination or sensor quality controls. The third reason is practical relevance. Automatic analysis contributes to a better understanding of factors that might mediate audience engagement in social video, i.e., the personality or mood projected by YouTubers, in addition to the words they use, have an influence in their audience. While research on behavioral analysis of conversational online video in the past has analyzed personality and mood impressions among vloggers [32] [7], the diversity of the population is definitely not reflective of the actual population of the world. India and China put together have one-third of the world's population; whereas the population in most previous work involving YouTube conversational videos are from the US and to a lesser degree from other western countries. This is due to a combination of factors, including the sheer volume of videos generated by US YouTubers, and a research trend to focus on anglophone video data. A key research question arises, namely if the models built using datasets primarily from the US are “universal” and generalize to other populations. This is a specially critical question given the growing body of literature that demonstrates the negative impact of data biases in machine-learning-driven inference systems [11] [24]. In this paper, we contribute to the emerging area of automatic analysis and understanding of online video communities' practices. As a specific phenomenon, we analyze BookTubing videos from Western and non-Western countries, particularly India.

Booktubing refers to YouTube channels produced by users usually between the ages of 15 and 25 who talk about and review literary works, most typically Young Adult (YA) novels. The tone and style of the videos is usually casual and videos often include “book reviews”; “Hauls” in which a vlogger showcases a collection of books recently purchased; “Read-alongs,” in which a vlogger hosts a live reading event; “TBR,” in which a vlogger discusses the books in their “To Be Read” pile; “Wrap-Ups,” in which a vlogger briefly discusses a group of books which the speaker recently read but had not yet individually reviewed [26].

To study the diversity of the population among Booktubers, we rely on nonverbal and verbal communication cues. Although nonverbal communication is a universal phenomenon [21], the meanings of nonverbal cues are different across cultures [2]. Nonverbal communication is dependent on the context including verbal, situational and cultural [28]. It varies across “high-context” (East Asian countries such as China, Japan, Korea, India) and “low-context” cultures (Western countries such as United States, United Kingdom, including Australia, New Zealand). High-context cultures rely heavily on nonverbal cues for communication whereas low-context depends more on words themselves [18] [2] [35]. Since the study is on Indian and Western population, where India is an example of a high-context culture and Western countries are cases of low-context cultures, the study of nonverbal and verbal cues is pertinent.

At the same time, Western BookTubers have been closely working with their communities for a long time, compared to Indian BookTubers. In other words, many Western BookTubers are likely to be experienced with this social video genre. As studied in [12] BookTubers attempt to bring in differences in style and tone of creating the videos to get recognition from their community. This aspect might also influence verbal and nonverbal cue usage.

In this paper, we study differences in the BookTube community with respect to their cultural context (Western and non-Western). This is a novel theme in contrast to previous work, which has analyzed emotional and personality aspects of generic US vloggers (not including BookTubers) [14] [8] [7] [23] [39], or that has examined basic practices of BookTubers without doing any automated analyses [27] [26] [1] [12]. The research questions addressed are:

- RQ1: Can a basic label related to the cultural setting of a BookTube video, specifically Indian versus Western, be automatically inferred from nonverbal and verbal cues? If so, what cues play a major role in this task?
- RQ2: As a byproduct of the above, can nonverbal and verbal cues tell anything about another key aspect of BookTube audiences, namely a binary level of popularity?

The contributions of the work are: (1) A large corpus consisting of 17318 videos with audio, video, and text data, with a detailed descriptive analysis of its basic characteristics. In this work, we conduct automatic analysis of BookTube video in contrast to existing BookTube literature which focuses more on defining BookTube itself from the media studies perspective. (2) We conduct a behavioral analysis of BookTube users as a specific category of online videos and in the context of cross-cultural comparisons, based on nonverbal and verbal cues. For this, we define a binary classification task to infer Indian vs. Western BookTubers. (3) Understanding the popularity of a YouTube video is an extremely complex task where multiple factors like video content, network effects, etc. play a major role. In this work, we add to the literature on the specific domain of conversational online videos by studying the contribution of BookTube verbal and nonverbal content towards discriminating binary level of popularity of BookTube videos.

The rest of the paper is organized as follows: Section 2 discusses the related work. Section 3 describes the dataset. Section 4 outlines the descriptive analysis of the data. Section 5 describes nonverbal and verbal features that we extracted for the tasks. Section 6 presents the cross-cultural classification task. The binary task of inferring the popularity is outlined in Section 7. The findings of the analysis are discussed in Section 8. Finally, we conclude the paper in Section 9.

2 RELATED WORK

In this section we discuss the existing works in three categories and how our work is novel. First, we discuss the previous works that describe BookTubing as an online video subgenre, then works related to the behavioral analysis of online videos, and finally works that discuss the factors affecting the popularity of online videos.

BookTubing as an online video subgenre

The role of online media and user-generated content has increased tremendously over the past ten years in social media platforms like YouTube. Also, ever than before, more focused communities are growing in YouTube. One such community is BookTube where readers review, critique and discuss Young Adult (YA) books. They create and share anything and everything related to books. The BookTube channels are not constrained by language; there are channels not only in English, but also in many other languages. In this way, “the physical and linguistic boundaries of BookTube are limitless” [27]. The topics discussed in the community are not just book reviews, but also “Hauls” in which vloggers showcase a collection of books they purchased recently; “Read-alongs,”

in which a live reading event is hosted; “TBR,” i.e., “To Be Read” books; and “Wrap-Ups,” in which a vlogger briefly discusses a group of books which was recently completed but had not yet been reviewed individually [26]. Another featured topic is Tag/Challenge in which creative prompts are shared with the community to stimulate conversation; often they are just a series of topical questions. Everyone is encouraged to respond, and Tags/Challenges are then shared with others through “tagging” friends, challenging them also to complete the Tag/Challenge [26].

In the literature, BookTube is a recent phenomenon and researchers have first focused on defining the BookTube phenomenon. There are not many writings in this area and because of the constant and rapid changes in the community, it is challenging to track the phenomenon. Semingson et al. [27] in their paper defined BookTube, describing the BookTube features and observing that “BookTubing represents a networked attempt to learn and discover literature in today’s digital and multimodal spaces. As a global phenomenon, the cross-cultural implications of booktubing are multidimensional”. Perkins in [26] examined the BookTube community and the most popular BookTubers to understand the different roles that they play in the community. The author observed that reading books have become a social event with the advent of the BookTube community. Albrecht [1] tried to position BookTube in the publishing industry and analyzed the role of the BookTubing community in the young reader book world.

Ehret et al. in [12] described the affective experiences and the participatory cultures of content creators especially in BookTube. Hughes in the work [19] researched Young Adult (YA) BookTube channels and interviewed YA BookTubers, to discern the pivotal trends in the functionality of the BookTube community. The author observed that the trends include the characteristics of the books that are discussed, the influential role of the BookTuber and the possibility of formation of YA canons due to the kind of titles discussed in the community. Sorenson et al. [30] conducted a genre-analysis of BookTube to understand the relationship between Networked Knowledge Communities (NKC) and the Networked Knowledge Society (NKS). The authors examined BookTube as a discourse community and observed that the members have shared rules, genres, hierarchies and values. Furthermore, by relating these discourse features to other latent educational possibilities of the NKS, the authors explored how BookTube might be usefully implemented to model NKC practices in more traditional face-to-face educational settings.

The works discussed in the previous paragraphs focus on qualitative studies of BookTube using a small sample, whereas our work focuses on quantitative analysis using a

significantly larger dataset. The literature discussed was centered on defining the community, understanding the characteristics, the genre and trends in the community. In contrast, our work focuses on the quantitative study of the audiovisual content generated by a set of BookTubers.

Behavioral analysis of online video

Biel et al. [7] studied personality impressions in social media from crowdsourced impressions and predicted the impressions using automatic nonverbal cues from conversational video logs extracted from YouTube. They also investigated the associations between crowdsourced personality impressions and the actual levels of attention that vloggers gather online, using several measures of attention estimates from YouTube metadata. Teijeiro et al. [32] proposed a model to predict personality impressions from statistical cues and activity cues computed over facial expressions. They studied the relation between emotions from face expressions and personality traits. Gatica-Perez et al. [14] studied and predicted personality impressions of conversational videos from the longitudinal data collected from YouTube. They observed that the inference about a vlogger from multiple videos achieved better performance than single videos. Biel et al. [9] addressed the personality prediction task with verbal content of videos. All these works looked into personality recognition.

More recently, the personality trait problem is approached with deep learning methods. In [39] Zhang et al. proposed an ensemble model with visual and audio data using early and late fusion of features extracted from convolutional neural networks (CNN) for personality analysis. Subramaniam et al. [31] also used deep models like 3D CNN and Long Short Term Memory (LSTM) for predicting the personality impressions from short videos. The method uses preprocessed audio-visual features. For the same task, Gucluturk et al. [17] utilized audio-visual deep residual network which is trained end-to-end on raw videos.

In [23] Park et al. tried to understand persuasiveness from online data. In this work the authors computed verbal and nonverbal descriptors from audio, visual and text data and investigated the usefulness of combining multimodal descriptors for thin slices for the prediction of persuasiveness. Nojavanasghari et al. [22] also used multimodal fusion to leverage the complementary information from individual modalities for predicting persuasiveness.

The above works focused on trait prediction, including personality and persuasiveness. These works also studied vlogs as a generic category, and most often with smaller datasets compared to our work which uses 4478 videos. Our work is in the direction of understanding the cultural context of YouTube videos in the context of BookTube

videos which is a sub-community within YouTube. Specifically, the BookTubing video genre has not been considered in previous automatic analysis studies.

Popularity of Online Video

Literature in psychology, social computing and machine learning tried to understand the factors affecting the popularity of a video in YouTube. Guadagno et al. in [16] examined the role of emotional response and video source on the likelihood of spreading an Internet video. The authors concluded that these results have implications for emotional contagion, social influence, and online behavior of the users. Content that generates stronger affective responses appear to be more likely to spread as a viral video. Wattenhofer et al. [36] examined the social and content facets of user popularity and found a stronger correlation between a user's social popularity and their most popular content as opposed to typical content popularity. Vallet et al. [34] focused on the observable dependencies between the virality of video content on Twitter and the popularity of such content on YouTube. The authors found that YouTube features capturing user engagement have more effective prediction capabilities. Biel et al. [7] studied the relationship between the personality traits and the attention gained by vloggers. They found that traits such as extroversion and openness to experience are correlated to average social attention. Biel et al. [6] studied possible correlation of nonverbal features to average social attention. The authors conducted a small sample study on a coarse audio-visual feature set, whereas we extract finer features on a much larger dataset for a binary classification task. Trzcinski et al. [33] trained a Support vector regression model using visual features and metadata like views and likes to predict the popularity of social media videos.

Recently, deep learning predictors of popularity of YouTube videos have started to appear. Bielski et al. [10] proposed the use of deep learning methods along with attention mechanisms for predicting the popularity of social media videos from video data and the video title. Since we focus more on the explainability of the features and the task, we are not using deep learning methods in this work. The literature shows that understanding the popularity of a YouTube video is an extremely complex task where multiple factors such as video content, network effects, affective response of the user, social influence and user behavior play a major role. In this work, we examined the role played by verbal and nonverbal content of the video that makes it popular. In short, we did a preliminary analysis of the role played by the affective behavior of the video content of the video with respect to discriminative power for recognition of binary popularity levels. This contributes to the complex task of inferring the popularity of a BookTube video on a larger dataset with 4478 videos in the BookTube community.

3 DATASET

Our dataset consists of 17318 BookTube videos. This section outlines description of the data along with an analysis of the metadata.

Data Collection

The Booktube videos consist of different categories like reviews, hauls, to-be-read, wrap-ups, tags, discussions and unboxing. The corpus has videos of Indian and Western BookTubers. There are videos from 28 BookTube Indian channels and 36 Western channels in the corpus. The total number of videos in the corpus is 17318 (Western: 15620 videos; Indian: 1698 videos). All BookTubers speak English. The average number of videos per channel in the Western community is 433 and Indian community is 60. The number of subscribers for each of the channels varies between 13k-390k in the Western community and 27-44k in the Indian community. The dataset also has metadata for each of the videos along with automatic audio transcriptions from YouTube. The corpus has videos uploaded between 10 December 2009 to 30 August 2018.

For the analysis, we curated the dataset by removing three categories of videos, namely "tag", "discussion" and "unboxing". These videos are very different from other categories, featuring with multiple people in the videos. The curation was done using the tags information available in the metadata. All the videos with user tags such as "discussions", "unboxing" and "tags" were removed. The curated dataset consist of 4478 videos from reviews, hauls, wrap-ups and tbr categories. There are only 55 channels in the curated set (21 Indian and 34 Western). The average duration of the videos is 7 minutes. The rest of the experiments are done on the curated dataset.

Metadata

The metadata for each video was downloaded from YouTube. This consists of information like upload date, duration, view count, like count, or dislike count. In this work, we use the metadata as ground truth for the experiments.

4 DESCRIPTIVE ANALYSIS OF METADATA

This section outlines the descriptive analysis of the metadata. There are 4478 videos, out of which 3830 are from Western users and 648 are from Indian users. The statistics of the data is given in Table 1. The analysis is as follows:

Population distribution: The corpus consists of 21 channels from India and 34 channels from Western BookTubers. The dataset includes several popular BookTubers. In the Western category 18 channels are from US, 7 from UK, 4 from Australia, 4 from Canada and 1 from Germany.

Table 1: Data Description

	No. of Channels	No. of Videos	No. of Videos per channel (mean, median, std)	#Subscribers per channel (mean, median, std)	#Views per video (mean, median, std)	% of reviews, hauls, tbr, wrap-ups
Indian	21	648	30, 15, 34	4740, 466, 10k	2067, 395, 8334	58, 17, 12, 13
Western	34	3830	112, 106, 79	109k, 67k, 101k	15k, 6604, 24k	38, 30, 9, 23

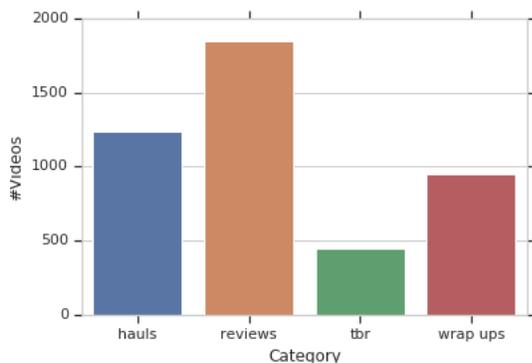


Figure 1: Category-wise distribution

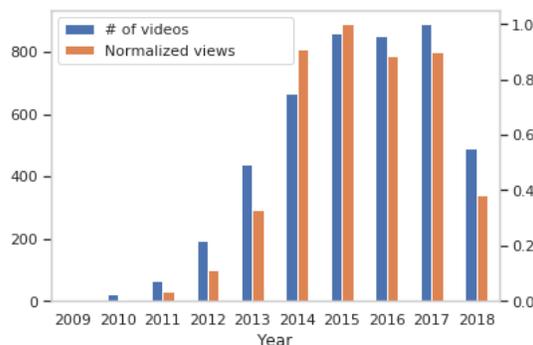


Figure 2: Distribution of videos uploaded and views over the years

Gender distribution within the channels: There are videos of 7 male BookTubers and 48 female BookTubers in the curated dataset, and hence it is biased towards female BookTubers.

Distribution of video categories: The majority of the channels have videos from all categories considered here, such as reviews, hauls, tbr, and wrap-ups. In general, BookTubers concentrate more on review videos as seen in Figure 1. Certain channels concentrate on review videos alone. The corpus has 41 percent of videos from the review category, followed by 28 percent from from hauls, 21 percent from wrap-ups, and 10 percent from tbr. The distribution is shown in Figure 1.

Distribution of videos and views over time: The BookTube corpus is from 2009 to 2018. The distribution in Figure 2 shows that BookTube is a growing community, and that the number of videos uploaded in the initial period is comparatively low. The popularity of the genre is also growing as seen in Figure 2.

Distribution of number of videos per channel: The average number of videos per channel in the corpus is 83. The number of videos for each channel varies between 1 and 343. The number of videos for each channel is shown in Figure 3a. An unpaired t-test (t-test can be applied in this setting since the sample size is large) for the number of videos per channel for the Indian and Western groups shows that there is a significant difference in the mean number of videos from

the two communities ($p < 0.0001$), indicated by the effect size (Cohen’s d): we have a large effect size with Cohen’s d of 1.24 and 95% confidence interval (CI) is [0.65, 1.83] for the effect size. The CI does not include zero and the CI range is large which implies that the effect apparently exists and may be large [38].

Distribution of number of subscribers and views for each channel: The distribution of number of subscribers for each channel (normalized with respect to the largest individual value) is shown in Figure 3b. The number of subscribers varies largely and the view count is proportional to the number of subscribers. Figure 4a shows that only few channels have a very large number of view count, which possibly makes them celebrated within the community [26]. Table 1 shows statistical differences in the mean number of views for Indian and Western BookTubers. A t-test shows that the mean number of views for Indian and Western groups is significantly different ($p < 0.0001$), with a medium effect size computed using Cohen’s d (0.58) and 95% confidence interval [0.49, 0.66]. The CI does not include zero and the CI is small which implies that the effect apparently exists [38].

Book category versus view count, like count, duration: Although the number of videos is higher for the review category, the maximum numbers of views and likes are for haul videos, as seen in Figure 4a and 4b. Hauls and wrap-ups are lengthy videos compared to reviews and tbr as in Figure

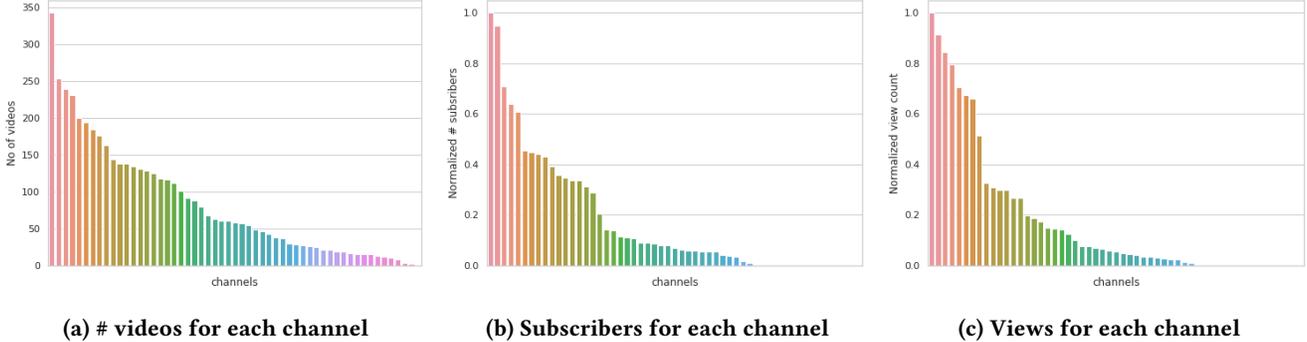


Figure 3: Distribution of number of videos, subscribers and views

4c. We also observe many outliers for reviews in terms of duration.

Word cloud: Figure 5 shows the word cloud for the Indian and Western videos obtained from automatic speech recognition. The figure shows that some of the most common words are the same (kind, know, now, one, think). Words as features will be studied in more detail in later sections.

5 NONVERBAL AND VERBAL FEATURE EXTRACTION

As discussed in the Introduction section, to differentiate between Indian and Western BookTubers, nonverbal features and verbal features are expected to be useful. In the sections below, we describe the audio, visual and textual features extracted for further analysis from audio, video and the automatically generated transcripts.

Audio features

The audio feature set consists of 64 low-level descriptors (LLDs) including energy, spectral and voicing related LLDs from the OpenSMILE toolbox [13]. The feature set also has functionals applied to the LLDs. In total the feature set consists of 5759 features used in the INTERSPEECH 2012 Speaker Trait Challenge [29]. The list of 64 LLDs are given in Table 2.

Visual Features

The visual features are based on headpose, eye gaze and intensities of facial action units. These features are relevant since high context culture inherently use these modes for nonverbal communication. The visual cues are extracted from the OpenFace toolbox [5][4][3][37]. Statistical properties like mean and variance are computed for rotation of head in x, y, z directions; eye gaze direction vector in x, y, z directions and facial action unit intensities for 17 action units. The final visual feature vector consists of 52 features.

Verbal Features

Verbal features also play an important role in communication. We use automatically generated subtitles from YouTube in Web Video Text Tracks (WebVTT) format. The spoken text from WebVTT files was extracted by parsing the text. In this section, we describe the linguistic features, count vectorizer, and term frequency inverse document frequency (tf-idf) features computed for the extracted verbal content.

Lexical Features. Lexical features are widely used in psychology and social computing. These features are computed from Linguistic Inquiry and Word Count (LIWC) [25] tool. It calculates the degree to which various words from 60 different categories are used in everyday speech, including words from psychological processes, linguistic processes, personal concerns and spoken categories.

Count vectorizer and tf-idf. The count vectorizer counts the number of word occurrences in a document. The term frequency-inverse document frequency (tf-idf) computes the importance of a word to the document in the corpus. The features computed are tf-idf word-level vector and tf-idf 2-gram vector. The features are computed using the sklearn package from Python. The documents in the corpus were used as they are, without preprocessing. The feature vector was 500-dimensional for both count vectorizer and tf-idf.

6 INFERRING INDIAN VS WESTERN BOOKTUBERS (RQ1)

In this section we address RQ1, which examines the task of inferring the cultural context of a BookTube video. The ground-truth for the cultural context was obtained from the location information given in the YouTube channels considered here. There was only one channel in the dataset in which the BookTuber is of Indian origin but the location is USA, but this person had moved to the US years ago. Therefore the use of location as the ground-truth for the context is sufficiently justified. All the videos belonging to channels

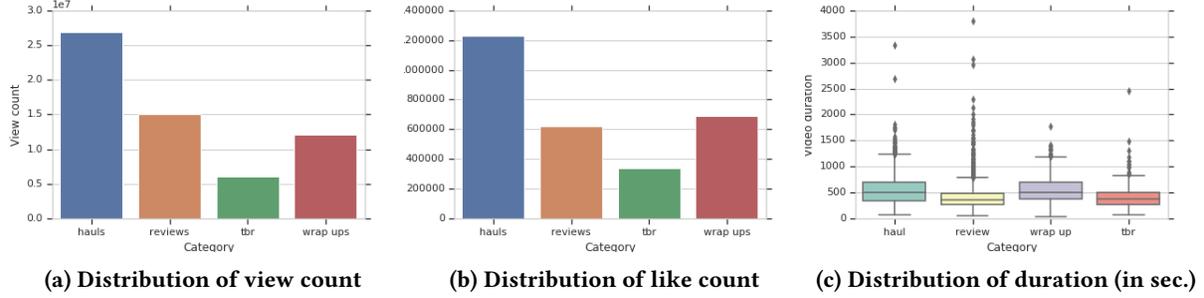


Figure 4: Distribution of metadata with respect to categories

Table 2: Audio features: 64 Low level descriptors (Table reproduced from [29]). The numbers in brackets indicate the number of features for each feature type

energy related LLD (4)
Sum of auditory spectrum (loudness), Sum of RASTA-style filtered auditory spectrum, RMS Energy, Zero-Crossing Rate
spectral LLD (54)
RASTA-style auditory spectrum bands 1-26 (0-8 kHz), MFCC 1-14, Spectral energy 250-650 Hz, 1 k-4 kHz, Spectral Roll Off Point 0.25, 0.50, 0.75, 0.90 Spectral Flux, Entropy, Variance, Skewness, Kurtosis, Slope, Psychoacoustic Sharpness, Harmonicity
voicing related LLD (6)
F0 by SHS + Viterbi smoothing, Probability of voicing, logarithmic HNR, Jitter (local, delta), Shimmer (local)

from Western BookTubers are considered as one class and videos from Indian channels are in the other class. There are 3830 videos in Western category and 648 videos in Indian category. The details of features, experiments and results are outlined in the following sections.

Feature selection

In this subsection, the feature selection is explained. The extracted features from each of the audio-visual-textual modalities are large in dimension. In order to reduce the dimensionality of the feature vector we used feature selection. This also reduces overfitting, model complexity, and reduce noise in the data. In the experiments, we used the extremely randomized tree (ExtraTree) classifier to compute feature importance and discard irrelevant features.

Audio features. We considered the top 20 features, after which the feature importance decreases. The top relevant features from the ExtraTree classifier feature selection for the audio data are voicing related features, spectral entropy, spectral roll-off, MFCCs and psychoacoustic sharpness with different functionals applied over these set of features.

Visual Features. For visual analysis, the top 20 features were selected. The mean of eye gaze vector in y, z directions; variance of eye gaze vector in z-direction; mean of the location of the head with respect to camera in y, z directions; mean of pitch of head; variance of head yaw; mean of action unit intensities such as AU04 (Brow lowerer), AU06 (cheek raiser), AU14 (dimpler), AU17 (chin raiser), AU25 (lips apart), AU45 (Blink) and variance of action unit intensities such as AU04 (Brow lowerer), AU06 (cheek raiser), AU12 (lid tightener), AU14 (Dimpler), AU25 (lips apart), AU26 (Jaw drop), AU45 (Blink).

Linguistic Features. The transcripts of the BookTube videos were analyzed using LIWC scores, word count and tf-idf vectorizer. The top 20 LIWC lexical features selected through attribute importance from the feature selection procedure are categories such as you, leisure, work, causation, social processes, exclusive, discrepancy, tentative, perception, religion, I, ingestion, relativity, biological processes, health, humans, hear, money, inhibition and space.

The tf-idf features are ranked with feature selection and top 100 features were selected. The tfidf vectorizer feature consists of words related to book, channel, adjectives, people etc.



(a) Indian



(b) Western

Figure 5: Word cloud for Indian and Western

Multimodal Fusion Features. In multimodal fusion, audio, visual and linguistic features are combined together. Experiments are done with feature selection. The feature set for the multimodal experiments are created by concatenating the top 20 audio features, 20 visual features and 20 textual features. The selected set of features are the top 30 features for which the feature importance score is greater than 0.01. The top features are given in Table 3.

Experiments

In this subsection, we describe the experiments conducted to address RQ1. Once the important features are computed, a binary classification is done to infer the cultural context of a BookTuber. The classification task uses different machine learning algorithms like Logistic Regression (LR), Support Vector Machine (SVM) and Random forest (RF) from sklearn package in Python. The dataset is divided into 70 % train set and 30 % test set with stratified sampling. The hyper parameters of the machine learning algorithms are optimized over 5-fold cross-validation of the train set using the grid search algorithm in sklearn package. The final inference is done on the 30% test set. The hyper parameters of the various models are given in Table 4. The SVM model is trained for linear kernel. The number of data points in each of audio, visual and text data is different. The audio features were computed on 4478 samples. If OpenFace toolbox did not detect a face on the majority of frames of a video, such samples were removed for the experiments. This resulted in 3281 samples in the visual feature set. The linguistic features were computed

on 3376 samples for which the word count was not zero. For the multimodal experiments, common samples from audio, visual and text sets were considered, resulting in 2910 samples.

Results

We now present the experimental results. The measures used for evaluation are average precision-recall score along with accuracy, since the binary classes had a large class imbalance. The dataset has roughly 15 % of sample points from the Indian population and 85% from the Western population.

Audio Features. The inference task is binary classification of Indian vs. Western BookTubers. The baseline for the task is the accuracy from the majority classifier created for the test set. The majority classifier classifies everything to the majority class, here Western. The results for different machine learning algorithms are summarized in Table 5. The accuracy for the baseline classifier is 85%. Random Forest does a good classification job with average precision-recall score of 0.93 resulting in 96% accuracy.

Visual Features. The classification result with visual features alone is given in Table 5. The baseline classifier resulted in an accuracy of 86%. Support Vector Machine and Random Forest classifiers are performing better than the baseline classifier. The average precision-recall score is 0.72 and 0.90 for SVM and Random Forest respectively.

Linguistic Features. With LIWC, count vectorizer and tf-idf features, the classification results are given in Table 6. The baseline for the task is 0.90. Logistic regression is performing better for LIWC and tf-idf-ngram features with an average precision-recall score of 0.84. For count vectorizer and word-level tf-idf features, Random Forest is giving better performance with an average precision-recall score of 0.93 and 0.94 respectively.

Multimodal Fusion Features. The multimodal fusion of audio, visual and text features resulted in a better separability of the classes. The text features considered for the multimodal fusion was the word-level tf-idf feature vector since it was performing better than other text-based features when considered alone. The results are tabulated in Table 5. As we can see, the fusion of features improved the average precision-recall score to 0.98 and accuracy to 98%.

7 INFERRING BINARY POPULARITY LEVEL (RQ2)

We now outline the methodology to address RQ2, i.e., inferring a binary popularity level of BookTube videos using nonverbal and verbal cues. The ground truth for the binary task of inferring the popularity is created from the normalized like count available in the metadata. The like count is normalized with the view count for each of the videos and

Table 3: Top features used for multimodal fusion of Indian vs. Western classification experiments

Top 30 Features
Functionals such as quartile, percentile, linear prediction (LP) coefficients and gain, standard deviation of rising slope, linear regression slope of voicing LLDs, computed for spectral roll-off point 0.25, 0.75, spectral entropy, psychoacoustic sharpness, MFCC1; mean of eye gaze direction vector in y-direction; mean of rotation of head in x-direction(pitch); variance of intensity of action unit 4, 6, 25; 'videos', 'kind', 'video', 'comment', 'bunch'

Table 4: Hyper parameter for the models (# Est-number of estimators is 500, MM-Multimodal)

Modality	SVM C	Random Forest		Criteria
		Max feat	Max depth	
Audio	0.1	'auto'	10	'entropy'
Visual	1	'auto'	15	'gini'
Lexical	0.1	'auto'	15	'entropy'
MM	0.1	auto	15	'entropy'

then scaled to 0-1 range. The resulting score is considered as the popularity score for a video. A popularity score greater than or equal to the median value (0.21 in this dataset) is considered as high popularity class, while below the median is considered as low popularity class.

Features

The features considered for this problem is also selected with feature importance from Extra Trees classifier. The top 20 audio features include functionals computed on LLDs such as sum of RASTA-style filtered auditory spectrum and MFCC 3, 7, 9. The top visual features are mean of intensity of action units AU1, AU4, AU7, AU6, AU12, AU14, AU25, AU45; variance of intensity of action units AU1, AU4, AU6, AU12, AU14; mean of eye gaze vector in x, y directions; mean of yaw, pitch, roll; mean of distance of head from camera in y, z directions. The important LIWC lexical features are words related to cognitive processes, tentative, assent, she/he, insight, I, social processes, achievement, word count, they, past, negations, number, we, prepositions, relativity, money, non-fluencies, leisure and words/sentence.

The top 30 features in the multimodal fusion set includes functionals computed on LLDs such as sum of RASTA-style filtered auditory spectrum, MFCC 3, 7, 9; mean of eye gaze vector in x direction; mean of head yaw; mean and variance of intensity of action unit AU12; words in the category of cognitive processes, relativity, insight, assent, tentative and negations.

Table 5: Indian vs. Western classification result with audio, visual and multimodal features.

(APS-Average Precision Score, Acc- Accuracy, MB-Majority Baseline, LR- Logistic Regression, SVM- Support Vector Machine, RF- Random Forest)

Classifier	Audio		Visual		MM	
	Acc	APS	Acc	APS	Acc	APS
MB	0.85	-	0.86	-	0.90	-
LR	0.87	0.50	0.86	0.52	0.91	0.44
SVM	0.88	0.53	0.91	0.72	0.90	0.40
RF	0.96	0.93	0.94	0.90	0.98	0.98

Experiments and Results

For the task of binary popularity classification, we used the same set of machine learning algorithms and procedures as explained in Section 5. The results for inferring the popularity are tabulated in Table 7, with Random Forest as the classifier. The majority baseline is accuracy is 0.52. The results show that multimodal fusion features are performing better compared to single modality features. The multimodal fusion resulted in an accuracy of 69% and average precision-recall score of 0.75.

8 DISCUSSION

Inferring Indian vs. Western BookTubers. The results in Table 5 show that audio features contribute more to the classification task of Indian versus Western categories when individual data modalities are considered. The visual features come next followed by linguistic features. The multimodal fusion features resulted in further improvement w.r.t. average precision-recall score and accuracy. The top features in the multimodal fusion features were all audio features followed by visual and linguistic features as seen in Table 3. However, there was a marginal difference of 0.01 in the average precision-recall score if linguistic features were removed completely from the set. This shows that verbal features are not contributing much to the task when used jointly with audio-visual features. The verbal features itself performed reasonably when considered alone, yet when combined to

Table 6: Indian vs. Western classification result with linguistic features (APS-Average Precision Score)

Classifier	LIWC		Count vectorizer		tf-idf word level		tf-idf-ngram	
	Accuracy	APS	Accuracy	APS	Accuracy	APS	Accuracy	APS
Majority Baseline	0.91	-	0.90	-	0.90	-	0.90	-
Logistic Regression	0.96	0.84	0.97	0.89	0.96	0.90	0.95	0.84
Support Vector Machine	0.96	0.82	0.96	0.88	0.96	0.92	0.95	0.83
Random Forest	0.92	0.77	0.95	0.93	0.96	0.94	0.94	0.79

Table 7: Popularity classification result with Random Forest classifier

Feature	Accuracy	Average Precision score
Majority Baseline	0.52	-
Audio	0.64	0.70
Visual	0.61	0.65
Lexical	0.60	0.66
Count vectorizer	0.66	0.71
Tf-idf word-level	0.67	0.69
Multimodal	0.69	0.75

audio and visual features, the verbal features do not seem to help. Overall, the results show that the studied cues can differentiate between the two classes.

The median of the top visual features of the Indian and Western Booktubers was computed to understand how it varies within the group. The median values show that head movement and eye movement are higher for Western BookTubers, whereas facial action unit intensities are higher for Indian BookTubers. While facial expression differences could be seen as in alignment with what has been reported in the past regarding high and low-context aspects, head movement and eye movement do not support this [18] [2] [35]. This issue needs to be investigated in more depth in future work.

The usefulness of multimodal feature fusion was discussed in the literature for predicting persuasiveness [23] [22]. This is clearly evident in the tasks described here also. Multimodal fusion achieves a better discrimination of Indian versus Western BookTubers.

Inferring Binary Popularity Level. Though there is an increase in accuracy and average precision-recall for the classification of popular videos with multimodal fusion features, the numbers in Table 7 show the complexity of the problem for even a simplified binary task. The results reveal that nonverbal cues and verbal cues also contribute to the task in addition to the other aspects discussed in Section 2. In

particular, Biel et al. [6] [7] studied possible correlations between aggregated nonverbal features and social attention of generic YouTubers on a somewhat related task. Some of the verbal features performed better than audio and visual features for inferring the popularity level. Unlike the task in RQ1 where audio features are dominant, here the word usage is also gaining prominence. The content of the video is also important along with the nonverbal aspects. Here the word usage gained prominence, i.e., the spoken content of the video was slightly better than the nonverbal aspects. Multimodal feature fusion resulted in the best performance, but the improvement is not as noticeable as in the previous task in RQ1. This can be due to the dependency of the popularity variable to several other important variables as described in [16] [36] [34] [15].

9 CONCLUSION AND FUTURE WORK

In this work, we presented the study of BookTube videos as a novel theme to study the cultural context of YouTube videos using nonverbal and verbal cues. Our conclusions are summarized in the following paragraphs.

RQ1 was posed to understand basic differences of YouTube videos based on the cultural context of BookTubers. The experiments revealed that Indian vs. Western users can be distinguished automatically using nonverbal and verbal cues. Each of the audio, video and textual features contributed to the inference task. Multimodal fusion further improved the results and showed the importance of both audio and visual cues in this task.

RQ2 was posed to understand the role played by the behavior and spoken words of BookTubers to infer a basic, binary popularity level of a BookTube video. The experiments showed that nonverbal and verbal cues provide some ability to discriminate, yet the task is very challenging. Among the modalities, verbal content was found to be slightly more relevant than audio-visual nonverbal cues; their combination produced the best performance. The results illustrates the difficulty of the problem, due to a number of factors involved in the concept of popularity as the literature shows.

While we showed that BookTubers can be differentiated based on verbal and nonverbal cues, our work did not explain the reasons why. This is one of the limitations of this work. Another limitation is the gender imbalance, which might be characteristic of popular BookTubers (or at least of BookTubers that are recommended by the search options of YouTube). Furthermore, the dataset could be made more diverse in terms of geographic distribution, which we plan to further expand in future work. Regarding the addressed tasks, inferring the popularity level can be posed as a multiclass classification problem, since the view count varies over a large range, or as a regression problem. Any of these formulations remain challenging given the complexity of online diffusion processes. Finally, a longitudinal analysis can be another direction to understand the changes in the BookTube community over the years, in terms of practices and expertise of BookTubers.

ACKNOWLEDGMENTS

This work was supported by a Mobility Grant awarded by the Swiss Leading House South Asia 2018. The first author would also like to thank Visvesvaraya PhD Scheme, Ministry of Electronics and Information Technology (MeitY), Government of India.

REFERENCES

- [1] Katharina Albrecht et al. 2017. *Positioning BookTube in the publishing world: An examination of online book reviewing through the field theory*. Master's thesis.
- [2] Peter A Andersen, Michael L Hecht, Gregory D Hoobler, and Maya Smallwood. 2003. Nonverbal communication across cultures. *Cross-cultural and intercultural communication* (2003), 73–90.
- [3] Tadas Baltrušaitis, Marwa Mahmoud, and Peter Robinson. 2015. Cross-dataset learning and person-specific normalisation for automatic action unit detection. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, Vol. 6. IEEE, 1–6.
- [4] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2013. Constrained local neural fields for robust facial landmark detection in the wild. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*. 354–361.
- [5] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 1–10.
- [6] Joan-Isaac Biel and Daniel Gatica-Perez. 2011. VlogSense: Conversational behavior and social attention in YouTube. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 7, 1 (2011), 33.
- [7] Joan-Isaac Biel and Daniel Gatica-Perez. 2013. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *IEEE Transactions on Multimedia* 15, 1 (2013), 41–55.
- [8] Joan-Isaac Biel, Lucía Teijeiro-Mosquera, and Daniel Gatica-Perez. 2012. FaceTube: predicting personality from facial expressions of emotion in online conversational video. In *Proceedings of the 14th ACM international conference on Multimodal interaction*. ACM, 53–56.
- [9] Joan-Isaac Biel, Vagia Tsiminaki, John Dines, and Daniel Gatica-Perez. 2013. Hi youtube!: Personality impressions and verbal content in social video. In *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 119–126.
- [10] Adam Bielski and Tomasz Trzcinski. 2018. Pay Attention to Virality: understanding popularity of social media videos with the attention mechanism. *arXiv preprint arXiv:1804.09949* (2018).
- [11] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. 77–91.
- [12] Christian Ehret, Jacy Boegel, and Roya Manuel-Nekouei. 2018. The Role of Affect in Adolescents' Online Literacies: Participatory Pressures in BookTube Culture. *Journal of Adolescent & Adult Literacy* 62, 2 (2018), 151–161.
- [13] Florian Eyben and Björn Schuller. 2015. openSMILE: the Munich open-source large-scale multimedia feature extractor. *ACM SIGMultimedia Records* 6, 4 (2015), 4–13.
- [14] Daniel Gatica-Perez, Dairazalia Sanchez-Cortes, Trinh Minh Tri Do, Dinesh Babu Jayagopi, and Kazuhiro Otsuka. 2018. Vlogging Over Time: Longitudinal Impressions and Behavior in YouTube. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*. ACM, 37–46.
- [15] Sharad Goel, Ashton Anderson, Jake Hofman, and Duncan J Watts. 2015. The structural virality of online diffusion. *Management Science* 62, 1 (2015), 180–196.
- [16] Rosanna E Guadagno, Daniel M Rempala, Shannon Murphy, and Bradley M Okdie. 2013. What makes a video go viral? An analysis of emotional contagion and Internet memes. *Computers in Human Behavior* 29, 6 (2013), 2312–2319.
- [17] Yağmur Güçlütürk, Umut Güçlü, Marcel AJ van Gerven, and Rob van Lier. 2016. Deep impression: Audiovisual deep residual networks for multimodal apparent personality trait recognition. In *European Conference on Computer Vision*. Springer, 349–358.
- [18] Edward Twitchell Hall. 1989. *Beyond culture*. Anchor.
- [19] Melina Hughes. 2017. BookTube and the Formation of the Young Adult Canon. (2017).
- [20] Andreas M Kaplan and Michael Haenlein. 2010. Users of the world, unite! The challenges and opportunities of Social Media. *Business horizons* 53, 1 (2010), 59–68.
- [21] Mark L Knapp, Judith A Hall, and Terrence G Horgan. 2013. *Nonverbal communication in human interaction*. Cengage Learning.
- [22] Behnaz Nojavanasghari, Deepak Gopinath, Jayanth Koushik, Tadas Baltrušaitis, and Louis-Philippe Morency. 2016. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 284–288.
- [23] Sunghyun Park, Han Suk Shim, Moitreyia Chatterjee, Kenji Sagae, and Louis-Philippe Morency. 2014. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 50–57.
- [24] Frank Pasquale. 2018. When machine learning is facially invalid. *Commun. ACM* 61, 9 (2018), 25–27.
- [25] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. *The development and psychometric properties of LIWC2015*. Technical Report.
- [26] Kathryn Perkins. 2017. The Boundaries of BookTube. *The Serials Librarian* 73, 3-4 (2017), 352–356.
- [27] A Booktubing Primer. 2017. Booktubing. *ALAN Review* (2017), 61.
- [28] Shatha N Samman, Michael Moshell, Bryan Clark, and Chantel Brathwaite. 2009. *Learning to decode nonverbal cues in cross-cultural interactions*. Technical Report. GLOBAL ASSESSMENT LLC ORLANDO FL.

- [29] Björn Schuller, Stefan Steidl, Anton Batliner, Elmar Nöth, Alessandro Vinciarelli, Felix Burkhardt, Rob van Son, Felix Weninger, Florian Eyben, Tobias Bocklet, et al. 2012. The interspeech 2012 speaker trait challenge. In *Thirteenth Annual Conference of the International Speech Communication Association*.
- [30] Karen Sorensen and Andrew Mara. 2014. Booktubers as a Networked knowledge community. In *Emerging Pedagogies in the Networked Knowledge Society: Practices Integrating Social Media and Globalization*. IGI Global, 87–99.
- [31] Arulkumar Subramaniam, Vismay Patel, Ashish Mishra, Prashanth Balasubramanian, and Anurag Mittal. 2016. Bi-modal first impressions recognition using temporally ordered deep audio and stochastic visual features. In *European Conference on Computer Vision*. Springer, 337–348.
- [32] Lucía Teijeiro-Mosquera, Joan-Isaac Biel, José Luis Alba-Castro, and Daniel Gatica-Perez. 2015. What your face vlogs about: expressions of emotion and big-five traits impressions in YouTube. *IEEE Transactions on Affective Computing* 6, 2 (2015), 193–205.
- [33] Tomasz Trzciński and Przemysław Rokita. 2017. Predicting popularity of online videos using support vector regression. *IEEE Transactions on Multimedia* 19, 11 (2017), 2561–2570.
- [34] David Vallet, Shlomo Berkovsky, Sebastien Ardon, Anirban Mahanti, and Mohamed Ali Kafaar. 2015. Characterizing and predicting viral-and-popular video content. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*. ACM, 1591–1600.
- [35] Fons J. R. van de Vijver. 2017. *Nonverbal Communication across Cultures*. American Cancer Society, 1–10. <https://doi.org/10.1002/9781118783665.ieicc0252>
- [36] Mirjam Wattenhofer, Roger Wattenhofer, and Zack Zhu. 2012. The YouTube Social Network.. In *ICWSM*.
- [37] Erroll et al. Wood. 2015. Rendering of eyes for eye-shape registration and gaze estimation. In *Proceedings of the IEEE International Conference on Computer Vision*. 3756–3764.
- [38] Koji Yatani. 2016. Effect Sizes and Power Analysis in HCI. In *Modern Statistical Methods for HCI*. Springer, 87–110.
- [39] Chen-Lin Zhang, Hao Zhang, Xiu-Shen Wei, and Jianxin Wu. 2016. Deep bimodal regression for apparent personality analysis. In *European Conference on Computer Vision*. Springer, 311–324.