Probabilistic Topic Models for Human Emotion Analysis

Prasanth Lade
Arizona State University
prasanthl@asu.edu

Vineeth N Balasubramanian
IIT, Hyderabad
vineethnb@iith.ac.in

Sethuraman Panchanathan
Arizona State University
panch@asu.edu

Abstract

Arousal and valence are two important dimensions that can capture the spectrum of human emotions and their prediction is a challenging task. In this work, we used existing and novel probabilistic topic models to understand and analyze human emotions from multimodal data. We extracted latent facial and vocal topic features from video and audio frames using existing topic models to predict arousal and valence. We evaluated these higher level features on three tasks viz. topic visualization, temporal topic evolution and emotion recognition. A new topic model based on Supervised Latent Dirichlet Allocation is proposed specifically for predicting changes in arousal and valence and is comparatively evaluated against existing state-of-the-art topic models.

1 Introduction

To facilitate a deeper analysis of human emotions, the affective analysis community has recently commenced working on continuous emotions instead of discrete emotions like sad, happy, disgusted, etc. There are two popular continuous emotion dimensions viz. arousal and valence, where arousal measures the amount of energy in the emotion and valence measures how negative or positive a given emotion is. Arousal and valence are typically rated in a range of -1 to +1 and can be any rational number between them. Selecting features that can predict these dimensions can be a very significant task in the analysis of human emotion. In this work, we evaluated latent facial and vocal topic features extracted using existing and novel probabilistic topic models towards the objective of human emotion analysis. These topic features are extracted from popular features (base features) used in existing emotion recognition work. We validated our features on two applications, viz. emotion recognition and emotion change detection. In emotion recognition, given a facial video, we predict arousal and valence at each frame; and in change detection, we predict whether a change has occurred or not in these 2 dimensions in a real-world multimodal data stream. We used Latent Dirichlet Allocation (LDA) [1] and supervised LDA (sLDA) [2] with the proposed features for the emotion recognition task; whereas for change detection, we proposed a new model called supervised LDA for change detection (sLDACd) which we compared against existing models.

2 Emotion Recognition

The arousal and valence of a person interacting in a video can be labeled using his/her facial and speech features. Among facial features, Local Binary Patterns (LBP) (appearance features), facial landmarks from Active Shape Models (shape features) have been extensively used for emotion recognition. Low level descriptors of prosodic, spectral and energy features are frequently used audio features for emotion recognition. Our objective is to extract these features at each time step and predict the real-valued quantities, arousal and valence. We hypothesize that topic models can be used to extract meaningful patterns that can: (i) efficiently predict emotions, (ii) be visualizable;

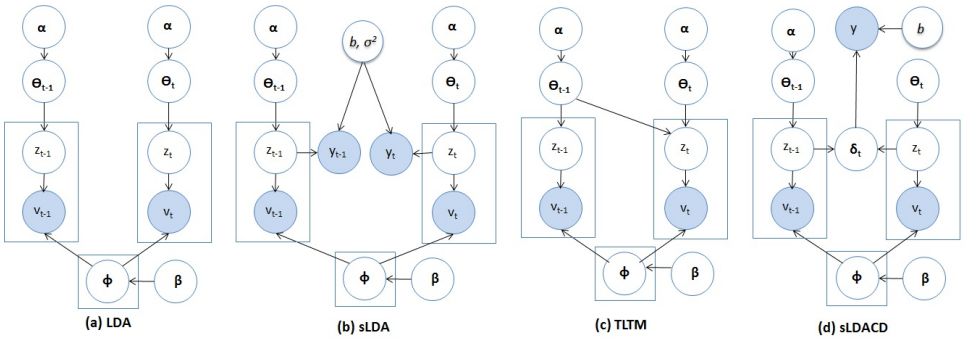

Figure 1: Different probabilistic topic models used in this work

and (iii) evolve with changes in emotion over time. We used LDA and sLDA to extract latent topic features for the emotion recognition task. To extract topic features from videos, we consider each image/audio frame as a document, and quantize the image/audio features as explained below.

2.1 Feature Quantization

Shape Features: 68 facial landmarks are extracted by tracking facial images, and the changes in x and y values of a point from its neutral position are calculated. We discretize the relative movement of landmarks into r and θ where θ is the direction of movement and r is the quantized scale of drift. θ is quantized to 4 bins viz. top left, top right, bottom left and bottom right. For example, given a landmark 18 that moved into top right, the corresponding word will be 18TopRight. This word is then repeated r times in the image document and the size of the vocabulary is 68*4 i.e. 272.

Appearance Features: LBP features are extracted from each image by dividing it into 10 x 10 blocks and appending histograms of 59 binary patterns calculated from each image block. We generate a word from each LBP in a block and repeat it using its histogram frequency. E.g. a word corresponding to LBP 6 in block 10 is given an id ((10-1)*59+6) and the frequency is given by the histogram of block 10. The size of the vocabulary is 5900 (59 LBPs from 100 blocks).

Audio Features: The audio features are real-valued descriptors and hence, we consider each feature and quantized all possible values to 50 bins using K-Means clustering. The bin to which a feature value belongs becomes a word. So for a total of 1242 descriptors in an audio document we generated 1242 words and the total size of the vocabulary is 62101.

2.2 Evaluation

Classification using Shape features: We tested our topic features obtained using the LDA model on the CK+ dataset that has 327 image sequences in which the final image is annotated with one of 7 emotions viz. sad, happy, angry, disgust, contempt and surprise (a classification problem). We used ASM features from the final image of a video, generated image documents and extracted topics using LDA (as described above). The obtained topics are visualized in the top part of Figure 2(a). The red point indicates the word and the length of blue line indicates the probability of the word in the topic. We did a 118 fold subject independent validation for 7 emotions and we achieved an accuracy of 85.62% against an accuracy of 66.68% given by the baseline ASM features.

Regression using Appearance and Audio features: For a deeper analysis we used the AVEC-12 dataset [4], which has 31 training and 32 development videos containing facial videos of people, annotated with arousal and valence values at each time step. We used LDA and sLDA models with Collapsed Gibbs sampling with LBP and audio features on this regression problem. For both LDA and sLDA we selected the topic size as K=50 using cross validation on training videos and α = 50/K. We trained using the training videos and tested on development videos. Plots of topics generated from appearance features are shown in the bottom of Figure 2(a) where the most active blocks of a topic are highlighted. In Figure 2(b) we have plotted the arousal of a video and the probability of Topic 30 (shown in inset) as time evolves and we see that the intensity of topic around the lip

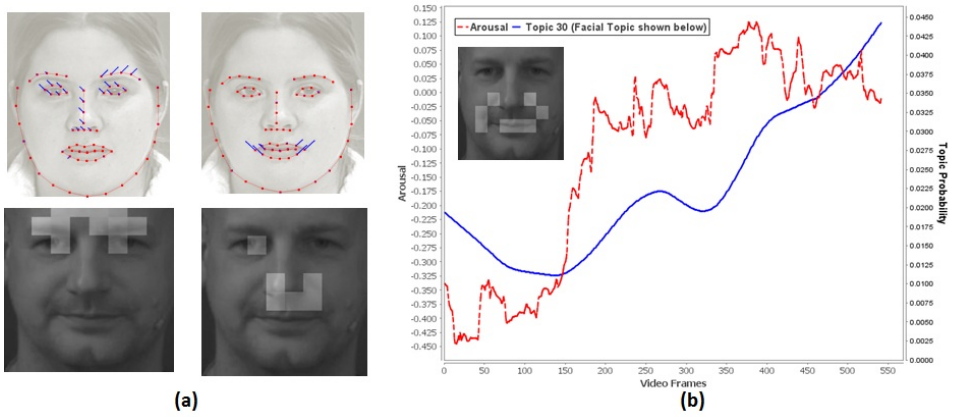

Figure 2: (a) Topics from shape, appearance features (b) Evolution of a topic with Arousal

Table 1: Mean Cross correlations using Topic features

Dimension	Feature	B-LR	B-SVR	LDA-LR	LDA-SVR	SLDA	PCA-LR	PCA-SVR
Arousal	Audio	0.285*	0.263*	0.287	0.272	**0.3**	0.049	0.045
	LBP	0.103	0.151	0.129	**0.18**	0.149	0.093	0.171
Valence	Audio	0.1	0.09	0.1	**0.126**	0.096	0.0115	0.011
	LBP	0.203	0.207	0.21	0.25	**0.303**	0.325*	0.291*

region has a direct effect on arousal. Table 1 shows the cross-correlations of predicted arousal and valence against the actual values averaged across all development videos (B=baseline, LR=Linear regression, SVR=Support Vector Regression; We also compared topic models with Principal Component Analysis - PCA - to evaluate them as dimensionality reduction techniques). We observe that LDA and sLDA have uniformly performed well across all dimensions but B and PCA outperform in arousal and valence prediction respectively. In results indicated with a *, the corresponding models predicted straight lines and due to the way correlations are calculated, straight lines that match the overall slope of the signal get high values. But we hypothesized that topic models predict the changes in the streams; to validate this, we evaluated them on a change detection task (below).

3 Emotion Change Detection

Detecting changes in emotion can play a significant role in identifying coherent substreams of data which can further be used for emotion recognition. Since no datasets exist with labeled emotion changes, we used the Cumulative Sum based CumSum method [5] to annotate arousal and valence signals for all AVEC12 videos with changes. We compared the performance of our topic features with PCA-based features and base features for change detection.

3.1 SLDACD model

For predicting change points, we need a dependency between two consecutive documents and since both LDA and sLDA models do not model this dependency we created a new supervised topic model based on SLDA which we call the sLDA for Change Detection as shown in Figure 1(d). In this model we assume that at a given time step, the change in emotion depends on both current and previous normalized topic probabilities $\bar{z}_{t-1}$ and $\bar{z}_t$ through the variable $\boldsymbol{\delta_t}$. y_t is the observed variable assuming 0 or 1 depending on whether change has occurred, and is modeled as a Logistic function of $\boldsymbol{\delta_t}$ and $\boldsymbol{b}$ where $\boldsymbol{b}$ are the regression coefficients. For each topic k, δ_{tk} is calculated from z_{t-1k} and z_{tk} as: (i) Scale z_{t-1k} and z_{tk} from $[0, 1]$ to $[min, max]$ where $min, max > 1$; (ii) Calculate the differences $a_1 = |z_{t-1k} - z_{tk}|$ and $a_2 = max - a_1$; and (iii) Calculate $\delta_{tk} = a_2/(a_1 + a_2)$. Note that for each topic k, δ_{tk} is the expected value of the $Beta(a_1, a_2)$ distribution. If the topic probabilities z_{t-1k} and z_{tk} are very close then a_1 increases and a_2 decreases and thus the distribution $Beta(a_1, a_2)$ will be right-skewed with a mean that moves towards 1 and vice versa.

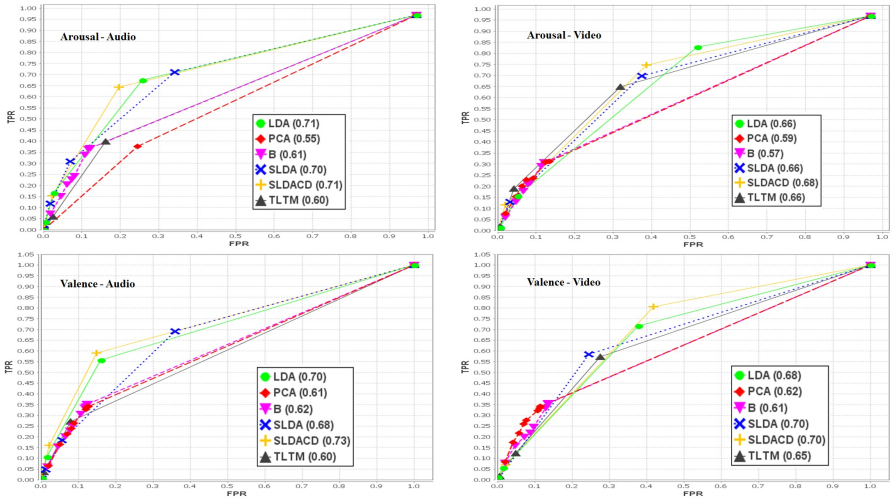

Figure 3: ROC curves for change detection with Area under curve values for each model in paranthesis

Thus δ_{tk} can be seen as the probability of z_{t-1k} and z_{tk} being similar. We use this topic wise similarity probability vector as features to predict change. To sample a topic z_{tw} for a word w of document t we use Collapsed Gibbs sampling equations based on LDA model but with an extra logistic probability as given below:

$$p(z_{tw} = k|\boldsymbol{v}, y_t, \bar{z}_{\neg tw}, \alpha, \beta) \sim (n_{tk}^{\neg w} + \alpha)\frac{(n_{wk}^{w} + \beta)}{(n_k + W\beta)}p(y_t|\boldsymbol{\delta_t^{\neg w}}, \boldsymbol{b})$$

where all the count variables have same meaning as in LDA and $p(y_t = 1|\boldsymbol{\delta_t^{\neg w}}, \boldsymbol{b}) \sim 1/(1 + exp(\boldsymbol{b}\delta_t^{\neg w})exp(b_k/N_t))$. We used iterative sampling similar to EM, where in the E-step we estimated the $\bar{z}_t$ and in the M-step we estimated the coefficients $\boldsymbol{b}$ using logistic regression.

3.2 Evaluation

We used the AVEC12 dataset for change detection and annotated changes at every frame using CumSum method. We used LDA, sLDA, TLTM (a topic model where the current word is influenced by both $\boldsymbol{\theta_t}$ and $\boldsymbol{\theta_{t-1}}$), our sLDACd, PCA and Baseline (B) methods. sLDACd predicts 0/1 directly and for other topic models we generated topic features, calculated $\boldsymbol{\delta_t}$ for each document and then used logistic regression for predictions. For Baseline and PCA methods, we used the actual predictions of arousal and valence and extracted changes at different confidence intervals. The ROC curves are shown in Figure 3. We observe that the Baseline and PCA methods do not perform as well as topic models on change detection, indicating that they possibly do not predict variations in emotions. The proposed sLDACd method gave the best AUC values for most dimensions except for arousal-audio and valence-video where LDA and sLDA equally performed well.

Conclusion: In this work, we explored the use of topic models for emotion analysis, and showed that topic features are visualizable and effective for not only emotion recognition, but also for change detection (such as detection of mood swings). In future we plan to explore strategies to transfer topics learnt from one set of videos to another and use optical motion flow as base features.

References

[1] Blei, D.M., Ng, A.Y. & Jordan, M.I. (2003) Latent dirichlet allocation. *JMLR*, (3): 993-1022.

[2] Blei, D.M. & AcAuliffe, J.D. (2007) Supervised topic models. *NIPS*.

[3] Shang, L & Chan, K.P. (2011) A temporal latent topic model for facial expression recognition. *ACCV*.

[4] Schuller, B, Valstar, M. et al (2012) Avec 2012 the continuous audio/ visual emotion challenge. *ICMI*.

[5] Wayne, A.T. http://www.variation.com/cpa/tech/changepoint.html