Temporally Selective Attention Model for Social and Affective State Recognition in Multimedia Content

ABSTRACT

The sheer amount of human-centric multimedia content has led to increased research on human behavior understanding. Most existing methods model behavioral sequences without considering the temporal saliency. This work is motivated by the psychological observation that temporally selective attention enables the human perceptual system to process the most relevant information. In this paper, we introduce a new approach, named Temporally Selective Attention Model (TSAM), designed to selectively attend to salient parts of human-centric video sequences. Our TSAM models learn to recognize affective and social states using a new loss function called speaker-distribution loss. Extensive experiments show that our model achieves the state-of-the-art performance on rapport detection and multimodal sentiment analysis. We also show that our speaker-distribution loss function can generalize to other computational models, improving the prediction performance of deep averaging network and Long Short Term Memory (LSTM).

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; *Supervised learning by regression*; • **Human-centered computing** → *Empirical studies in HCI*;

KEYWORDS

Affective state recognition; Temporally selective attention; Speaker-distribution loss

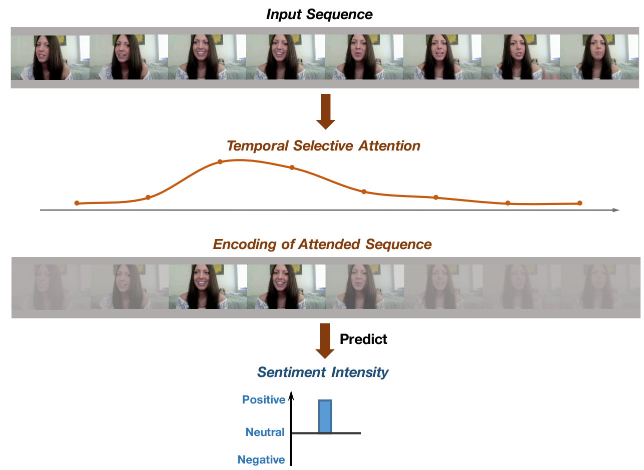


Figure 1: An illustration of the attention and encoding steps in our Temporally Selective Attention Model (TSAM). To understand the speaker's affect, we first identify the task-relevant parts in the input sequence and then encode the sequence while filtering the non-relevant parts (shown as gray shading). Our TSAM approach allows to recognize affect and social states from unsegmented video sequences.

1 INTRODUCTION

The success of video-sharing and social network websites has led to greatly increased posting of online multimedia content, with a large proportion of the these videos being human-centric. The sheer amount of such data promotes research on behavior understanding that can effectively discover the affective and social states within human-centric multimedia content. Various applications can benefit from this behavior understanding. Multimodal sentiment analysis allows for mining large numbers of online videos to extract the expressed opinions about products or movies [34]. In education, with the advent of online learning platforms, students are interacting increasingly remotely with peers and tutors. Better understanding of the social dynamics during these remote interactions has the potential to increase engagement and learning gains [56].

Automatically recognizing affective and social states in multimedia contents has some unique characteristics which bring new technical challenges. The first characteristic of recognizing affective and social states, such as users' mood, sentiment, or rapport, is that they are usually perceived over a long period of time. For example, previous work trying to recognize rapport, i.e. a harmonious relationship in which people are coordinated and understand each other, annotated the ground truth of rapport with a minimum of 30-second time windows [57]. This first characteristic brings with it the technical challenge that not everything happening during the video-recorded interaction will be relevant to recognize the affective and social states. According to the psychologists [39][29], the human perceptual system is able to process the most relevant information by the rapid modulation of *temporally selective attention*. Most existing approaches in affective multimedia analysis do not address this issue. Many researchers simply compute summary statistics of behavior features over the whole video [36]. In recent emotion recognition approaches, these systems will either work on very short segments or even individual frames [58], or process sequentially all available frames in the video sequence without a temporal attention process [3]. With the recent advances in recurrent neural networks, LSTM (Long Short-Term Memory) [14] models are gaining popularity in affective computing and were applied to affect recognition in multimedia contents [36, 43, 49]. While LSTM models are great at memorizing sequence information, they do not include an explicit mechanism to perform temporally selective attention.

A second characteristic of social and affective datasets is that they often contain more than one training sequence with the same

Conference'17, Washington, DC, USA
2017. 123-4567-24-567/08/06…$15.00
DOI: 10.475/123_4

speaker (or with the same dyad if the dataset contains dyadic social interactions). The conventional approach for training recognition models is to ignore this fact and learn the model parameters using a loss function which sum over all sequences, independent of the speaker grouping. For example, the square loss function will penalize differences between predictions and ground truth labels for each training sequence individually and then sum all these squared differences. These conventional loss functions do not take advantage of the natural grouping found in social and affective datasets. For example, when learning a rapport level predictor, predictions from sequences of a friend dyad should have a different distribution than if these sequences were from a stranger dyad.

In this paper, we propose a novel approach, named Temporally Selective Attention Model (TSAM), designed to infer the social and affective states in unsegmented multimedia contents (see Figure 1). TSAM's attention mechanism localizes the task-relevant part of the input sequence and filters out the noisy time-steps. Our TSAM approach is composed of three components: the *attention module*, the *encoding module* and the *speaker-distribution loss*. The attention module localizes the task-relevant part from the input sequence, allowing us to filter out the noisy or irrelevant time-steps. The encoding module integrates the attention scores to represent the sequence. Finally, our speaker-distribution loss function encourages the model predictions for a specific speaker (or dyad) to follow the same distribution of that speaker's ground truth labels.

In summary, our proposed temporally selective attention model has the following advantages over prior work:

(1) It automatically localizes the task-relevant parts from the unsegmented multimedia sequences, improving the performance for affective and social state recognition.
(2) The attention scores, inferred by our TSAM model, are easily interpretable and allow to identify the relevant input observations.
(3) Our proposed speaker-distribution loss function takes advantage of speaker's individual label distribution during training. Our experiments show that it generalizes to other computational models.
(4) Our TSAM model outperforms previous state-of-the-art algorithms for two multimedia datasets: multimodal sentiment analysis with monadic interactions and rapport level estimation with dyadic interactions. We also show generalization of our attention and encoding modules on the widely popular task of text-only sentiment analysis.

The structure of this paper is as follows. We first discuss the related work in Section 2. Our model is introduced in Section 3. In Section 4 and 5, we evaluate our model and compare it to the baseline methods. The paper is concluded in Section 6.

2 RELATED WORK

2.0.1 Affective Computing. Affective phenomena such as emotions, moods, and personality modulate our everyday interactions with our friends, family members and colleagues [40]. In the broad sense, affective phenomena include a large number of processes, states, and traits. Of these, emotions are often characterized as relatively short-term reactions to an event, and are contrasted to moods which are also less event-centric. Preferences, attitudes and sentiments often represent a judgment or disposition toward a specific object or stimulus. In the community of multimedia affective computing, *emotion recognition* and *sentiment analysis* are the most widely-studied research areas.

Automatic emotion recognition performs affect classification from a predefined taxonomy of emotions. It plays a crucial role in tasks within affect-sensitive human computer interaction (HCI) systems, customer services, intelligent automobile systems, and entertainment industries [55]. Most researchers studying emotion recognition use audio [37][48][12], video [17], physiological signals [22], or the combination of multiple modalities [18][28].

Sentiment analysis [30] is a widely studied topic in natural language processing. Functionally, research in sentiment analysis can be split into word-level sentiment word identification [8], document-level opinion mining [9], and aspect-level sentiment classification [27]. Early work [45][15] was mostly based on hand-crafted sentiment lexicons, which are hard to collect. The recent trend of deep learning has enabled various kinds of neural network models for sentiment classification. This includes semantic compositionality [42], sentiment embeddings [47], and memory network [46]. With the advent of mobile social media, people are sharing ever-greater quantities of video, image and audio data, in addition to text. As such, there is an increasing number of datasets explicitly designed for *multimodal sentiment analysis*, including YouTube[25], MOUD[33], ICT-MMMO[51], and MOSI[53]. As such, there is also a growing body of work concerned with *multimodal sentiment analysis* [35][54]. The state-of-the-art performance is achieved by Wang et al. [50] that aims to improve the generalizability of neural networks across datasets.

2.0.2 Social Interaction. Recent work has also studied various aspects of interpersonal social dynamics, in addition to intrapersonal affect modeling. Zhao et al. [57] developed a dyadic rapport detector for reciprocal peer tutoring. Based on the human-annotated social strategies, this paper focused on the discovery of temporally co-occurring and contingent behavioral patterns that signal interpersonal rapport. Neubauer et al. [26] proposed a method to assess team behaviors that develop resilience to stress by utilizing nonverbal and linguistic measures from team members. Damian et al. [6] explored how automatic behavioural feedback loops during social interactions might enable users to improve their behavior quality by analyzing their behaviours in real time. Moreover, multimodal interaction analysis have been proposed for health care systems, e.g. detection of early stages of dementia [20].

2.1 Recurrent Neural Networks

Recurrent neural networks (RNNs) are a generalization of feed-forward neural networks sharing weights on variable lengths of sequences. Gated Recurrent Unit (GRU) [4] and Long Short-Term Memory (LSTM) [14] are among the most popular architectures due to their effective solutions to the vanishing gradient problem. Specifically, LSTM can keep long-term memory by training proper gating weights. The fundamental idea, using a memory cell updating and storing the information, makes LSTM capture long-distance dependencies more effectively than standard RNNs. Its effectiveness

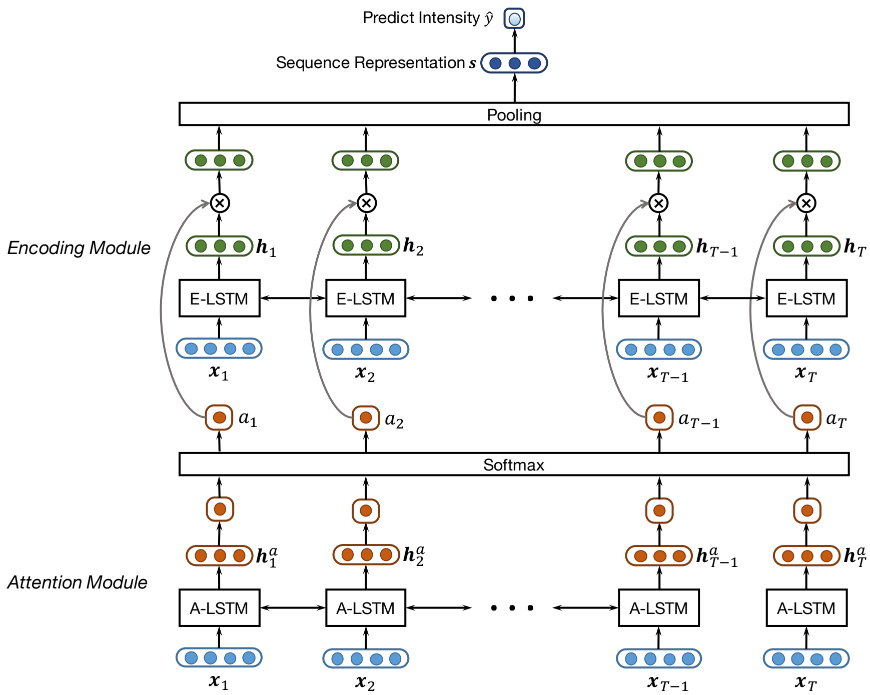


Figure 2: The framework of the temporally selective attention model. The attention module selects the attended steps with attention weights. The encoding module encodes the time-steps, and represent the sequence by integrating the attention weights over all encoded steps. The prediction is determined based on the sequence representation.

has been empirically shown on a wide range of problems, including machine translation [44], speech recognition [13], dependency parsing [10], and video activity detection [24], etc.

2.2 Attention Models

Attention mechanisms are an effective way for neural networks to enhance their capability and interpretability. In visual question answering, attention networks allow the model to locate the objects and concepts referred to in the question [52]. In summarization and machine translation, an attention-based encoder is developed to learn a latent alignment over the input text [38][1]. In aspect-level sentiment analysis, the attention gates enable the model to concentrate on the key parts of a sentence, given the aspect.

Pei et al. [31] is the recent work most relevant this paper. They deployed the attention mechanism to RNNs. The recurrent attention-gated units accumulate the summative hidden states, and represent the sequence as the last state. In their model, if a time-step is assigned with a high attention value during the temporal encoding, the model would forget the previous time-steps. Our model avoids such information information loss since our encoding module encodes the sequence with weighted attention.

3 TEMPORALLY SELECTIVE ATTENTION MODEL

In this section, we discuss our Temporally Selective Attention Model (TSAM) for social and affective state recognition. The temporally selective attention mechanism enables our model to localize the task-relevant part of the input sequence. Figure 2 shows an overview of our framework. Our TSAM model consists of three components: (1) an attention module that determines the attention scores indicating the relevance of each time-step, (2) an encoding module that represents the whole sequence by integrating the attention weights, (3) the speaker-distribution loss function that shapes the predictive distribution to take advantage of natural speaker grouping in the training set.

To formally define our TSAM approach, we will focus on regression problems where the affective or social state is described with a real-value. As shown later in our experiments, our TSAM models can easily be extended to classifications tasks with a discrete set of state labels. We define the input sequence as $\mathbf{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_T\}$ where $\mathbf{x}_t \in \mathbb{R}^D$ is the feature vector representing the t-th time-step and T is the sequence length. Our model predicts the real-value affective or social state $y \in \mathbb{R}$.

In this section, we first present our attention module to identify relevant parts in unsegmented sequences. Then we discuss how to obtain the sequence representation through a detailed discussion of each module in the framework. Finally, we present the speaker-distribution loss and compare it to the standard square loss function.

3.1 Attention Module

Since not every time-step of the sequence is relevant for the prediction, the model should extract the salient parts from the noisy time-steps. For example, to detect the dyadic rapport, the model is expected to deal with 30-second video segments of conversation, containing 900 frames (at an average frame rate of 30 fps). The attention mechanism helps the model to select the salient time-steps by explicitly assigning attention weights.

Our attention module takes advantage of bi-directional Long-Short Term Memory (LSTM) network to preprocess the sequence. LSTM is able to process an input sequence via the recursive application of a transition function. To address the problem of vanishing gradient, the LSTM model uses a memory cell and a hidden state variable that are passed from one unit to the next one.

Let the dimensionality of the hidden state variable for both forward and backward LSTMs be D_H. The hidden state output of each time-step is denoted as $\mathbf{h}_t^a = [\overrightarrow{\mathbf{h}}_t^a; \overleftarrow{\mathbf{h}}_t^a] \in \mathbb{R}^{2D_H}$, the concatenation of hidden outputs of the left-to-right LSTM $\overrightarrow{\mathbf{h}}_t^a \in \mathbb{R}^{D_H}$ and the right-to-left LSTM $\overleftarrow{\mathbf{h}}_t^a \in \mathbb{R}^{D_H}$. $\overrightarrow{\mathbf{h}}_t^a$ and $\overleftarrow{\mathbf{h}}_t^a$ are calculated as following:

$$\begin{pmatrix} \overrightarrow{\mathbf{i}_t} \\ \overleftarrow{\mathbf{i}_t} \end{pmatrix} = \sigma \begin{pmatrix} \overrightarrow{\mathbf{W}}_i \mathbf{x}_t + \overrightarrow{\mathbf{U}}_i \overrightarrow{\mathbf{h}}_{t-1}^a + \overrightarrow{\mathbf{b}_i} \\ \overleftarrow{\mathbf{W}}_i \mathbf{x}_t + \overleftarrow{\mathbf{U}}_i \overleftarrow{\mathbf{h}}_{t-1}^a + \overleftarrow{\mathbf{b}_i} \end{pmatrix} \tag{1}$$

$$\begin{pmatrix} \overrightarrow{\mathbf{f}_t} \\ \overleftarrow{\mathbf{f}_t} \end{pmatrix} = \sigma \begin{pmatrix} \overrightarrow{\mathbf{W}}_f \mathbf{x}_t + \overrightarrow{\mathbf{U}}_f \overrightarrow{\mathbf{h}}_{t-1}^a + \overrightarrow{\mathbf{b}_f} \\ \overleftarrow{\mathbf{W}}_f \mathbf{x}_t + \overleftarrow{\mathbf{U}}_f \overleftarrow{\mathbf{h}}_{t-1}^a + \overleftarrow{\mathbf{b}_f} \end{pmatrix} \tag{2}$$

$$\begin{pmatrix} \overrightarrow{\mathbf{o}_t} \\ \overleftarrow{\mathbf{o}_t} \end{pmatrix} = \sigma \begin{pmatrix} \overrightarrow{\mathbf{W}}_o \mathbf{x}_t + \overrightarrow{\mathbf{U}}_o \overrightarrow{\mathbf{h}}_{t-1}^a + \overrightarrow{\mathbf{b}_o} \\ \overleftarrow{\mathbf{W}}_o \mathbf{x}_t + \overleftarrow{\mathbf{U}}_o \overleftarrow{\mathbf{h}}_{t-1}^a + \overleftarrow{\mathbf{b}_o} \end{pmatrix} \tag{3}$$

$$\begin{pmatrix} \overrightarrow{\mathbf{u}_t} \\ \overleftarrow{\mathbf{u}_t} \end{pmatrix} = tanh \begin{pmatrix} \overrightarrow{\mathbf{W}}_u \mathbf{x}_t + \overrightarrow{\mathbf{U}}_u \overrightarrow{\mathbf{h}}_{t-1}^a + \overrightarrow{\mathbf{b}_u} \\ \overleftarrow{\mathbf{W}}_u \mathbf{x}_t + \overleftarrow{\mathbf{U}}_u \overleftarrow{\mathbf{h}}_{t-1}^a + \overleftarrow{\mathbf{b}_u} \end{pmatrix} \tag{4}$$

$$\begin{pmatrix} \overrightarrow{\mathbf{C}_t} \\ \overleftarrow{\mathbf{C}_t} \end{pmatrix} = \begin{pmatrix} \overrightarrow{\mathbf{f}_t} \times \overrightarrow{\mathbf{C}_{t-1}} + \overrightarrow{\mathbf{i}_t} \times \overrightarrow{\mathbf{u}_t} \\ \overleftarrow{\mathbf{f}_t} \times \overleftarrow{\mathbf{C}_{t-1}} + \overleftarrow{\mathbf{i}_t} \times \overleftarrow{\mathbf{u}_t} \end{pmatrix} \tag{5}$$

$$\begin{pmatrix} \overrightarrow{\mathbf{h}}_t^a \\ \overleftarrow{\mathbf{h}}_t^a \end{pmatrix} = \begin{pmatrix} \overrightarrow{\mathbf{o}_t} \times tanh(\overrightarrow{\mathbf{C}_t}) \\ \overleftarrow{\mathbf{o}_t} \times tanh(\overleftarrow{\mathbf{C}_t}) \end{pmatrix} \tag{6}$$

where $\times$ denotes the element-wise product, and $\sigma(\cdot)$ denotes the sigmoid function. $(\overrightarrow{\mathbf{i}_t}, \overleftarrow{\mathbf{i}_t})$, $(\overrightarrow{\mathbf{f}_t}, \overleftarrow{\mathbf{f}_t})$, $(\overrightarrow{\mathbf{o}_t}, \overleftarrow{\mathbf{o}_t})$ are the input gates, forget gates, and output gates respectively. $\{\overrightarrow{\mathbf{W}}_z, \overleftarrow{\mathbf{W}}_z, \overrightarrow{\mathbf{U}}_z, \overleftarrow{\mathbf{U}}_z, \overrightarrow{\mathbf{b}_z}, \overleftarrow{\mathbf{b}_z}\}_{z \in \{i,f,o,u\}}$ are the LSTM parameters. $\overrightarrow{\mathbf{C}_t}, \overleftarrow{\mathbf{C}_t}$ are the memory cells at time-step t.

Collecting the processed sequence, the attention weight vector $\mathbf{a} \in \mathbb{R}^T$ is then computed as:

$$\mathbf{a} = softmax(\mathbf{H}^{a\intercal} \mathbf{w}_a), \tag{7}$$

where $\mathbf{H}^a \in \mathbb{R}^{2D_H \times T}$ is the matrix composed by the hidden vectors $[\mathbf{h}_1^a, \cdots, \mathbf{h}_T^a]$. $\mathbf{w}_a \in \mathbb{R}^{2D_H}$ is the projection vector which will be jointly trained with LSTM parameters. The element a_t in vector $\mathbf{a}$ represents the attention weight for step t.

3.2 Encoding Module

In this section, we train a second bi-directional LSTM to encode the all the sequence observations from $\mathbf{X}$. Let the hidden state outputs of the bi-LSTM be $[\mathbf{h}_1, \cdots, \mathbf{h}_T]$, where $\mathbf{h}_t = [\overrightarrow{\mathbf{h}}^t; \overleftarrow{\mathbf{h}}^t]$ denotes the outputs of the t-th LSTM unit calculated similar to Equation (1) - (6).

Unlike most prior work using the last output $\mathbf{h}_T$ of the LSTM, our model represents $\mathbf{X}$ as the attention-weighted combination of all outputs computed from the encoding module. We derive the final representation $\mathbf{s}$ of the sequence as:

$$\mathbf{s} = \sum_{t=1}^{T} a_t \mathbf{h}_t. \tag{8}$$

As for prediction, we calculate the predicted score $\hat{y}$ by projecting the representation to a real-value scalar:

$$\hat{y} = \mathbf{w}_s^\intercal \mathbf{s} + b_s. \tag{9}$$

3.3 Speaker-Distribution Loss Function

In this section, we introduce a new loss function for regression models named Speaker-Distribution Loss (SDL). The intuition behind this loss function is to take advantage of the natural grouping often present in affective and social datasets. These datasets will often contain more than one labeled sequence for each speaker. For social interaction datasets, the same dyad may have more than one labeled sequence. Our speaker-distribution loss function takes advantage of this natural grouping to improve the distribution of the predicted labels.

A second motivating factor of our speaker-distribution loss function is that we observed empirically that common loss functions such as square loss may end up being too conservative in their prediction and always predict the average sequence label. Our speaker-distribution loss function encourages the model's predictions to follow the same distribution as the training data. A regression model which always predict the average label will be penalized (unless all training samples have the same label). Our speaker-distribution loss function goes a step further by performing this enforcement in a speaker-spcific manner. The model's predictions for a specific speaker (or dyad) should follow the same distribution of that speaker's ground truth labels.

Problem Formulation Suppose we are given a training set $\mathcal{D} = \{(\mathbf{X}_1, y_1), \cdots, (\mathbf{X}_n, y_n)\}$ containing n sequences $\mathbf{X}_i$ with variable lengths and their corresponding labels y_i. The traditional way of calculating the loss function is to aggregate the square distances of all pairs of prediction and ground-truth $(\hat{y}_i, y_i)$. Here we define "*discrepancy*" of $\hat{y}$ and y as the square distance measurement, i.e. $\delta(\hat{y}, y) = \|\hat{y} - y\|^2$. The square loss can be written as:

$$\mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \delta(\hat{y}_i, y_i) = \frac{1}{n} \sum_{i=1}^{n} \|\hat{y}_i - y_i\|^2. \tag{10}$$

Although Equation (10) is the common choice for most regression tasks, minimizing it does not always guarantee high correlation between predictions and ground-truth values. As the square penalty is sensitive to minor changes of the distance, it is highly possible to induce conservative predictions that close to the mean values $\hat{y}_i \approx \frac{1}{n} \sum_{i=1}^{k} y_i$ with small variations. To deal with this, some models (e.g. SVR [41]) relax the penalty by adding a slack term or decreasing the penalty order. Such methods introduce hyper-parameters, e.g. slack weights and tolerant threshold, making the model harder to tune.

Speaker-Distribution Loss To overcome the potential problems of square loss, we propose the *Speaker-Distribution Loss*. To help us with notation, we define an utility function $S(i)$ which returns all indices of the sequences from the same speaker (or dyad).

If speaker information is not available, then this function will return all sequence indices from the training set (in our experiments, we call this loss *Global-Distribution Loss*). The core component of our *Speaker-Distribution Loss* is the $D_{expected}$ function which compares the distribution of the model predictions with the distribution of the ground truth labels for the same speaker (or dyad). Formally, we define $D_{expected}$ function as:

$$D_{expected} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \in S(i)} \delta(\hat{y}_i, y_j). \tag{11}$$

Then, our speaker-distribution loss is defined as the square loss divided by our $D_{expected}$ function:

$$\mathcal{L}_{SDL} = \frac{\frac{1}{n}\sum_i \delta(\hat{y}_i, y_i)}{\frac{1}{n^2}\sum_i \sum_{j \in S(i)} \delta(\hat{y}_i, y_j)} = \frac{n\sum_i \delta(\hat{y}_i, y_i)}{\sum_i \sum_{j \in S(i)} \delta(\hat{y}_i, y_j)} \tag{12}$$

The numerator part of our speaker-distribution loss function will minimize the squared distance between predicted and ground truth labels ($\hat{y}_i$ and y_i), where these distances are computed independently of other training sequences. The denominator part of our speaker-distribution loss function will enforce the distributions of predicted and ground truth labels to be closer. This enforcement is performed by grouping sequences per speaker (or per dyads).

Our TSAM approach is not constrained to regression problems. For classification tasks, $\mathcal{L}_{SDL}$ can be easily redefined by changing the discrepancy function to cross-entropy error.

4 EXPERIMENTAL SETUP

To show the effectiveness of our model, we experiment on affect state and social state datasets with unimodal and multimodal settings. In the experiment, our model is evaluated on three different tasks: interpersonal rapport detection, multimodal sentiment analysis, and text sentiment analysis. While the main focus of our approach is regression tasks given our speaker-distribution loss function, we also show generalization when using our attention and encoding modules for classification. In general, We expect to study the following research questions:

(1) How well does our model generalize to different tasks, from regression (rapport detection and multimodal sentiment analysis) to classification (multimodal and text sentiment analysis)?

(2) How does our model perform in multimodal and unimodal settings?

(3) Are attention weights able to select the task-relevant time-steps?

(4) In which cases does the speaker-distribution loss improve over the square loss?

4.1 Datasets

4.1.1 Rapport Dataset. The "Rapport in Peer Tutoring" dataset (RPT or Rapport dataset for short) was collected to understand the dynamics of rapport formation and the impact of rapport on peer tutoring and learning. RPT is comprised of audio and video data from 14 dyads of students in two hour-long peer tutoring sessions, for a total of 28 hours of data. We followed a similar experimental setup as the "Rapport 2013" [56] dataset. However, unlike "Rapport 2013", the students worked together via a live video chat software, and they were all dyads of strangers prior to the first session, unlike the friends and strangers in [56]. Half of the dyads were pairs of boys and half were pairs of girls, with a mean age of 13.5. The RPT corpus was segmented into 30-second "thin-slices" (3,363 in total), which were given to naive observers on Amazon Mechanical Turk to rate the rapport for the dyad in each slice. Each Turker was shown a definition of rapport and asked to rate the rapport in 10 randomly selected video slices on a 7-point Likert scale, with 1 being very low rapport, and a 7 being very high rapport. Each slice was rated by 3 Turkers, with an average Krippendorff's alpha across all Turkers' slices of 0.61, and the average rating used as the final measure of rapport.

4.1.2 MOSI. The Multimodal Opinion Sentiment Intensity (MOSI) dataset [53] is proposed as a benchmark for multimodal sentiment intensity analysis. This dataset is collected from YouTube movie reviews and it contains 2,199 video segments from 89 distinctive speakers. Sentiment intensity is defined from strongly negative to strongly positive with a linear scale from −3 to 3. The sentiment intensity of each video segment is annotated by five online workers from Amazon Mechanical Turk website and the final rating is the average of all 5 workers. Three different modalities: audio, video, and text, are provided.

4.1.3 IMDB Movie Review. To evaluate our model on text, we leverage the IMDB dataset [23], which is a benchmark for sentiment analysis. This corpus contains 50,000 movie reviews taken from IMDB, each comprised of several sentences. 25,000 instances are labeled as training data and 25,000 instances are labeled as test data. There are two types of labels (positive and negative), and they are balanced in both the training and test set.

4.2 Comparison Methods

To answer the research questions, we compare the following methods in our experiments:

TAGM [31]: TAGM (Temporal Attention-Gated Model) is the latest attention model for sequence classification. It is specifically designed for salience detection. Different from our work, TAGM developed the recurrent attention-gated units to accumulate the summative hidden states and learn the sequence representation as the last time-step.

DAN [16]: DAN (Deep Averaging Network) is a deep neural network that models a sequence by averaging the embeddings associated with an input sequence. DAN is a simplified model that weights each time-step equally. By comparing our model with DAN, we can study the necessity of our attention mechanism.

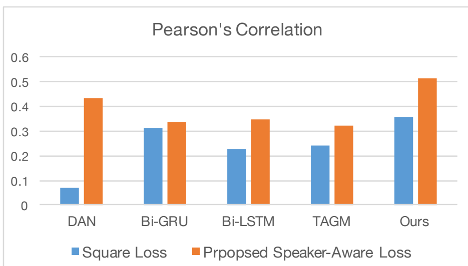
Bi-LSTM and **Bi-GRU**: LSTM [14] and GRU [4] are now popular techniques for sequence modelling. In the experiments, we will test the bi-directional LSTM and GRU. The number of hidden states are set to be same as our model.

The state-of-the-art methods: We compare our model to the state-of-the-art results for each dataset.

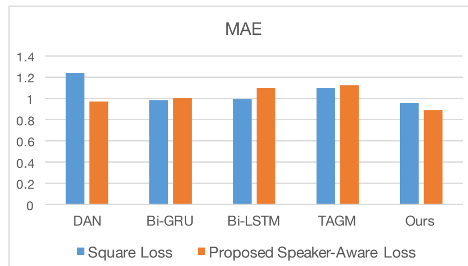
- **SAL-CNN** [50]: SAL (Select-Additive Learning) is designed f or improving the generalizability of multimodal sentiment analysis. SAL-CNN specifically addresses the confounding factor problem for convolutional neural networks.
- **PVEC** [21] and **SA-LSTM** [5]: PVEC (Paragraph Vectors) and SA-LSTM (Sequence Autoencoder initialized LSTM) are able to learn the text representation with unlabeled

		Audio		Video		A + V	
		MAE	Pearson	MAE	Pearson	MAE	Pearson
Baselines	DAN	1.006	0.413	1.236	0.204	0.979	0.431
	Bi-GRU	1.057	0.304	1.130	0.224	1.006	0.337
	Bi-LSTM	1.103	0.346	1.282	0.180	1.101	0.347
	TAGM [31]	1.331	0.297	1.251	0.203	1.124	0.323
Our Models	TSAM w/o Att	1.189	0.450	1.466	0.183	1.178	0.466
	SL-TSAM	0.967	0.351	1.029	0.065	0.968	0.355
	GDL-TSAM	0.956	0.483	1.092	0.175	0.936	0.486
	TSAM	**0.937**	**0.489**	**1.005**	**0.336**	**0.894**	**0.512**
	Human	MAE: 1.183					

Table 1: The regression performance on Rapport dataset. All models are trained with the speaker-distribution loss. Pearson's Correlation (higher is better) and MAE (lower is better) are the evaluation metrics. The rapport scores are between 1 and 7.



(a) Comparing speaker-distribution loss and square loss on Pearson's correlation.



(b) Comparing speaker-distribution loss and square loss on Mean Absolute Error.

Figure 3: Comparing effect of square loss and proposed speaker-distribution loss. Figure (a) and (b) shows the model performances under Pearson's correlation and MAE respectively. We report the fused results (A + V) for all methods.

corpus. Both methods achieve strong performance for text classification benefitting from the pretraining strategy on external data.

Our proposed model with its variants:

- **TSAM**: Our proposed Temporally Selective Attention Model with speaker-distribution loss as described in Section 3.
- **SL-TSAM**: Our proposed model with the square loss.
- **GDL-TSAM**: Our proposed model with the global-distribution loss as described in Section 3.3.
- **TSAM w/o Attention**: To test the effect of attention weights in our model, we remove the attention module. The weight of every frame is a constant $\frac{1}{T}$. The sequence representation of Equation (8) is substituted as $\mathbf{s} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{h}_t$.

4.3 Feature Extraction

For visual features, we use OpenFace [2] to extract facial appearance features. We perform a similarity transform from the currently facial landmarks to a representation of frontal landmarks from a neutral expression. We then extract Histograms of Oriented Gradients (HOGs) features [11] from the aligned face. This leads to a 4,464 dimensional vector describing the face. In order to reduce the feature dimensionality, we use PCA to keep 95% of explained variability, leading a basis of 1,391 dimensions.

For acoustic features, we utilize COVAREP [7] (version 1.4.1) to extract commonly used speech features, such as Mel-Frequency Cepstral Coefficients (MFCCs) and prosodic/voice quality features.

For text, the words are represented as the pretrained *Glove* [32] word embedding.

4.4 Evaluation Metrics

We evaluate the regression tasks with *Mean Absolute Error (MAE)* and *Pearson's Correlation* and evaluate the classification tasks with *accuracy*.

4.5 Training

Our model is trained in an end-to-end fashion with *Adam* [19] as the optimizer. When minimizing Equation (12), we use minibatch scheme to approximate the expected discrepancy of Equation (11), i.e. summation over predictions and ground truth values within the minibatch.

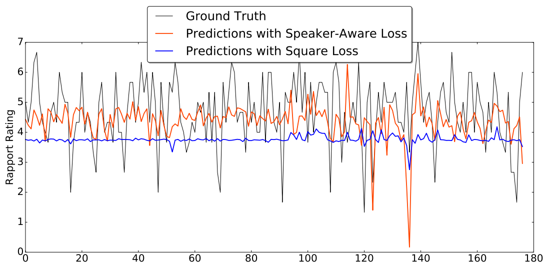
5 RESULTS AND DISCUSSION

5.1 Rapport Detection

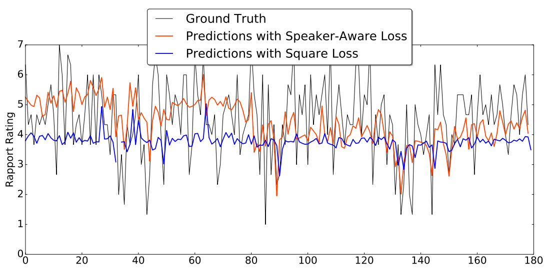
We perform a speaker-independent leave-one-dyad-out cross testing following the leave-one-dyad-out cross validation. The system

	Audio		Video		Text		A + V + T	
	MAE	Pearson	MAE	Pearson	MAE	Pearson	MAE	Pearson
DAN	1.362	0.133	**1.342**	0.146	1.288	0.434	1.265	0.446
Bi-GRU	1.367	0.108	1.384	0.106	1.212	0.474	1.208	0.485
Bi-LSTM	1.374	0.131	1.376	0.132	1.157	0.499	1.124	0.512
TAGM [31]	**1.348**	**0.169**	1.390	0.051	1.068	0.590	1.035	0.587
TSAM w/o Att	1.373	0.127	1.363	0.170	1.076	0.578	1.014	0.596
TSAM	1.366	0.136	1.360	**0.172**	**0.986**	**0.630**	**0.967**	**0.641**

Table 2: Multimodal sentiment regression results on MOSI dataset.



(a) Dyad 1.



(b) Dyad 2.

Figure 4: Illustration of the predictions of square loss and proposed speaker-distribution loss. Figures show the prediction curves of speaker-distribution loss and square loss from two dyads, as well as the ground truth. The datapoints on the curves represent predicted/ground truth values of a 30-second slice. All slices are predicted independently, and the curves are plotted by connecting the predictions/ground truths in chronological order.

outputs the testing predictions when it reaches the best performance on the validation dyad. In this way, the splits of the dataset is disjoint with respect to speakers.

For each method, we report a simple fusion model that fuses the results from two modalities with linear combination: $P_{fused} = \alpha P_A + (1 - \alpha) P_V$, where P_i is the prediction value of modality i, and α is the learnable coefficient.

5.1.1 Performance of Regression. Table 1 presents the regression performance on the Rapport dataset. In general, our model achieves the best results in both unimodal and multimodal settings. In comparison with TSAM w/o Att, we can see that the attention mechanism improves the model performance. Although training

	Audio	Video	Text	A + V + T
DAN	0.572	0.628	0.650	0.692
Bi-GRU	0.591	0.586	0.710	0.714
Bi-LSTM	0.584	0.578	0.668	0.691
TAGM [31]	0.579	0.597	0.713	0.716
SAL-CNN [50]	**0.618**	**0.636**	0.732	0.730
TSAM w/o Att	0.588	0.615	0.723	0.744
TSAM	0.609	0.618	**0.745**	**0.751**

Table 3: Multimodal sentiment classification results on MOSI dataset.

TSAM with square loss has competitive MAE scores, the Pearson's Correlation drops dramatically. Furthermore, the superiority of TSAM over GDL-TSAM verifies the importance of natural grouping of speaker-aware social states.

All models achieve better results with acoustic features than with visual features. Also, the combination results under the multimodal setting outperform the unimodal results. It is interesting to note that the MAE of human raters is larger than most models. It might be because that raters have their own preferences of rating scales.

5.1.2 Effect of Speaker-Distribution Loss. We study the effect of speaker-distribution loss proposed in Section 3.3. The speaker-distribution loss and the square loss are compared across different models in Figure 3. All models trained with the speaker-distribution loss consistently outperform the square loss under Pearson's correlation. Most models gain huge improvements of more than 10%. On the other side, using speaker-distribution loss does not significantly affect the performance in terms of MAE. Although the objective of square loss is to minimize the pairwise discrepancy (which corresponds with the MAE metric), the speaker-distribution loss is capable of achieving similar effects.

To investigate the prediction properties of speaker-distribution loss, we illustrate the prediction curves of our model when trained with two losses in Figure 4. Each datapoint on the curve represents the prediction value of a 30-second slice. The ground truth curve is also plotted as reference. Not only does the speaker-distribution loss curve fit the distribution of ground truth better, it also produces more distinguishable outputs to avoid conservation predictions (always average rating). Dyad 1 (Figure 4(a)) especially reflects this observation.

5.1.3 Visualization of Attended Frames. In this experiment, we visualize the task-relevant frames the attention module captures.

Figure 5: Visualization of attention weights of an 30-second slice in Rapport dataset. The arrow shows the change of attention weights with respect to time. Red denotes high weights while green denotes low weights. Representative frames from different parts are illustrated.

	DAN	Bi-GRU	Bi-LSTM	TAGM	PVEC [21]	SA-LSTM [5]	TSAM w/o Att	TSAM
Accuracy	0.894	0.890	0.887	0.905	0.926	**0.928**	0.898	0.917

Table 4: The binary classification performance on the IMDB dataset. The state-of-the-art methods, SA-LSTM [5] and PVEC [21] utilize the unlabeled corpus to train the text representation. Other methods are trained with the labeled reviews.

Figure 5 presents an example slice with the learned attention weights of time-steps. We use the color variations to indicate the magnitudes of attention weights. Compared with the unattended frames (green), the attended part (red) corresponds to the frames showing good interactions of two speakers which are signs for high rapport.

5.2 Multimodal Sentiment Analysis

We conduct both binary classification and regression experiments for multimodal sentiment analysis on MOSI dataset.

5.2.1 Binary Classification and Regression. The results of binary classification and regression are presented in Table 3 and 2. These tables show that TSAM achieves the best performance among all models with multimodal fusion and text modality. Consistent with the results in rapport detection, our full model TSAM consistently outperforms the one without attention.

5.3 Text Sentiment Analysis

We perform a binary classification task on IMDB movie review dataset. In this experiment, we show generalization using our attention and encoding modules for written language.

5.3.1 Results on Binary Classification. The experimental results on IMDB dataset are reported in Table 4. In this experiment, we include the state-of-the-art methods, SA-LSTM [5] and PVEC [21], which utilize external review documents to pre-train the text representation. Although less data are used for training, our model achieves comparable results. Moreover, with the same amount of resources (using the labeled corpus only), our model outperforms the other baselines.

5.3.2 Visualization of Attended Words. We illustrate the top-20 most attended words in Figure 6. As the sentence length T varies in the corpus, the attention weight of a word is normalized by dividing

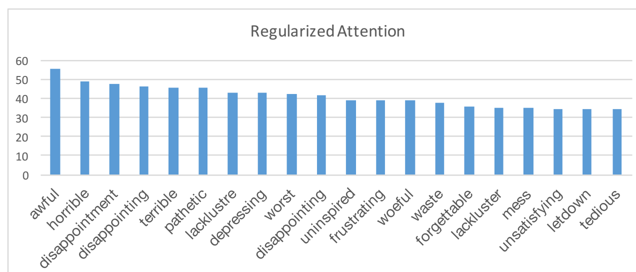


Figure 6: The top-20 attended words. We calculate the mean regularized attention weights of the words. All attended words are sentiment words.

the expected weight $\frac{1}{T}$. Then we can measure the importance of a word in sentiment analysis by calculating the mean normalized attention over all time-steps where it appears. We can see that all the top ranked attended words are sentiment words that express the individual attitude.

6 CONCLUSIONS

In this paper, we propose an temporally selective attention model for social and affective state recognition. The attention mechanism combined with the encoding module enables our model to attend to salient parts of human-centric video sequences. Taking the advantage of natural grouping of speaker-aware labels, we develop a speaker-distribution loss for model training. In the experiments, our model achieves the state-of-the-art performance on different tasks with both unimodal and multimodal settings.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. *ICLR* (2015).

[2] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. 2016. Openface: an open source facial behavior analysis toolkit. In *WACV*.

[3] Chen Chen, Zuxuan Wu, and Yu-Gang Jiang. 2016. Emotion in Context: Deep Semantic Feature Fusion for Video Emotion Recognition. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 127–131.

[4] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[5] Andrew M Dai and Quoc V Le. 2015. Semi-supervised sequence learning. In *Advances in Neural Information Processing Systems*. 3079–3087.

[6] Ionut Damian, Tobias Baur, and Elisabeth André. 2016. Measuring the impact of multimodal behavioural feedback loops on social interactions. In *ICMI*.

[7] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP - A collaborative voice analysis repository for speech technologies. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 960–964.

[8] Zhi-Hong Deng, Hongliang Yu, and Yunlun Yang. 2016. Identifying Sentiment Words Using an Optimization Model with L1 Regularization. In *AAAI*.

[9] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive Recursive Neural Network for Target-dependent Twitter Sentiment Classification. In *ACL*.

[10] Chris Dyer, Miguel Ballesteros, Wang Ling, Austin Matthews, and Noah A Smith. 2015. Transition-based dependency parsing with stack long short-term memory. *ACL* (2015).

[11] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. 2010. Object detection with discriminatively trained part-based models. *TPAMI* 32, 9 (2010), 1627–1645.

[12] Sayan Ghosh, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2016. Representation Learning for Speech Emotion Recognition. *Interspeech* (2016).

[13] Alex Graves, Navdeep Jaitly, and Abdel-rahman Mohamed. 2013. Hybrid speech recognition with deep bidirectional LSTM. In *ASRU*.

[14] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[15] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *SIGKDD*.

[16] Mohit Iyyer, Varun Manjunatha, Jordan Boyd-Graber, and Hal Daumé III. 2015. Deep Unordered Composition Rivals Syntactic Methods for Text Classification. In *ACL*.

[17] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, and others. 2016. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces* 10 (2016), 99–111.

[18] Yelin Kim and Emily Mower Provost. 2016. Emotion spotting: Discovering regions of evidence in audio-visual emotion expressions. In *ICMI*.

[19] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *ICLR* (2015).

[20] Alexandra König, Aharon Satt, Alexander Sorin, Ron Hoory, Orith Toledo-Ronen, Alexandre Derreumaux, Valeria Manera, Frans Verhey, Pauline Aalten, Phillipe H Robert, and others. 2015. Automatic speech analysis for the assessment of patients with predementia and Alzheimer's disease. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* 1 (2015), 112–124.

[21] Quoc V Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents.. In *ICML*, Vol. 14. 1188–1196.

[22] Christine L Lisetti and Fatma Nasoz. 2002. MAUI: a multimodal affective user interface. In *MM*.

[23] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 142–150.

[24] Alberto Montes, Amaia Salvador, and Xavier Giro-i Nieto. 2016. Temporal Activity Detection in Untrimmed Videos with Recurrent Neural Networks. *NIPS Workshop* (2016).

[25] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. 2011. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *ICMI*.

[26] Catherine Neubauer, Joshua Woolley, Peter Khooshabeh, and Stefan Scherer. 2016. Getting to know you: a multimodal investigation of team behavior and resilience to stress. In *ICMI*.

[27] Thien Hai Nguyen and Kiyoaki Shirai. 2015. PhraseRNN: Phrase Recursive Neural Network for Aspect-based Sentiment Analysis. In *EMNLP*.

[28] Behnaz Nojavanasghari, Tadas Baltrušaitis, Charles E Hughes, and Louis-Philippe Morency. 2016. EmoReact: a multimodal approach and dataset for recognizing emotional responses in children. In *ICMI*.

[29] Lauri Oksama and Jukka Hyönä. 2008. Dynamic binding of identity and location information: A serial model of multiple identity tracking. *Cognitive psychology* 56 (2008), 237–283.

[30] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *EMNLP*.

[31] Wenjie Pei, Tadas Baltrušaitis, David MJ Tax, and Louis-Philippe Morency. 2017. Temporal Attention-Gated Model for Robust Sequence Classification. (2017).

[32] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global Vectors for Word Representation.. In *EMNLP*, Vol. 14. 1532–1543.

[33] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-Level Multimodal Sentiment Analysis. In *ACL*.

[34] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. 2017. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion* 37 (2017), 98–125.

[35] Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain. 2016. Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing* 174 (2016), 50–59.

[36] Soujanya Poria, Iti Chaturvedi, Erik Cambria, and Amir Hussain. 2016. Convolutional MKL based multimodal emotion recognition and sentiment analysis. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*. IEEE, 439–448.

[37] Jia Rong, Gang Li, and Yi-Ping Phoebe Chen. 2009. Acoustic feature selection for automatic emotion recognition from speech. *Information processing & management* 45, 3 (2009), 315–328.

[38] Alexander M Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. *EMNLP* (2015).

[39] Lisa D Sanders and Lori B Astheimer. 2008. Temporally selective attention modulates early perceptual processing: Event-related potential evidence. *Attention, Perception, & Psychophysics* 70, 4 (2008), 732–742.

[40] Klaus R Scherer. 2005. What are emotions? And how can they be measured? *Social science information* 44, 4 (2005), 695–729.

[41] Alex J Smola and Bernhard Schölkopf. 2004. A tutorial on support vector regression. *Statistics and computing* 14 (2004), 199–222.

[42] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, and others. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *EMNLP*.

[43] Bo Sun, Qinglan Wei, Liandong Li, Qihua Xu, Jun He, and Lejun Yu. 2016. LSTM for dynamic emotion and group emotion recognition in the wild. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 451–457.

[44] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

[45] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics* 37 (2011), 267–307.

[46] Duyu Tang, Bing Qin, and Ting Liu. 2016. Aspect level sentiment classification with deep memory network. In *EMNLP*.

[47] Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification. In *ACL*.

[48] Imen Trabelsi, D Ben Ayed, and Noureddine Ellouze. 2016. Comparison Between GMM-SVM Sequence Kernel And GMM: Application To Speech Emotion Recognition. *Journal of Engineering Science and Technology* 11, 9 (2016), 1221–1233.

[49] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Björn Schuller, and Stefanos Zafeiriou. 2016. Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 5200–5204.

[50] Haohan Wang, Aaksha Meghawat, Louis-Philippe Morency, and Eric P Xing. 2016. Select-Additive Learning: Improving Cross-individual Generalization in Multimodal Sentiment Analysis. *arXiv preprint arXiv:1609.05244* (2016).

[51] Martin Wöllmer, Felix Weninger, Tobias Knaup, Björn Schuller, Congkai Sun, Kenji Sagae, and Louis-Philippe Morency. 2013. Youtube movie reviews: Sentiment analysis in an audio-visual context. *Intelligent Systems* 28 (2013), 46–53.

[52] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *CVPR*.

[53] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos. *arXiv preprint arXiv:1606.06259* (2016).

[54] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *ntelligent Systems* 31 (2016), 82–88.

[55] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. 2009. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *TPAMI* 31 (2009), 39–58.

[56] Ran Zhao, Alexandros Papangelis, and Justine Cassell. 2014. Towards a dyadic computational model of rapport management for human-virtual agent interaction. In *International Conference on Intelligent Virtual Agents*. Springer, 514–527.

[57] Ran Zhao, Tanmay Sinha, Alan W Black, and Justine Cassell. 2016. Socially-aware virtual agents: Automatically assessing dyadic rapport from temporal patterns of behavior. In *IVA*.

[58] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, Wenlong Xie, Xiaolei Jiang, and Tat-Seng Chua. 2016. Predicting personalized emotion perceptions of social images. In *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 1385–1394.