Transformer-Based Physiological Feature Learning for Multimodal Analysis of Self-Reported Sentiment

Shun Katada
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
s2040005@jaist.ac.jp

Shogo Okada
Japan Advanced Institute of Science and Technology
Nomi, Ishikawa, Japan
okada-s@jaist.ac.jp

Kazunori Komatani
The Institute of Scientific and Industrial Research (SANKEN), Osaka University
Ibaraki, Osaka, Japan
komatani@sanken.osaka-u.ac.jp

ABSTRACT

One of the main challenges in realizing dialog systems is adapting to a user's sentiment state in real time. Large-scale language models, such as BERT, have achieved excellent performance in sentiment estimation; however, the use of only linguistic information from user utterances in sentiment estimation still has limitations. In fact, self-reported sentiment is not necessarily expressed by user utterances. To mitigate the issue that the true sentiment state is not expressed as observable signals, psychophysiology and affective computing studies have focused on physiological signals that capture involuntary changes related to emotions. We address this problem by efficiently introducing time-series physiological signals into a state-of-the-art language model to develop an adaptive dialog system. Compared with linguistic models based on BERT representations, physiological long short-term memory (LSTM) models based on our proposed physiological signal processing method have competitive performance. Moreover, we extend our physiological signal processing method to the Transformer language model and propose the Time-series Physiological Transformer (TPTr), which captures sentiment changes based on both linguistic and physiological information. In ensemble models, our proposed methods significantly outperform the previous best result ($p < 0.05$).

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**.

KEYWORDS

Time-Series Processing, Physiological Signals, Sentiment Analysis

ACM Reference Format:
Shun Katada, Shogo Okada, and Kazunori Komatani. 2022. Transformer-Based Physiological Feature Learning for Multimodal Analysis of Self-Reported Sentiment. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3536221.3556576

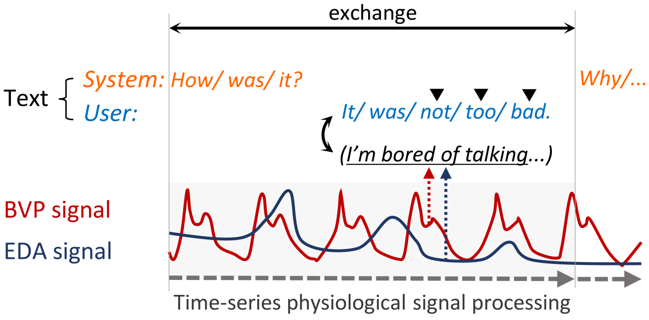


Figure 1: Example of capturing self-sentiment changes by using linguistic information and physiological signals at the exchange level. The user token sequences "not/too/bad" include both neutral and positive sentiments. However, the true self-sentiment in his or her mind is "bored" (tentative example). This masked negative sentiment is accompanied by reduced arousal levels and would be captured by time-series physiological signals.

1 INTRODUCTION

The development of an adaptive dialog system that can recognize a user's state in real time is necessary to ensure enjoyable conversations in human-agent interactions. During a chat dialog, the system should behave according to the real-time state of the user. For example, if a user is bored with the current topic, the system should explore other topics, like human behavior. However, there are several reasons why this task is challenging. For example, self-reported sentiment (hereafter referred to as self-sentiment) cannot necessarily be expressed with the linguistic information obtained from user utterances. Users may mask their self-sentiment in their mind and not express their true sentiment as an utterance or behavior due to their emotional intelligence [18].

Peripheral physiological signals have been investigated in psychophysiology and affective computing [20]. These signals can potentially reflect emotional changes by capturing physiological changes in the autonomic nervous system (ANS). For example, a faster phasic component in the electrodermal activity (EDA), which is derived from the activity of the sweat glands, can be used to detect emotional arousal [5, 13]. Since the ANS is involuntary, i.e., it cannot be controlled consciously, physiological changes during dialog are difficult to mask. Therefore, physiological signals may be suitable for capturing self-sentiment changes that cannot be

ICMI '22, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9390-4/22/11…$15.00
https://doi.org/10.1145/3536221.3556576

represented by linguistic information in user utterances and can function as complementary information.

However, investigations into the effectiveness of time-series physiological signals for estimating self-sentiment during dialog exchanges have been limited. Most studies on the use of physiological signals to estimate emotion/sentiment have induced emotional stress with visual stimuli over a relatively long time period (several minutes). Thus, there is a need to investigate whether signals detected in shorter time periods (approximately 10 seconds) are effective for online emotion/sentiment estimation.

In addition, although it is assumed that short-time physiological signals can complement spoken linguistic information in self-sentiment detection, there have been no studies that show an effective method for combining time-series physiological signals with token sequences represented by state-of-the-art (SOTA) language models, such as the Bidirectional Encoder Representations from Transformers (BERT) model [6]. Thus, the exploration of effective methods for fusing physiological signals and token representations is valuable for developing adaptive dialog systems.

The aforementioned issues and the approach presented in this study are summarized in Fig. 1. An "exchange" is defined as a segment that begins at the start of a system utterance and ends at the start of the next system utterance. In this case, a model based solely on user token sequences is insufficient for estimating self-sentiment. We expect that time-series physiological signals could be used to capture self-sentiment changes that are not expressed in linguistic information, and a time-series model that combines physiological signals and language representations can improve the sentiment estimation performance, as these data are complementary.

In this study, we propose an effective method for processing physiological signals and combine this method with a language model. We focus on linguistic information and physiological signals since the models based on BERT representations or physiological signal had dominant performance compared to audiovisual models (described in Section 5.1). The contributions of our work are as follows:

- We propose a time-series physiological signal processing method for exchange-level sentiment estimation. The models based on the time-series data of the EDA phasic component capture short-time sentiment changes during exchanges, showing competitive performance to a linguistic model based on SOTA computational representations, i.e., BERT representations (Section 5.1).
- We introduce the Time-series Physiological Transformer (TPTr), which combines time-series physiological signals with BERT representations to capture short-time sentiment changes based on both textual aspects and physiological changes in the user (Section 5.2). As a result, our proposed ensemble model outperforms the previously reported best result.
- Our proposed model is extended and validated by using a variety of physiological signals, including the blood volume pulse (BVP). The performance is further improved with the ensemble method, as shown in Section 5.3.

2 RELATED WORKS

This section specifically focuses on research related to the Transformer language model and multimodal models.

Text-based approaches are critical in sentiment analysis, and neural network models such as LSTM are widely used [30]. However, the Transformer model, which was developed by [28], has become the de facto standard and most used language model. The best Transformer-based model is BERT [6], which achieved numerous successes with sentiment estimation tasks with datasets such as the Stanford Sentiment Treebank (SST-2) [25]. When BERT is pretrained with a large-scale dataset, representations can be extracted from text data (referred to as BERT representations), and BERT representations can be used as input feature vectors in other architectures. This method allows BERT representations to be easily combined with audiovisual features and is often used in multimodal sentiment analysis.

Although several Transformer-based multimodal models for affective computing and sentiment analysis have recently been proposed [4, 8, 22], a Multimodal transformer called MulT was the first model proposed in multimodal sentiment analysis research [27]. Language, video and audio modalities, as well as sentiment labels annotated by third parties, were used to demonstrate the effectiveness of the proposed crossmodal attention model, which latently adapts streams from one modality to another. Although physiological signals were not included in these studies, it has been suggested that Transformer-based models could capture cross-modal attention between text and audiovisual signals. Multimodal Adaptation Gate (MAG) was introduced in [22] and is applied to the Transformer architecture of BERT/XLNet. The MAG allows to shift the language-only position (representation) of the word to the new position by injecting audio-visual information. The core component of the MAG is a non-verbal displacement vector derived from the audio and visual vectors with their respective gating vectors. [8] proposed modality-invariant and -specific representations, which project language, audio and visual modalities to two distinct subspaces. The respective representations are stacked into a matrix, and Transformer is used to perform a multi-headed self-attention on the matrix.

Compared to linguistic and audiovisual modalities, there are very few publicly available physiological signal datasets for emotion/sentiment research. However, several datasets that include physiological signals have been created while viewing emotional videos [14, 19, 26] or conversations [15, 23]. In [15], a multimodal human-agent dialog corpus that included linguistic, audiovisual, and physiological information was created. The participants interacted with an agent, and sentiment labels were retrospectively annotated for each exchange by both the participants and a third party. The collected nonverbal signals (audio, visual, and physiological signals) in this dataset were used for sentiment estimation with support vector machine or feedforward deep neural network models, and the results showed that physiological signals, particularly features based on SC signals, were useful for exchange-level sentiment estimation, as reported by [11]. Our previous study was extended by [12], which reported a comprehensive analysis of the effectiveness of physiological signals in multimodal sentiment analysis. Since our proposal in this paper is an effective hybrid algorithm that combines physiological features and the Transformer language model, our study differs considerably from these previous studies, which used conventional neural networks [11, 12].

To the best of our knowledge, there is no publicly available dataset that includes textual and physiological information during dialog exchanges, except for [15]. As mentioned above, text-based approaches are the most common sentiment analysis methods, and multimodal language models using Transformer and BERT have been proposed. Physiological signals are promising candidates for capturing subtle sentiment changes that cannot be detected in the speaker's explicit information, i.e., text and audiovisual information. Nevertheless, an effective method that combines a SOTA language model and physiological signals has not yet been developed, most likely because of dataset limitations.

We propose the use of physiological signals with a SOTA language model to estimate sentiment during human-agent interactions. The Hazumi1911 dataset [15], which is the only publicly available dataset that includes time-series textual and physiological information, enables us to evaluate the effectiveness of the combination of physiological signals and text. We propose a time-series physiological signal processing method that effectively combines physiological signals and token sequences of utterances. We show that our proposed method is useful for exchange-level sentiment estimation, and our results are comparable to those of a model based on BERT representations. Then, we show how the time-series physiological signals can be incorporated into a SOTA language model, and proposed model were compared with the previously reported best performing model.

3 PROPOSED METHODS

This section presents our proposed methods for incorporating time-series physiological signals at the exchange level. In this study, the physiological signals included the EDA, BVP, heart rate (HR) and skin temperature (TEMP). The EDA is a measure of the electrical activity in human skin and reflects sweat gland activity. The BVP is based on spectral analyses of the skin (blood vessels) and reflects physiological changes in cardiovascular activity. In this study, the raw EDA signal (skin conductance (SC), denoted as EDA_{SC}) was decomposed into a fast phasic component (EDA_{fast}) and a tonic component ($\text{EDA}_{\text{tonic}}$) with the same method as in [11]. In Section 3.1, we describe a physiological signal processing method for calculating fine-segmented physiological changes. Since each physiological signal has a different sampling rate, a simple segmentation and averaging method was applied. In Section 3.2, to evaluate the effectiveness of the processed data, time-series machine learning models are introduced. Specifically, we propose a Time-series Physiological Transformer (TPTr) model in which the encoder is based on attention weights from the token representations and corresponding physiological signals. We expect this encoder to capture sentiment changes by using both linguistic and physiological information, as sentiment changes cannot be detected with only linguistic information.

3.1 Time-Series Physiological Signal Processing

To roughly align physiological signals within the exchanges with the token, a unit of language models, we divide each physiological signal during each exchange by the number of tokens. Let one exchange duration be s, the sampling rate of the physiological signal in Hz be h, and the number of tokens in one exchange be n. The number of samples per token m is determined by rounding $\frac{sh}{n}$ down to the nearest integer. Then, from the start of the exchange, the raw sampling data per m are averaged in order (i.e., m is the variable window size). Thus, the physiological signal $\boldsymbol{p}$ in the ith exchange is denoted as an n-dimensional vector:

$$\boldsymbol{p}_i^{\alpha} = (p_{i1}^{\alpha}, \ldots, p_{in}^{\alpha})^T \tag{1}$$

where α indicates the physiological submodality such as EDA_{fast}, $\text{EDA}_{\text{tonic}}$, EDA_{SC}, BVP, HR, TEMP.

We note that our proposed preprocessing method is not the strict word-level alignment method that has been proposed in prior works [7, 29]. In contrast to acoustic signals, physiological signals do not necessarily have a significant co-occurrence property with the uttered words because the physiological changes may relate to words spoken in the past or future. Thus, physiological signals are not simply weighted with a specific token in this study. Rather, the aim is to extract representations from fine segments of physiological signals with token sequences, which could shift the original representations at the *exchange level*. More details and examples of our experiment are shown in Section 5.4.

3.2 Time-Series Modeling of Physiological Signals

(1) Physiological LSTMs: The LSTM and bidirectional LSTM models are applied to validate whether our proposed time-series preprocessing method performs comparably to models based on BERT representations, which have deep bidirectionality [6]. An LSTM [10] model based on physiological signals $\boldsymbol{p}_i$ at time t can be represented as

$$\begin{aligned}
\begin{pmatrix} \boldsymbol{f}_t \\ \boldsymbol{g}_t \\ \boldsymbol{\iota}_t \\ \boldsymbol{o}_t \end{pmatrix} &= \begin{pmatrix} \sigma \\ \tanh \\ \sigma \\ \sigma \end{pmatrix} W \begin{pmatrix} \boldsymbol{p}_t \\ \boldsymbol{h}_{t-1} \end{pmatrix} \\
\boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{g}_t \odot \boldsymbol{\iota}_t \\
\boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t)
\end{aligned} \tag{2}$$

where $\boldsymbol{f}_t$, $\boldsymbol{\iota}_t$ and $\boldsymbol{o}_t$ are the forget, input, and output gates, respectively; σ is the sigmoid function; W is the weighting parameter; $\boldsymbol{c}_t$ is the memory cell; $\boldsymbol{h}_t$ is the hidden state; and $\odot$ is the Hadamard product. Note that the time t corresponds to the number of tokens n, as described in Section 3.1. $\boldsymbol{p}_t$ is denoted as a vector in the above equation: however, this variable corresponds to a scalar when the selected physiological submodality is single.

After the preprocessing methods described in Section 3.1 were carried out, the raw physiological data of each participant were normalized by Z score normalization. In other words, we normalized each feature over all the samples collected from a participant in the training or testing data during preprocessing. Following this, zero padding was performed since the token length of each exchange differs. Then, the result was fed into the input layer of the LSTM model. The final LSTM block outputs $\boldsymbol{h}_t$ are connected to the final output layer in a mode known as many-to-one, and finally, the estimated values are obtained.

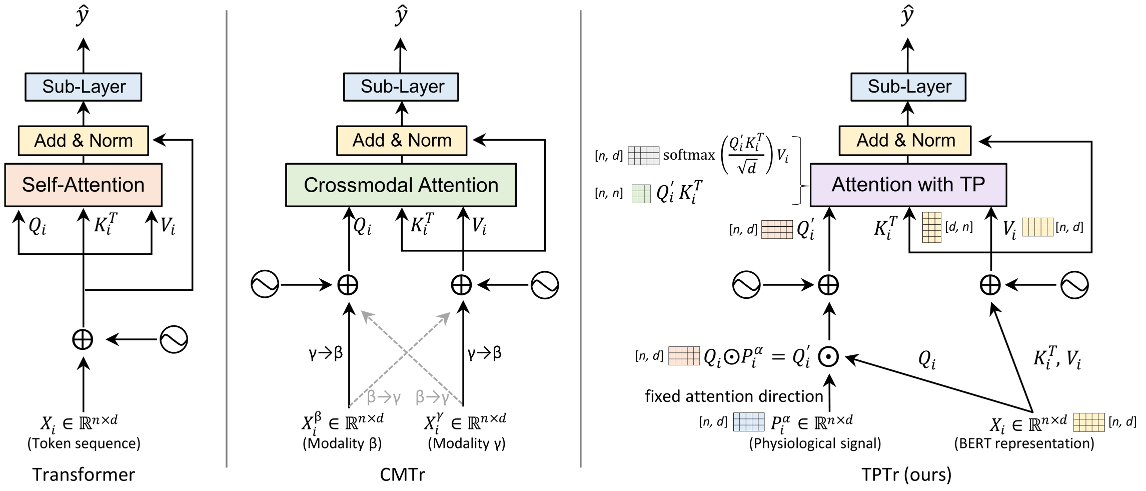

$\hat{y}$
Sub-Layer
Add & Norm
Self-Attention
Q_i K_i^T V_i
$X_i \in \mathbb{R}^{n\times d}$
(Token sequence)
Transformer
Crossmodal Attention
γ→β
β→γ
$X_i^\beta \in \mathbb{R}^{n\times d}$
(Modality β)
$X_i^\gamma \in \mathbb{R}^{n\times d}$
(Modality γ)
CMTr
[n, d] softmax $\left(\frac{Q_i' K_i^T}{\sqrt{d}}\right) V_i$
[n, n] $Q_i' K_i^T$
Attention with TP
[n, d] Q_i'
K_i^T [d, n]
V_i [n, d]
[n, d] $Q_i \odot P_i^\alpha = Q_i'$
fixed attention direction
Q_i
K_i^T, V_i
[n, d] $P_i^\alpha \in \mathbb{R}^{n\times d}$
(Physiological signal)
$X_i \in \mathbb{R}^{n\times d}$ [n, d]
(BERT representation)
TPTr (ours)

Figure 2: Conventional Transformer [28] (left) architecture, CrossModal Transformer [27] architecture based on two modalities, β and γ (CMTr, center), and our proposed Time-series Physiological Transformer (TPTr, right) architecture. In our proposed model (right), exchange-level physiological signals and BERT representations derived from user and system utterances are combined by applying the Transformer encoder. This model allows physiological information to be continuously linked to linguistic information (performing attention with time-series physiological signal processing) and can capture physiological aspects that cannot be detected with linguistic information alone. The number in the bracket indicates the dimension of the corresponding matrix. For a detailed description of the Transformer and CMTr architectures, please see Section 4.1.

(2) Time-Series Physiological Transformer: After the effectiveness of the physiological LSTM models were confirmed (as described in Section 5.1), we extended our proposed method to fuse time-series physiological signals with SOTA language representations, i.e., BERT representations, by using the Transformer encoder [28]. A summary of the proposed TPTr architecture is shown in Fig. 2. Exchange-level BERT representations are extracted with a pretrained BERT model[1], which is represented as a matrix of dimension $\mathbb{R}^{n\times d}$, where d is hidden size of the BERT representations and n is number of tokens. The time-series physiological signal $\boldsymbol{p}_i^\alpha$ was turned into $P_i^\alpha \in \mathbb{R}^{n\times d}$ by repeating array along the axis to match the dimension of the BERT representations. To incorporate the time-series physiological signals into the linguistic information, a dot-product attention mechanism [28] was applied. The dot-product attention mechanism is composed of a query Q, a key K, and a value V. We consider the BERT representation in the ith exchange as $Q_i = K_i = V_i$, where $Q_i \in \mathbb{R}^{n\times d}$. The dot product between Q_i and K_i^T is computed as the similarity to calculate the attention weight. To combine the time-series physiological signals with the BERT representations, we use the Hadamard product between P_i^α and Q_i, denoted as Q_i'. We hypothesize that this modification may shift the attention weight and could provide representations that differ from conventional BERT representations. The output of the dot-product attention operation is:

$$\text{Attention}(P_i^\alpha, Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i' K_i^T}{\sqrt{d}}\right) V_i \tag{3}$$

where the scaling factor $\sqrt{d}$ is used. The PEs are sinusoidal and identical to the modules proposed in [28]:

$$PE_{(pos,2j)} = sin(pos/10000^{2j/d}) \tag{4}$$

$$PE_{(pos,2j+1)} = cos(pos/10000^{2j/d}) \tag{5}$$

where pos is the position of the token and j is the dimension of the hidden layer. The PEs are added to Q_i', K_i^T and V_i to carry information about the position of the tokens. The weighting parameters $W^{Q'} \in \mathbb{R}^{n\times d}, W^K \in \mathbb{R}^{n\times d}$, and $W^V \in \mathbb{R}^{n\times d}$ are also implemented.

Like [28], the Transformer encoder is composed of two sublayers. The first sublayer is the aforementioned dot-product attention mechanism, and the second sublayer is a fully connected feedforward neural network (FNN). Each sublayer has a skipping connection [9] and layer normalization [1], denoted as "Add" and "Norm" in Fig. 2, respectively.

In summary, our proposed preprocessing method converts data, allowing the model to combine physiological signals with BERT representations, which are both represented as matrices during each exchange. These representations are fed into the Transformer model, with the token positions providing the attention weights, thus allowing physiological changes to be considered during exchanges.

4 EXPERIMENTAL SETTINGS

This section describes the experimental settings for the evaluation of our proposed model. One of the strengths of our proposed method is that our method applies short-time episodes (approximately 10 seconds), which enables dialog systems to adaptively respond to sentiment changes in the user in a timely manner. Only one publicly available dataset includes both the time-series physiological signals and linguistic information of the user at the exchange level:

[1] https://github.com/cl-tohoku/bert-japanese

the Hazumi1911 dataset [15]. We use this dataset to evaluate our proposed methods, and Section 4.3 summarizes the dataset. Section 4.1 describes the models used as baselines for comparison, and the evaluation procedure is described in Section 4.2.

4.1 Baselines and hyperparameters

As described in Section 3, the proposed time-series physiological signal processing method was evaluated by using the LSTM, BiLSTM, or TPTr models as inputs. This subsection describes the baseline models that were used for comparisons with our proposed method.
(1) Feedforward Neural Network (FNN): The FNN architecture was used as one of our baselines. The FNN was composed of an input layer, four fully connected layers with dropout in each layer, and an output layer. The FNN has two lower intermediate layers with 64 units and two higher intermediate layers with 32 units. The dropout rate was set to 0.3. The ReLU function was used as the activation function.
(2) Long Short-Term Memory Models (LSTMs): In the LSTM model, the number of LSTM blocks was set to 3, with 64 hidden units (in the BiLSTM model, the number of hidden units was set to 128 in total). No dropout was applied. The activation functions (sigmoid and hyperbolic tangent) are described in Section 3.2.
(3) Transformer (Tr): A conventional Transformer encoder [28] was used as a baseline. This model used only linguistic information (i.e., BERT representations) for sentiment estimation. As shown in Fig. 2 (left), the Transformer encoder was composed of two sublayers. The first sublayer was a self-attention mechanism, and the second sublayer was an FNN. Each sublayer had a skipping connection and layer normalization. The number of Transformer encoder blocks and attention heads is 1. The dimensionality of the input and output is 768, corresponding to the BERT model. The number of units in the pointwise FNN is 128. The dropout rate was set to 0.3. The Tr×3 model has three identical parallelized Transformer blocks. $\mathrm{FNN}_{\mathrm{L+P}}$ (described in Section 4.3) was used to combine with the Transformer models. CrossModal Transformer (CMTr) and our proposed TPTr(×3) are described below.
(4) CrossModal Transformer (CMTr): The CMTr is a core component of the MulT and was proposed in [27]. The MulT model captures multimodal signals according to crossmodal attention and achieves SOTA results in multimodal sentiment estimation. The CMTr model applied crossmodal attention with linguistic, audio, or video modalities, as reported in [27]. The two modalities β and γ, as denoted in Fig. 2 (center), correspond to linguistic, audio, or video modalities. The transfer of information from modality γ to modality β is denoted as "$\gamma \rightarrow \beta$" in Fig. 2 (center). The CMTr model also includes reverse attention, which is denoted as "$\beta \rightarrow \gamma$", in which information is assigned to another Transformer block, allowing modality γ to receive information from modality β. Thus, the attention direction is variable. On the other hand, our proposed TPTr model applies attention with linguistic and physiological modalities and has a fixed attention direction. Therefore, the CMTr and TPTr models use different modalities, and the attention mechanism also differs. For a fair comparison, we fuse BERT representations and physiological signals when using the CMTr architecture in this study. The CMTr model has two Transformer encoder blocks that pass information as $\gamma \rightarrow \beta$ and $\beta \rightarrow \gamma$. The output of each CMTr block was concatenated (64 units in total) and connected to the final output layer. The other parameter settings of the CMTr and Transformer models are identical.
(5) Time-Series Physiological Transformer (TPTr): The TPTr and Transformer models have the same parameter settings. The TPTr×3 model has three extended parallelized Transformer blocks. The output of each TPTr×3 block was concatenated (96 units in total) and connected to the final output layer. Other than these settings, we use the same parameter settings in the Tr(×3), CMTr, and TPTr(×3) models to facilitate a fair comparison.

For late fusion models, each higher intermediate layer in the model is concatenated and connected to the output layer. For ensemble models, the output values of each model were averaged and used as the final estimated value. Late fusion and ensemble methods are both widely used in multimodal machine learning [2]. In consideration of the computational cost, the maximum token length was set as 64 in this study. The other hyperparameters were set as follows: a learning rate of 0.001 with the Adam optimizer and a batch size of 32. The FNN model and models other than the FNN model were trained with 30 and 3 epochs, respectively. Mean squared error was used as a loss function in all experiments. All models were implemented in Keras with TensorFlow backend on NVIDIA GeForce RTX 2060.

4.2 Evaluation Procedure

A leave-one-person-out cross-validation (LOPOCV) method was used in our evaluation. In the LOPOCV method, the samples corresponding to each exchange between a participant and the dialog system were used as the test data, and the remaining samples of the other twenty-five participants were used as the training data. This procedure ensured that the test data of one participant were completely excluded from the training dataset, thereby preventing leakage and overestimation. The mean absolute error (MAE) and Pearson correlation coefficient (Corr) were calculated for each evaluation. The average MAE and Corr values with the LOPOCV method are reported. All experiments were performed three times with random initializations, and the evaluation values were calculated as the average value across the three repetitions. These evaluation values were then compared among the models.

4.3 Dataset

The Hazumi1911 dataset [15], a multimodal human-agent dialog corpus, was used in this study. The data were collected while participants chatted with an agent that operated using the Wizard of Oz method. Data from 26 of the participants and 2468 total exchanges were used in our experiment, and the data are denoted in the same manner as in [11]. The participants annotated the labels for each exchange while watching videos of themselves after the experiment. The labels were assigned as sentiment scores ranging from 1 (no enjoyment of the dialog) to 7 (enjoyment of the dialog) and used in regression tasks.

In the Hazumi1911 dataset, the participants' utterances were manually transcribed into text data. The language representations were extracted by BERT, as described in Section 3. In addition, physiological signals were recorded using an Empatica E4 wristband (Empatica Inc., Cambridge, MA, USA) developed by Empatica Inc.

The E4 device is worn like a wristwatch; it causes neither disturbance nor discomfort during dialog and has been widely used in affective computing research, such as in [17, 21, 31]. Thus, this device is suitable for the evaluation of our proposed methods. The EDA, BVP, HR and TEMP data were recorded at 4, 64, 1 and 4 Hz, respectively. Each time-series physiological signal was preprocessed as described in Section 3.1. Following [11], statistics such as the mean, standard deviation and maximum values of the physiological signals were used for comparisons with baseline models.

Acoustic and visual features were also extracted in the same manner as described in [12]. In brief, the INTERSPEECH 2009 Emotion Challenge feature set (IS09) [24] was extracted from participant's utterances as acoustic features using OpenSMILE software[2]. A total of 384 acoustic features were extracted. Based on the video data, facial landmarks near the eyes, mouth, and eyebrows were identified with the OpenFace library [3], and the velocity and acceleration at each point were calculated to use as facial features. Based on motion data of the hands, shoulders and head recorded with Microsoft Kinect sensors, the velocity and acceleration were calculated to use as motion features. In total, 86 visual features were extracted from the facial expressions and motion activity. These acoustic and visual features were used for model comparisons based on each modality. Models based on each feature are as follows:

(1) FNN$_\text{L}$: FNN model based on BERT representations
(2) FNN$_\text{P}$: FNN model based on EDA$_\text{fast}$ statistics
(3) FNN$_\text{A}$: FNN model based on acoustic features
(4) FNN$_\text{V}$: FNN model based on visual features
(5) FNN$_\text{L+P}$: FNN model based on BERT representations and EDA$_\text{fast}$ statistics
(6) (Bi)LSTM$_\text{P}$: (Bi)LSTM model based on time-series EDA$_\text{fast}$ signals

5 RESULTS AND DISCUSSION

First, we show the effectiveness of the models based on our proposed time-series physiological signal processing method. The physiological LSTM and BiLSTM models perform better than the conventional FNN model based on the statistics. Furthermore, ensembles with linguistic and physiological modalities further improve the estimation performance (Section 5.1). Second, a SOTA language model, namely, the Transformer model, was used to combine the time-series data derived from the physiological and linguistic information. This novel approach captures representations that depend on both token sequences and time-series physiological changes, resulting in further performance improvement with the ensemble model (Section 5.2). Third, to explore other effective time-series physiological signals, the TPTr model based on various physiological signals was evaluated in our proposed framework, and its usefulness was demonstrated (Section 5.3). This analysis reveals that the time-series BVP signal is another useful physiological signal for sentiment estimation. Fourth, to clarify the effect of the physiological signals, the differences in the attention weights between the conventional Transformer and TPTr models was shown (Section 5.4). Finally, a qualitative example of the estimation pattern is shown in Section 5.5 to visualize sequential dynamic sentiment changes and the behavior of each model.

Table 1: Sentiment estimation results of physiological LSTM models based on EDA$_\text{fast}$. The sentiment estimation results of the feedforward deep neural network (FNN) are also shown as a baseline for comparison. For the FNN, the subscript "L" represents models based on BERT representations; "P" represents models based on EDA$_\text{fast}$; "L+P" represents models based on both modalities; and "A" and "V" represent models based on acoustic and visual features, respectively. The experimental results based on the model reported in [12] and the results of our proposed models (ours) are also depicted.

Model		MAE	Corr
Single model	FNN$_\text{L}$ [12]	1.086	**0.254**
	FNN$_\text{P}$ [12]	1.069	0.091
	FNN$_\text{A}$ [12]	1.196	0.145
	FNN$_\text{V}$ [12]	1.166	0.145
	LSTM$_\text{P}$ (ours)	1.067	0.179
	BiLSTM$_\text{P}$ (ours)	1.069	0.176
Late fusion model	FNN$_\text{L+P}$ [12]	1.079	0.178
	FNN$_\text{L+P}$+LSTM$_\text{P}$ (ours)	1.062	0.184
	FNN$_\text{L+P}$+BiLSTM$_\text{P}$ (ours)	1.047	0.191
Ensemble model	FNN$_\text{L+P}$ [12]	1.047	0.238
	FNN$_\text{L+P}$+LSTM$_\text{P}$ (ours)	**1.041**	0.250
	FNN$_\text{L+P}$+BiLSTM$_\text{P}$ (ours)	**1.041**	0.249

5.1 Performance of Physiological LSTM Models

Table 1 shows the regression performance of the unimodal FNN models (FNNs trained with BERT representations, EDA$_\text{fast}$ statistics, acoustic features and visual features are depicted as FNN$_\text{L}$, FNN$_\text{P}$, FNN$_\text{A}$ and FNN$_\text{V}$, respectively) using the model reported in [12] (rows 2 to 5 in Table 1). Our proposed model, that is, the LSTM models trained on time-series physiological signals (LSTM$_\text{P}$ and BiLSTM$_\text{P}$), are shown in rows 6 and 7 in Table 1. In the single model results, our proposed physiological LSTM models have higher Corr values than the conventional FNN$_\text{P}$ (rows 3, 6 and 7 in Table 1). Although the FNN$_\text{L}$ model has the best Corr value of 0.254, the physiological models (FNN$_\text{P}$, LSTM$_\text{P}$ and BiLSTM$_\text{P}$) have lower MAEs than FNN$_\text{L}$ (1.086). The FNNs based on conventional acoustic and visual features (FNN$_\text{A}$ and FNN$_\text{V}$) do not outperform FNN$_\text{L}$, LSTM$_\text{P}$ or BiLSTM$_\text{P}$.

In terms of the MAE, further performance improvement was observed by combining the linguistic and physiological models (late fusion and ensemble models). The ensemble model FNN$_\text{L+P}$+LSTM$_\text{P}$ achieved an MAE of 1.041 and a Corr of 0.250.

These results suggest that our proposed physiological signal processing method is effective for exchange-level sentiment estimation, even if linguistic modalities are not included (LSTM$_\text{P}$ and BiLSTM$_\text{P}$). Compared to the experimental condition, which uses emotional stimuli, the estimation of self-sentiment in natural dialog is a difficult task. Nevertheless, our proposed method achieved competitive performance with an FNN trained on BERT representations (FNN$_\text{L}$). Furthermore, our proposed multimodal models based on linguistic and physiological information efficiently complement each modality. These results indicate that our proposed physiological signal

[2] https://www.audeering.com/opensmile/

Table 2: Sentiment estimation results for the Transformer model and its variant. Tr, Transformer; CMTr, CrossModal Transformer [27]; TPTr, our proposed Time-series Physiological Transformer. "×3" means triplicated Transformer blocks.

Model	Single model		Late fusion with FNN$_{\mathrm{L+P}}$		Ensemble with FNN$_{\mathrm{L+P}}$	
	MAE	Corr	MAE	Corr	MAE	Corr
Tr	1.082	0.227	1.057	0.221	1.042	0.259
Tr×3	1.109	0.219	1.069	0.230	1.053	0.257
CMTr [27]	1.083	0.190	1.138	0.198	1.040	0.254
TPTr (ours)	1.114	0.228	1.099	0.223	1.051	0.261
TPTr×3 (ours)	**1.068**	**0.232**	**1.045**	**0.240**	**1.033**	**0.262**

processing method can potentially capture sentiment changes that cannot be represented by BERT representations alone.

5.2 Performance of TPTr

Table 2 shows the regression performance of the conventional Transformer model, the CMTr model proposed in [27], and our proposed TPTr model. The single models and late fusion models did not outperform the abovementioned ensemble model FNN$_{\mathrm{L+P}}$+LSTM$_{\mathrm{P}}$ (Table 1). However, all the ensemble models showed higher performance than the single models. In particular, ensemble model FNN$_{\mathrm{L+P}}$+TPTr×3 achieved the best results, with an MAE of 1.033 and a Corr of 0.262. In a previous study that used the same dataset and machine learning task as we presented here, it was shown that the ensemble model FNN$_{\mathrm{L+P}}$ achieved a better performance than other multimodal models [12]. We show here that our proposed ensemble model (FNN$_{\mathrm{L+P}}$+TPTr×3) significantly outperforms the previously reported best model (FNN$_{\mathrm{L+P}}$, $p < 0.05$, Wilcoxon signed-rank test), suggesting the effectiveness of our proposed method. In addition, we observed significant performance improvement for the TPTr×3 model compared to the Tr×3 model by further experimental repetitions ($p < 0.05$, Wilcoxon signed-rank test).

These results indicate that incorporation of time-series physiological changes into the Transformer language model, which was achieved with our proposed TPTr model, can capture different representations that cannot be captured by using only FNN$_{\mathrm{L}}$ or FNN$_{\mathrm{P}}$ or the ensemble model FNN$_{\mathrm{L+P}}$. As shown in Section 3, only the dot product of the query and key differs between the conventional Transformer model and our proposed TPTr model, and this difference can affect the TPTr estimation result. The details of the attention weight are analyzed and discussed in Section 5.4.

5.3 TPTr Based on Other Submodalities

We investigated whether the TPTr model based on other physiological signals and its ensembles were effective for sentiment estimation. We evaluate the following models:

(1) Single model: This model is our proposed TPTr×3 model, which was trained on each preprocessed signal from the physiological submodality α, as shown in Section 3. A total of five single models were constructed.

Table 3: Sentiment estimation results of the TPTr model based on physiological submodality to explore other effective submodalities. EDA$_{\mathrm{tonic}}$, tonic component of EDA; EDA$_{\mathrm{SC}}$, skin conductance; BVP, blood volume pulse; HR, heart rate; TEMP, skin temperature.

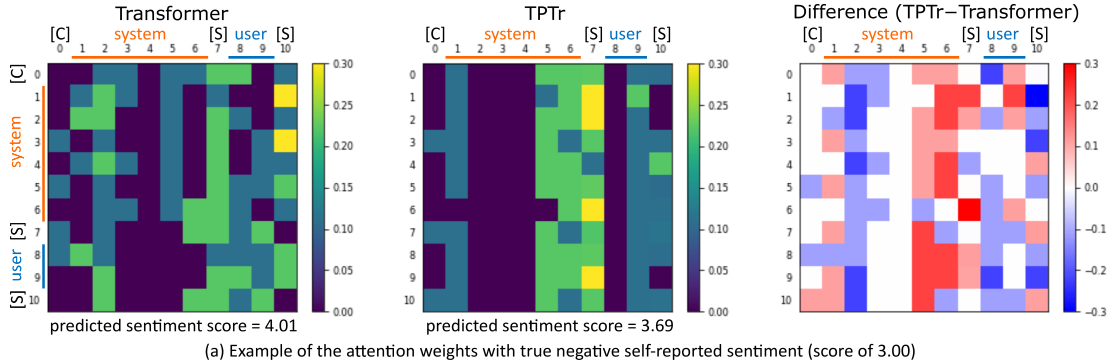
Model	TPTr submodality	MAE	Corr
Single model	EDA$_{\mathrm{tonic}}$	1.113	0.225
	EDA$_{\mathrm{SC}}$	1.115	0.237
	BVP	1.080	0.258
	HR	1.112	0.232
	TEMP	1.100	0.221
Ensemble model (3 models)	EDA$_{\mathrm{tonic}}$	1.052	0.261
	EDA$_{\mathrm{SC}}$	1.052	0.264
	BVP	1.041	0.269
	HR	1.050	0.264
	TEMP	1.053	0.259
Ensemble model (4 models)	EDA$_{\mathrm{fast}}$	1.041	0.268
	EDA$_{\mathrm{fast}}$ and BVP	**1.033**	**0.276**
Human		1.008	0.406

(2) Ensemble of 3 models: The ensemble was constructed using the FNN$_{\mathrm{L+P}}$ (i.e., FNN$_{\mathrm{L}}$ and FNN$_{\mathrm{P}}$), and TPTr×3 models trained on physiological submodalities.

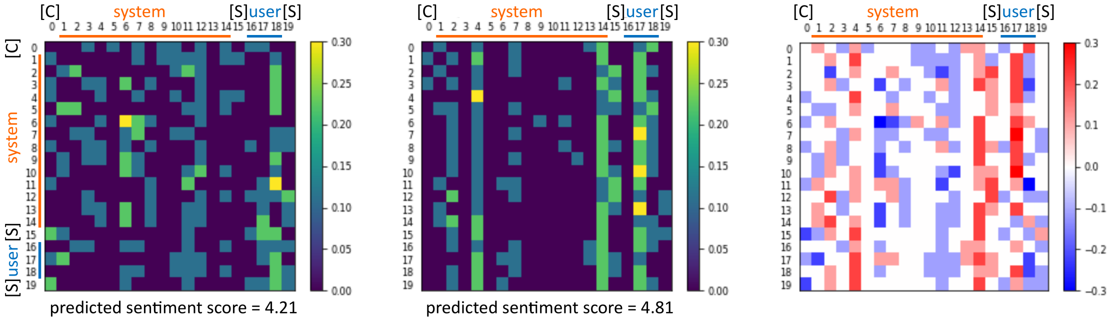
(3) Ensemble of 4 models: The ensemble was constructed using FNN$_{\mathrm{L+P}}$, and two models selected from Tr×3 or TPTr×3 trained on physiological submodalities. To compare the conventional Tr×3 model with our proposed TPTr×3 models, two sets of ensembles were evaluated: FNN$_{\mathrm{L+P}}$, TPTr×3 trained on EDA$_{\mathrm{fast}}$, Tr×3; FNN$_{\mathrm{L+P}}$, TPTr×3 trained on EDA$_{\mathrm{fast}}$, TPTr×3 trained on BVP.

Table 3 presents the estimation results of the abovementioned models. Among the five single models based on each submodality, the TPTr model based on the BVP signal has the best result (row 4 in Table 3). The TPTr model based on the BVP signal also had the best result for the ensemble of 3 models, with an MAE of 1.041 and a Corr of 0.269 (row 9 in Table 3). Finally, we evaluated the ensemble of 4 models: FNN$_{\mathrm{L+P}}$, TPTr×3 based on the EDA$_{\mathrm{fast}}$, and TPTr×3 based on the BVP signal (the second row from the bottom in Table 3). This ensemble model achieves the best result in this study, with an MAE of 1.033 and a Corr of 0.276. The ensemble of 4 models including Tr×3 has a worse performance in terms of the MAE (1.041) than the model without Tr×3 (MAE of 1.033, the last row in Table 2).

EDA$_{\mathrm{fast}}$ is known to be related to emotional arousal, and we have presented its effectiveness (Tables 1 and 2); however, the BVP signal could also be useful for sentiment estimation with our proposed framework. Both the EDA and BVP signals are related to the ANS; however, the EDA signal reflects changes in sweat gland activity, while the BVP signal reflects physiological changes in the cardiovascular system. Further improvement was achieved with the ensemble of 4 models by using the TPTr model based on the BVP signal; thus, different physiological submodalities may reflect different aspects of sentiment changes that cannot be explicitly represented by using linguistic information alone, resulting in the ensemble of 4 models achieving further performance improvement.



(a) Example of the attention weights with true negative self-reported sentiment (score of 3.00)



(b) Example of the attention weights with true positive self-reported sentiment (score of 5.00)

Figure 3: Example of the attention weights extracted from Transformer (left) and TPTr (center), and the difference between the two (right). Each square matrix is the attention weight computed from the $Q_iK_i^T$ (left) or $Q'_iK_i^T$ (center, please see equation 3). The dimension is equal to the total number of tokens including special tokens in one exchange. (a) Example of attention weights with true negative self-reported sentiment. (b) Example of attention weights with true positive self-reported sentiment. [C] and [S] indicate special tokens of BERT `[CLS]` and `[SEP]`, respectively.

On the other hand, other submodalities appeared to have little effect on the estimation performance. Thus, other time-series processing or feature extraction methods should be considered for these submodalities to determine whether they contribute to the sentiment estimation performance.

The last row in Table 3 depicts the sentiment estimation performance by five human annotators (the Cronbach alpha value was 0.83 for the third-party annotation, indicating the reliability of the third-party annotation). Our best MAE of 1.033 is close to the human performance, which had an MAE of 1.008, although there is still a gap between the correlation coefficients (0.276 vs. 0.406). Thus, the preprocessing method and neural network architecture could be improved. We focused on physiological signals in this study since physiological signals can capture sentiment changes that cannot be expressed by textual, acoustic and visual features. The combination of our proposed method and other nonverbal subnetworks for audiovisual modalities, such as those proposed in [29], may further improve the sentiment estimation performance; thus, additional investigations are needed.

5.4 Analysis of the Attention Weight

It is assumed that the incorporation of physiological signals into the Transformer architecture leads to changes in the attention weights since time-series physiological signals shift the query from Q to Q' in our proposed module (Fig. 2). Thus, we compared the attention weights between the Transformer and TPTr models. Test samples were used to extract attention weights from the learned model. Fig. 3 shows examples of attention weights with negative sentiment (Fig. 3(a)) and positive sentiment (Fig. 3(b)) derived from the Transformer (left) and TPTr (center) models, as well as their difference (right). The example shown in Fig. 3(a) has a true self-sentiment score of 3.00 (i.e., a negative example), and the estimated scores of the Transformer and TPTr models are 4.01 and 3.69, respectively. In this example, the segmented Japanese tokens of the system are "SO/NA/N/DESU/NE/," (number of tokens $n = 6$), which means "I got it" in English, and the Japanese token of the user is "HAI/," ($n = 2$), which means "Yes" or "Well", which generally functions as a filler and has a neutral or positive meaning. This ambiguous user utterance makes it difficult to estimate negative sentiment using only linguistic information; however, the TPTr model gives less attention to this neutral/positive token and estimates a value of 3.69, which is close to the true negative sentiment score. Conversely, the TPTr model pays more attention to user utterances in other cases, as shown in Fig. 3(b). This example has the system utterance "Which do you like better, sweet or spicy?" and the user utterance "I like both" in English. This example has a true self-sentiment score of 5.00 (i.e., a positive example), and the estimated scores of the Transformer and TPTr models are 4.21 and 4.81, respectively. Thus, the TPTr model may change the attention weight more flexibly

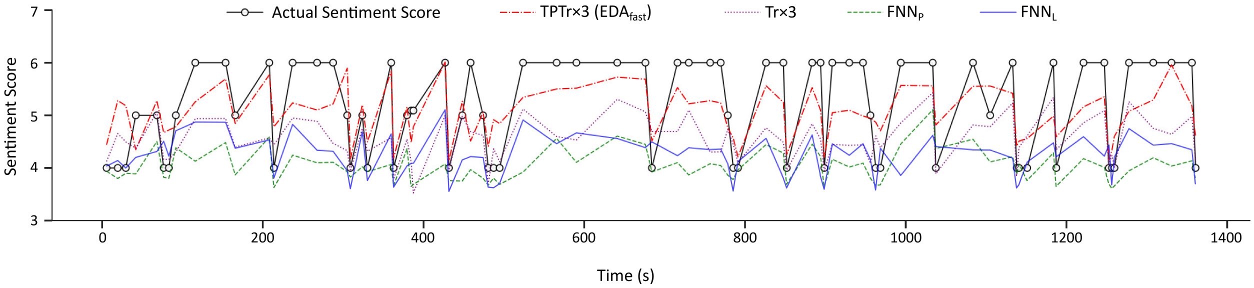


Figure 4: Qualitative example of the estimation pattern of each model. The black line with open circles indicates the actual sentiment score of a participant during each exchange in a dialog session. Each colored line indicates the estimated score based on each model during each exchange.

than the Transformer model, which may improve the ensemble model performance.

Taken together, our proposed TPTr architecture intuitively allows for shifting BERT representations to the physiology-related subspace, resulting in better estimation performance in the ensemble models. Our proposed models allow physiological information to be continuously linked to linguistic information and has a fixed attention direction, which is different from the prior works [4, 22, 27]. In the preliminary experiment, other architectural designs of the TPTr, such as another attention direction, degraded (or at least did not improve) the estimation performance. Thus, the time-series physiological signals play a supporting role to the Transformer based on the BERT representations (denoted as Q'_i in Section 3.2) by capturing self-sentiment changes that cannot be represented by linguistic information, although a further thorough investigation is needed.

5.5 Analysis of the Exchange-Level Estimation Pattern

To visualize exchange-level self-sentiment changes and differences in the estimation patterns among the models, an example of the estimation results during a dialog session is shown in Fig. 4. As shown by the black lines, the participant's self-sentiment changes dynamically during the dialog. Thus, self-sentiment estimation is a difficult task, and dialog systems should recognize and adapt to these sentiment changes at the exchange level. In this example, the conventional FNN$_L$ (blue line in Fig. 4, MAE of 0.954) and FNN$_P$ (green dashed line, MAE of 1.077) models cannot dynamically estimate the participant's sentiment and perform conservatively (estimated scores are almost neutral scores of 4). In addition, the conventional Transformer model (purple dotted line, MAE of 0.715) is insufficient for estimating positive sentiment, although some performance improvement is observed. On the other hand, the TPTr model (red dot-dashed line, MAE of 0.576) is effective in detecting subtle sentiment changes, particularly positive sentiment changes, which cannot be achieved by any of the other models presented in this example. Thus, the TPTr model could represent different aspects of sentiment changes that cannot be captured by using BERT representations or conventional Transformer.

5.6 Limitations and Future Works

There is no publicly available dataset that includes exchange-level self-sentiment labels and linguistic and physiological information except for the Hazumi dataset used in this study. Thus, we cannot evaluate our proposed model with another dataset, which will be considered in future work. Although our proposed method could contribute toward capturing short-time sentiment changes during individual exchanges (i.e., intraframe), our methods do not consider time-series changes in the overall dialog data (i.e., interutterance). Thus, the effectiveness of representations based on exchange sequences and attention mechanisms that capture more context and physiological changes merit further investigation. Additionally, there is a need to investigate effective methods for adding time-series audiovisual signals into TPTr (i.e., four modalities in total), and comparison with other SOTA language models such as RoBERTa [16] is also needed.

6 CONCLUSION

We showed that the model based on our proposed time-series physiological signal processing method has a comparable performance to linguistic-based models. Furthermore, the TPTr model, which introduced time-series physiological signals into a SOTA language model, significantly outperforms the previously reported best result. Furthermore, we presented that adding the BVP signal into the TPTr model based on the EDA$_{fast}$ signal resulted in further estimation performance improvement. It seemed that attention weights based only on the language modality can be changed by the injection of the physiological signals into TPTr which capture self-sentiment changes that are not expressed in linguistic information. Thus, our proposed framework could be valuable for developing novel techniques for extracting representations not only from linguistic modality but physiological modality.

ACKNOWLEDGMENTS

This work was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI (Grant Numbers 22H04860 and 22H00536) and JST AIP Trilateral AI Research, Japan (Grant Number JPMJCR20G6). This work was also partially supported by the Research Program of "Five-star Alliance" in "NJRC Mater & Dev.".

REFERENCES

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. https://doi.org/10.48550/ARXIV.1607.06450

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[3] Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. 2018. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, IEEE Computer Society, Los Alamitos, CA, USA, 59–66.

[4] Haifeng Chen, Dongmei Jiang, and Hichem Sahli. 2021. Transformer Encoder With Multi-Modal Multi-Head Attention for Continuous Affect Recognition. *IEEE Transactions on Multimedia* 23 (2021), 4171–4183. https://doi.org/10.1109/TMM.2020.3037496

[5] Michael E Darson, Anne Schell, and Diane L Filion. 2007. The electrodermal system. In *Handbook of psychophysiology*. Cambridge University Press, New York, NY, USA, 159–182.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[7] Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. Multimodal Affective Analysis Using Hierarchical Attention Strategy with Word-Level Alignment. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 2225–2235. https://doi.org/10.18653/v1/P18-1207

[8] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. MISA: Modality-Invariant and -Specific Representations for Multimodal Sentiment Analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*. Association for Computing Machinery, New York, NY, USA, 1122–1131. https://doi.org/10.1145/3394171.3413678

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 770–778. https://doi.org/10.1109/CVPR.2016.90

[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

[11] Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. 2020. Is She Truly Enjoying the Conversation? Analysis of Physiological Signals toward Adaptive Dialogue Systems. In *Proceedings of the 2020 International Conference on Multimodal Interaction*. Association for Computing Machinery, New York, NY, USA, 315–323. https://doi.org/10.1145/3382507.3418844

[12] Shun Katada, Shogo Okada, and Kazunori Komatani. 2022. Effects of Physiological Signals in Different Types of Multimodal Sentiment Estimation. In *IEEE Transactions on Affective Computing*. IEEE Computer Society, Los Alamitos, CA, USA. https://doi.org/10.1109/TAFFC.2022.3155604

[13] Jonghwa Kim and Elisabeth André. 2008. Emotion recognition based on physiological changes in music listening. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30, 12 (2008), 2067–2083.

[14] Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. 2011. DEAP: A database for emotion analysis; using physiological signals. *IEEE Transactions on Affective Computing* 3, 1 (2011), 18–31.

[15] Kazunori Komatani and Shogo Okada. 2021. Multimodal Human-Agent Dialogue Corpus with Annotations at Utterance and Dialogue Levels. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE Computer Society, Los Alamitos, CA, USA, 1–8. https://doi.org/10.1109/ACII52823.2021.9597447

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach.

[17] Marco Maier, Daniel Elsner, Chadly Marouane, Meike Zehnle, and Christoph Fuchs. 2019. DeepFlow: Detecting Optimal User Experience From Physiological Data Using Deep Neural Networks. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, Palo Alto, CA, USA, 1415–1421. https://doi.org/10.24963/ijcai.2019/196

[18] John D. Mayer and Peter Salovey. 1993. The intelligence of emotional intelligence. *Intelligence* 17, 4 (1993), 433–442. https://doi.org/10.1016/0160-2896(93)90010-3

[19] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. 2021. AMIGOS: A Dataset for Affect, Personality and Mood Research on Individuals and Groups. *IEEE Transactions on Affective Computing* 12, 2 (2021), 479–493. https://doi.org/10.1109/TAFFC.2018.2884461

[20] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. 2001. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 10 (2001), 1175–1191.

[21] Jessica Sharmin Rahman, Tom Gedeon, Sabrina Caldwell, Richard Jones, Md Zakir Hossain, and Xuanying Zhu. 2019. Melodious micro-frissons: detecting music genres from skin response. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, IEEE Computer Society, Los Alamitos, CA, USA, 1–8.

[22] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. 2020. Integrating Multimodal Information in Large Pretrained Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2359–2369. https://doi.org/10.18653/v1/2020.acl-main.214

[23] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, IEEE Computer Society, Los Alamitos, CA, USA, 1–8.

[24] Björn Schuller, Stefan Steidl, and Anton Batliner. 2009. The interspeech 2009 emotion challenge. In *Tenth Annual Conference of the International Speech Communication Association*. ISCA, Brighton, United Kingdom, 312–315.

[25] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Seattle, Washington, USA, 1631–1642. https://aclanthology.org/D13-1170

[26] Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. 2011. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* 3, 1 (2011), 42–55.

[27] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6558–6569. https://doi.org/10.18653/v1/P19-1656

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. Curran Associates, Inc., Red Hook, NY, USA, 5998–6008.

[29] Yansen Wang, Ying Shen, Zhun Liu, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2019. Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence* (Honolulu, Hawaii, USA) *(AAAI'19/IAAI'19/EAAI'19)*. AAAI Press, Palo Alto, CA, USA, Article 886, 8 pages. https://doi.org/10.1609/aaai.v33i01.33017216

[30] Ashima Yadav and Dinesh Kumar Vishwakarma. 2020. Sentiment analysis using deep learning architectures: a review. *Artificial Intelligence Review* 53, 6 (2020), 4335–4385.

[31] Heath Yates, Brent Chamberlain, Greg Norman, and William H. Hsu. 2017. Arousal Detection for Biometric Data in Built Environments using Machine Learning. In *Proceedings of IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing (Proceedings of Machine Learning Research, Vol. 66)*, Neil Lawrence and Mark Reid (Eds.). PMLR, Cambridge, MA, USA, 58–72. https://proceedings.mlr.press/v66/yates17a.html