Master's Thesis

Conversational Context-Aware Physiological and Linguistic Fusion for Self-reported Sentiment Analysis

TRAN MINH DUC

Supervisor: Prof. SHOGO OKADA

Graduate School of Advanced Science and Technology
Japan Advanced Institute of Science and Technology
(Information Science)

September, 2023

Abstract

Developing dialog systems capable of dynamically adapting to a user's sentiment state in real time is a challenging task. Existing multimodal models have demonstrated impressive performance in estimating third-party labeled sentiment levels by incorporating features from linguistic, visual, and speech modalities. On the other hand, physiological signals play a crucial role in estimating self-reported sentiment as they exhibit involuntary changes associated with emotions. Previous studies have shown the effectiveness of fusing physiological and linguistic features for self-reported sentiment estimation.

However, these studies often neglect the contextual interaction between exchanges, where each exchange consists of a pair of system and user utterances. To address this gap, we propose an efficient approach that incorporates interplay of physiological features between exchanges in dialogue, which consists of a system utterance followed by a user utterance. Specifically, we introduce a framework that combines linguistic and physiological signals across exchanges. Our approach employs attention mechanisms to capture contextual information and long-term dependencies within the dialog, enabling a comprehensive understanding of sentiment evolution. Additionally, we leverage convolutional neural networks (CNNs) to learn robust representations from physiological signals, enhancing the interpretation of the user's emotional changes.

Through extensive experiments, our approach surpasses existing multimodal models in sentiment estimation. Our findings highlight the importance of inter-exchange learning for effective sentiment adaptation in dialog systems. By considering time-series changes in linguistic and physiological features across multiple exchanges, our approach captures the dynamical changes of sentiment level, leading to more accurate and adaptive dialog interactions.

Acknowledgements

I would like to take this opportunity to express my heartfelt gratitude to the individuals who have been instrumental in making my academic journey a fulfilling and enriching experience.

First and foremost, I am indebted to Professor Shogo Okada for his exceptional mentorship and guidance. His expertise, patience, and dedication to teaching have profoundly influenced my academic growth. I am grateful for his valuable insights, constructive feedback, and belief in my potential, which have significantly shaped my research and career aspirations.

I extend my heartfelt appreciation to all the members of the Okada Lab. Your camaraderie, collaboration, and shared passion for research have made the lab an inspiring and dynamic environment to work in. The collective knowledge and support within the lab have contributed immensely to my growth as a researcher. I would also like to thank my friends for being pillars of support, providing a sense of belonging, and making my academic journey enjoyable.

Lastly, I am deeply thankful to my family for their unwavering love, support, and encouragement throughout my educational pursuit. Their belief in my abilities and constant encouragement have been a source of strength and motivation during both challenging and joyous times.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 General Introduction

Creating an adaptive dialog system capable of accurately recognizing and adapting to a user's real-time state is important for encouraging interesting and entertaining human-agent interactions. The system should dynamically modify its behavior based on the user's present emotional state during a chat session. For instance, the algorithm should proactively steer the conversation away from the present topic if the user is getting bored with it, imitating human behavior. This can be achieved by using Natural Language Processing (NLP) techniques such as sentiment analysis or emotion detection in order to detect changes in emotions within conversations.

Incorporating NLP techniques, such as sentiment analysis or emotion detection, plays a crucial role in achieving this level of adaptability. By leveraging these techniques, the dialogue system can detect changes in the user's emotions within conversations. For instance, it can discern shifts from positive to negative sentiment or fluctuations in emotional intensity during the course of the conversation. Additionally, sentiment analysis and emotion detection can assist the system in better understanding the underlying context and tone of the user's messages, allowing it to respond more appropriately and empathetically. When the user expresses joy or satisfaction, the system can acknowledge their positive sentiment and provide positive reinforcement or relevant content. Conversely, if the user appears frustrated or upset, the system can offer empathetic responses or suggest solutions to address their concerns.

An adaptive dialogue system can also benefit from multimodal sentiment analysis, where the integration of linguistic information with nonverbal cues, such as facial expressions and vocal tones, enables a more comprehensive understanding of the user's emotional state. This holistic approach helps the system to accurately gauge the user's emotions and tailor its responses accordingly, further enhancing

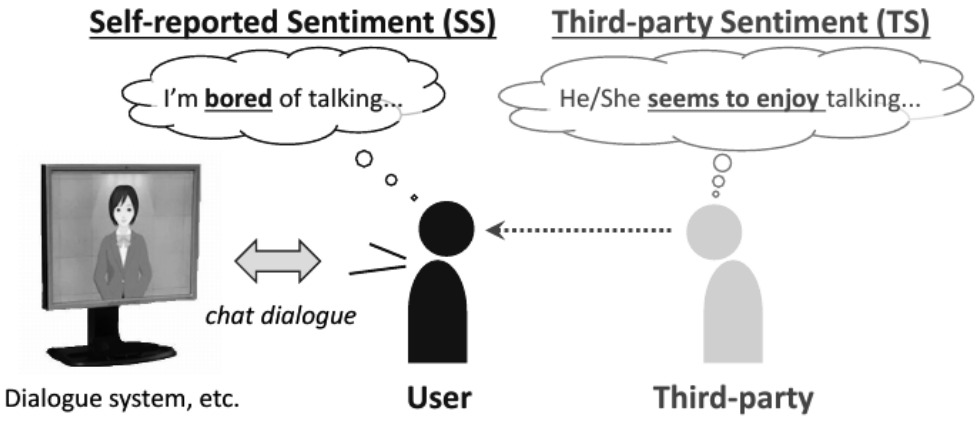

Figure 1.1: An example of the discrepancy between self-reported sentiment and third-party sentiment annotations.

the quality of interactions. Moreover, ongoing research in the field of affective computing and emotion-aware dialogue systems can contribute to refining the system's abilities. Continual advancements in machine learning algorithms and large-scale datasets for emotion recognition can lead to improved accuracy and robustness in detecting emotions from user inputs.

In conclusion, building an adaptive dialogue system that can recognize and adapt to a user's real-time emotional state is essential for creating engaging and meaningful human-agent interactions. By leveraging NLP techniques, such as sentiment analysis and emotion detection, alongside multimodal approaches, the system can dynamically adjust its behavior to provide more empathetic and personalized responses. This advances the vision of human-like conversational AI and fosters a more enjoyable and satisfying user experience.

1.2 Challenges and Motivations

Building adaptive dialogue system is difficult for numerous reasons. One major challenge arises from the fact that a user's self-reported sentiment may not always manifest explicitly in the linguistic information derived from their utterances. Self-reported sentiment is the user's own perception of their emotional state, which may not always be accurately reflected and recognized by others. Figure 1.1 shows an example of self-reported sentiment and third-party sentiment annotations. In this example, the user's self-reported sentiment is negative, but the third-party

sentiment is positive. People have different levels of emotional intelligence and may purposefully hide their genuine feelings in their thoughts instead of expressing them [1]. This phenomenon, known as "emotional masking", poses a significant hurdle to accurately identifying a user's emotional state solely based on linguistic cues. To address this challenge, researchers have explored alternative sources of information, such as voice [2] and facial expressions [3, 4], to uncover hidden emotional nuances that may not be recognized by linguistic analysis. By leveraging these additional aspects of human communication, dialog systems can gain deeper insights into a user's underlying emotional state and adapt their responses accordingly, thereby enhancing the overall effectiveness and naturalness of the conversation.

Despite the existing investigations into the use of physiological signals for self-sentiment estimation and the integration of linguistic information from state-of-the-art language models, an important aspect that has often been overlooked is the incorporation of inter-exchange information within a conversation. Dialog systems should be able to capture the dynamics and evolving sentiment across multiple exchanges to adapt their responses effectively. However, current approaches such as [5, 6] primarily focus on individual exchanges and fail to consider the interplay and contextual dependencies between them. This limitation hinders a comprehensive understanding of sentiment evolution and restricts the system's ability to adapt in real-time. Therefore, it is necessary to resolve the lack of inter-exchange information in sentiment estimation models to enable a more accurate and comprehensive representation of sentiment adaptation in dialog systems. By incorporating inter-exchange learning, dialog systems can capture the nuanced aspects of sentiment evolution, enabling more precise and adaptive interactions with users.

Figure 1.2 illustrates the challenges addressed in this study regarding the integration of physiological signals and the interplay of exchanges in sentiment estimation. The figure depicts a sequence of dialog exchanges between a user and a dialog system, with each exchange represented by a textual utterance. In each exchange, the corresponding physiological signals, e.g. Blood Volume Pulse, and user's sentiment score are shown. The visualization highlights the need for capturing inter-exchange information to understand the change of sentiment during the conversation. In addition, another issue in multimodal sentiment analysis is that there are a limited number of works that fully leverage these signals and jointly learn their feature representation along with linguistic data in an end-to-end manner.

Several studies used deep neural networks that are trained end-to-end to extract physiological features [7, 8], but they applied to emotion recognition task. While [9, 6] exploited multimodal learning for sentiment estimation, they hand-designed all the physiological signal features. These approaches can be limiting, as they may not capture the full potential of the physiological signals and may require

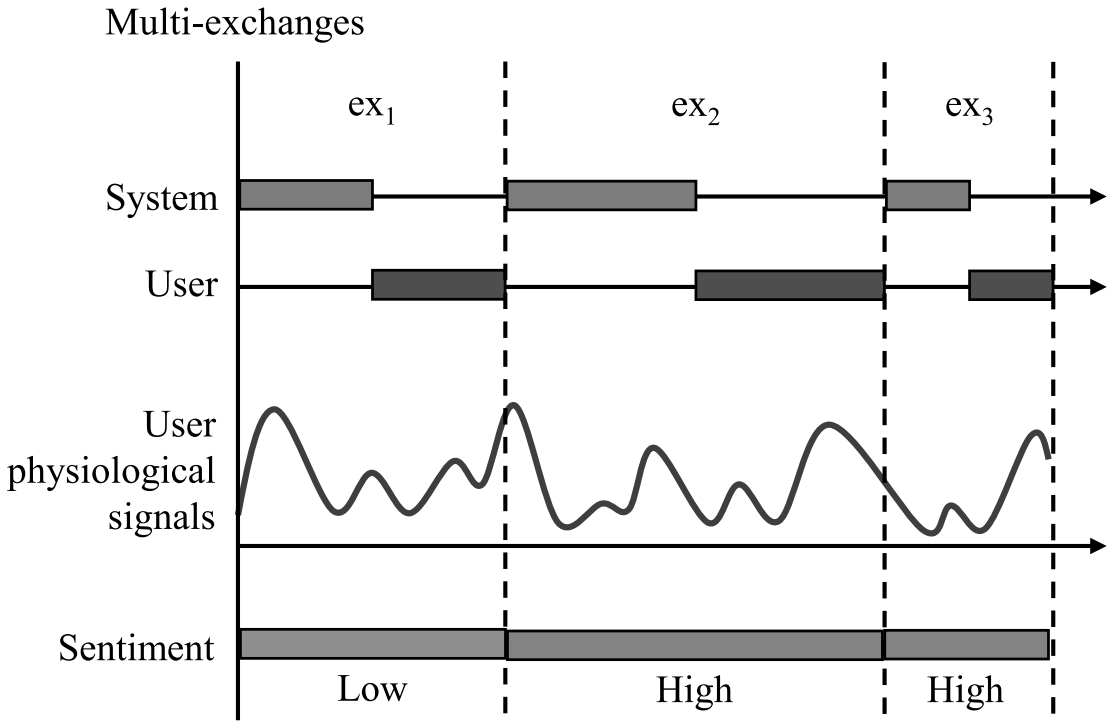

Figure 1.2: Example of capturing self-reported sentiment changes by using multiple exchanges information. The users' physiological signal may change slowly, taking effect after several turns.

domain expertise to design effective features. To overcome this limitation, our proposed approach employs convolutional neural networks (CNNs) to learn robust representations directly from physiological signals. By leveraging the power of deep learning, our framework enables an end-to-end learning process, allowing the model to automatically extract relevant features from the physiological signals, enhancing the utilization of these signals for sentiment adaptation in dialog systems.

1.3 Originality

In conclusion, this study primarily centers on two key contributions: the incorporation of inter-exchange information and physiological signals in sentiment estimation for dialog systems. We prioritize linguistic information and physiological signals, as previous research has shown that models utilizing these modalities outperformed visual and audio modalities [5]. The key highlights of our contributions include:

- Physiological signal representation with CNNs: The proposed approach lever-

ages convolutional neural networks (CNNs) to learn robust representations directly from physiological signals. The model automatically extracts relevant features from the physiological signals, enhancing the utilization of these signals for sentiment adaptation.

- Incorporation of inter-exchange information: This study addresses the limitations of existing sentiment estimation models by proposing an approach that captures inter-exchange information within dialog systems. By considering the dynamic changes of sentiment across multiple exchanges, the proposed approach enables more accurate and adaptive interactions with users.

1.4 Thesis Organization

We organize the structure of this report into five chapters. The first chapter provides an overview of the research topic, its significance, and the research objectives of this study. The remaining chapters are organized as follows:

- **Chapter 2: Related Works and Background** conducts a comprehensive analysis of the relevant literature, establishing a theoretical framework and identifying gaps for the current research.
- **Chapter 3: Methodology** outlines the chosen research approach and details the proposed network architecture.
- **Chapter 4: Experiments and Results** presents the datasets, experimental setup, and results of experiments.
- **Chapter 5: Conclusion and Future Works** summarizes the key findings, draws conclusions, highlights the study's contributions, and provides recommendations for future research.

Chapter 2

Related Works and Background

This chapter provides an overview of the related works in the field of multimodal sentiment analysis and physiological signal processing. We first introduce the text based sentiment analysis in Section 2.1. Afterthat, we discuss the physiological signal based sentiment analysis in Section 2.2. Section 2.3 discusses the multimodal sentiment analysis. Finally, Section 2.4 presents the background knowledge of several deep learning techniques, which were used in this thesis.

2.1 Text based Sentiment Analysis

Sentiment analysis, a critical area of NLP, has witnessed significant advancements with the rise of neural network models. While conventional lexicon-based or handcrafted feature methods have been extensively explored, this section concentrates on recent state-of-the-art (SOTA) neural network models that have revolutionized sentiment analysis.

In recent times, neural network models have gained immense popularity in sentiment analysis tasks [10]. Notably, Convolutional Neural Networks, Long Short-Term Memory networks, and their variants have demonstrated impressive performance. Kim [11] introduced simple CNN models that achieved SOTA results on multiple datasets, including the Stanford Sentiment Treebank v2 (SST-2) dataset, boasting an accuracy of 88.1%. A notable contribution in the realm of contextual word representations came from Peters et al. [12], who proposed embeddings from language models (ELMo) based on a Bidirectional Long Short-Term Memory (BiLSTM) approach. These deep contextualized word representations led to new SOTA performances on six NLP tasks, including sentiment analysis.

In addition to LSTM-based approaches, Vaswani et al.[13] introduced a revolutionary network architecture called the transformer. Built solely on an attention mechanism, transformers achieved substantial improvements in computational effi-

ciency through parallelization and demonstrated superior performance in machine translation tasks.

Following the transformer's success, Devlin et al.[14] developed BERT (Bidirectional Encoder Representations from Transformers), a groundbreaking language model based on a multilayer bidirectional transformer encoder. BERT remarkably advanced the state-of-the-art on eleven NLP tasks, achieving an impressive 94.9% accuracy on the SST-2 dataset. Consequently, BERT has become the standard for NLP tasks and is widely adopted in various applications.

While these text-based neural network models have showcased exceptional performance, it is essential to recognize that sentiment expression can be influenced by multiple factors. Spoken language, often noisier and less structured than written language[15], requires additional considerations beyond linguistic information for sentiment analysis in dialogues. Explicit user sentiment, expressed as textual information, can undoubtedly benefit from powerful NLP tools like BERT. However, to comprehensively understand user sentiment in dialogues, a multimodal approach becomes imperative.

A multimodal approach to sentiment analysis involves the integration of information from various modalities, such as text, audio, images, and videos. By fusing signals from different sources, we can capture nuanced emotions and sentiment cues that may not be adequately conveyed through text alone. This integration can potentially lead to more accurate and holistic sentiment estimation, especially in dynamic and interactive contexts like dialogues.

In conclusion, while text-based neural network models have significantly advanced sentiment analysis, we recognize the importance of a multimodal approach, considering the complexity and diversity of factors that influence sentiment expression in spoken language. Integrating powerful tools like BERT with information from other modalities holds promise in enhancing the understanding of user sentiment during dialogues and facilitating more contextually aware human-computer interactions.

2.2 Physiological Signal based Sentiment Analysis

This section specifically focuses on research related to the use of deep neural networks for modeling the physiological signal

The comprehension of emotions and sentiments is a multifaceted process that can be significantly enhanced by incorporating information from various modalities, including language, visual cues, audio, and physiological signals. Among these modalities, physiological signals offer a distinctive advantage in extracting

implicit emotional responses, as they are not easily controllable by individuals on a conscious level. For example, when observing someone watching a movie or participating in a video game, their emotional state may not be overtly expressed through external cues, making it difficult to gauge their level of interest and engagement solely based on external appearances.

In recent years, there has been a growing trend toward utilizing multiple modalities, including visual, audio, and text, to recognize user sentiment [16]. This trend can be attributed to the significant advancements in deep learning techniques for processing and analyzing these modalities. However, they are outward signals that could be masked by users [17, 8]. To overcome the limitations of traditional modalities, researchers have turned their attention to physiological signals as a valuable source of information for affective computing. For example, previous work has leveraged Electroencephalogram (EEG), Electrocardiogram (ECG), etc. to recognize user emotions [18, 19]. These signals are regarded as physiological modalities, different from behavioral modalities such as facial expressions. Specifically, [18] used frontal EEG signals to identify four emotional states: happy, fear, peace, and disgust. They let the participants watch the corresponding video while recording their brain signals.

Many approaches have been proposed to utilize physiological modalities to detect an individual's emotion, from hand-crafted feature modeling to applying deep neural networks. For instance, Zhao et al [17] extracted the distinct emotion-related features from the time domain, frequency domain, and nonlinear analysis. Xu et al. [18] applied similar approaches to EEG data by extracting features from the time domain, frequency domain, and space domain. They then filtered to reduce the number of features using various feature selection techniques. These features were then used as input to traditional machine learning algorithms like support vector machines or feedforward neural networks. However, these approaches often required domain expertise and limited the ability to capture complex patterns in the data. Thus, recent studies focused on developing deep neural networks to learn informative representations from physiological signals.

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), such as long short-term memory (LSTM) networks, have been employed to capture the temporal dynamics and spatial patterns in physiological data. For example, Wang et al. [20] and Yang et al. [21] proposed using CNNs, Zitouni et al. [22] proposed using LSTM to extract the high-level feature representation from four kinds of physiological signals and predict the users' arousal-valence states. These deep learning models have shown promising results in capturing the nuanced aspects of emotion and sentiment encoded in physiological signals. Even so, one major issue is that many works focus on recognizing users' emotion scores, but sentiment analysis using physiological signals is not widely studied. Katada et al. [6] studied

the effects of physiological signals in different types of multimodal sentiment estimation. They proposed using LSTM to verify the effect of physiological signals on sentiment recognition, then adapted a multimodal Transformer architecture to jointly learn with other modalities. However, there is a limitation is that they overprocessed the raw sigals, which might cause a loss of information.

In conclusion, physiological signals are a valuable source of information for sentiment analysis. Therefore, in this work, we propose a novel deep learning model to extract the high-level feature representation from physiological signals and predict the users' sentiment scores.

2.3 Multimodal Sentiment Analysis

Nonverbal information processing is a vital technique employed to extract a user's sentiment from sources other than linguistic data. Human communication encompasses not only natural language but also nonverbal behaviors like facial expressions [23, 24], vocal behavior [25], and gestures [26]. In the domain of facial expressions, Ekman and Friesen's Facial Action Coding System (FACS) is widely used to map emotions, contributing to various affective computing research endeavors. Vocal behavior, on the other hand, reveals emotions through acoustic features such as loudness, pitch, and rhythm, and significant efforts have been dedicated to understanding the relationship between vocal behavior and emotion [27]. While there are relatively fewer gesture-based emotion studies compared to facial expressions and vocal behavior, gestures are also essential in conveying emotions. For example, high-frequency hand clapping often expresses joy and satisfaction [28]. These nonverbal cues are collectively known as social signals [27], and their processing, termed social signal processing, is frequently employed to build automatic user state estimation models for adaptive dialogue systems [29]. In sentiment analysis, social signals such as facial expressions, body gestures, and prosody are frequently utilized as nonverbal information sources [30].

More recently, researchers have noted the limitations of single modalities, especially outward modalities. They shifted their focus to combining multiple types of data in the hopes of gaining a deeper understanding of human emotions. The introduction of Transformer models, in particular BERT, has transformed language modeling and text-based sentiment estimation [13, 14, 31]. The incorporation of audiovisual features into multimodal Transformer models has also been proposed [32, 33, 34]. MulT [35] was the first model for sentiment analysis across multiple modalities. MulT introduced a crossmodal attention mechanism that efficiently adapts information across modalities. Later on, Devamanyu et al. [33] improved feature representations by introducing modality-invariant and modality-specific representation modules.

Although previous studies have focused on text and audiovisual modalities, there is a growing interest in investigating the incorporation of physiological signals into multimodal affective analysis. Several datasets have been created for emotion and sentiment research using videos or conversational interactions [36, 37, 38, 39, 40]. These datasets offer valuable resources for investigating the effectiveness of physiological signals in capturing subtle sentiment changes that may not be evident in explicit textual or audiovisual information. However, [40] was the only dataset that contained both textual and physiological information for the sentiment analysis task. The dataset was generated by allowing participants to chat with agents, and the participants themselves annotated the sentiment labels. To capture long-term dependencies between physiological signals and the corresponding linguistic tokens, a time-series multimodal Transformer model, which was inspired by Mult, was proposed in [5]. However, these methods only considered the temporal dependencies between signals within a single utterance, limiting their ability to capture the dynamic changes across exchanges.

In the first step of our proposed method, CNNs are used to represent the physiologically significant features. Then, we propose a method that leverages LSTM to model the interaction between utterances for estimation of sentiment during human-agent interactions. Our methodology is intended to capture the dynamic nature of sentiments within a conversation by utilizing the sequential information from exchanges. To evaluate the efficacy of our method, we use the Hazumi1911 dataset [40], the only publicly available dataset containing both time-series textual and physiological information.

2.4 Background

2.4.1 Convolutional Neural Networks

Convolutional Neural Networks (CNNs) have revolutionized the field of computer vision, enabling remarkable advancements in image recognition tasks. Originally developed to process 2D grid-like data, CNNs have since been extended to handle various data types, including time series data. In this section, we introduce the fundamental concepts of CNNs and delve into their innovative application in time series feature representation.

Convolutional Neural Networks

Convolutional Neural Networks belong to a category of deep learning models that possess the ability to automatically acquire hierarchical representations from input data. They excel in tasks that involve spatial relationships, making them particularly well-suited for image processing. At the heart of a CNN lies the convolutional layer, responsible for applying filters or kernels across the input

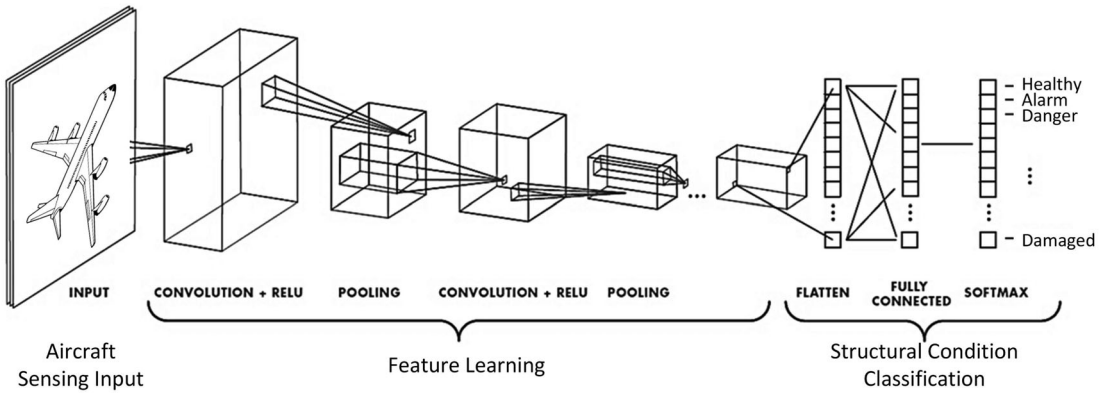

Figure 2.1: An example of a convolutional neural network.

data to identify and extract local patterns and distinctive features. These filters slide over the input, performing element-wise multiplications and summing the results to create feature maps that capture specific patterns. The ability of CNNs to learn hierarchical features, where lower layers detect simple patterns like edges, and deeper layers recognize complex structures, makes them powerful tools for feature extraction in various data domains. A visualization of the convolutional layer is shown in Figure2.1. The input is a 2D grid-like structure, such as an image, and the convolutional filters slide over the input, performing local feature extraction at each location. The filters are typically small in size, such as 3x3 or 5x5, and are applied across the entire input. The filters are learned during the training process, and the resulting feature maps are passed to the next layer in the network.

Application of CNNs in Time Series Feature Representation

While initially designed for image processing, CNNs have been adapted to handle time series data effectively. Time series data is a sequence of values recorded at successive time points, making it fundamentally different from grid-like data. To apply CNNs to time series data, we need to reinterpret the 1D temporal structure as a 2D grid-like structure, where one axis represents time steps and the other axis corresponds to different features or channels. To capture temporal patterns, the convolutional filters slide over the time series, performing local feature extraction at each time step. By learning relevant patterns and features, CNNs excel in recognizing temporal dependencies and capturing characteristic patterns present in the data.

In summary, the adaptation of Convolutional Neural Networks to time series data allows for automatic feature extraction, hierarchical representation, and robustness to time shifts. These advantages make CNNs highly effective tools for time series feature representation, enabling advanced analyses and predictions in

diverse domains.

2.4.2 Recurrent Neural Networks

In this section, we delve into the world of Recurrent Neural Networks (RNNs), a specialized class of deep learning models designed to tackle sequential data. RNNs differ from traditional feedforward neural networks in their ability to capture temporal dependencies, making them highly effective for modeling dynamic sequences such as time series data, natural language, and audio.

At the core of RNNs lies their recurrent nature, which enables them to maintain hidden states that persist across time steps. Each RNN unit processes the input at the current time step along with the hidden state from the previous time step, allowing the model to retain memory of past information. This feedback loop allows RNNs to understand the sequential nature of the data and make predictions based on the context learned from preceding elements in the sequence.

The simple formulation of RNNs can be represented as follows:

$$h_t = \sigma(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$
$$y_t = \sigma(W_{hy}h_t + b_y)$$

where x_t is the input at time step t, h_t is the hidden state at time step t, y_t is the output at time step t, W_{xh} is the weight matrix for the input, W_{hh} is the weight matrix for the hidden state, W_{hy} is the weight matrix for the output, b_h is the bias vector for the hidden state, b_y is the bias vector for the output, and σ is the activation function.

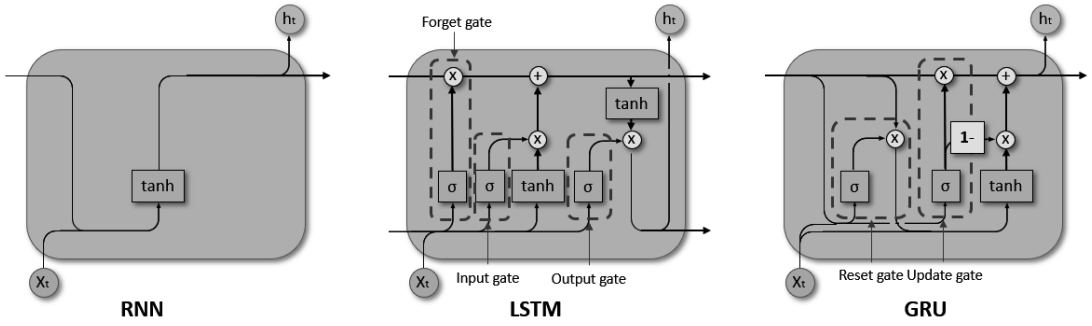

Figure 2.2: Visualization of RNN, LSTM, and GRU.

Over time, various variants of RNNs have been developed to address some of their limitations, such as difficulty in capturing long-term dependencies. Some of the popular RNN variants are visualized in Figure2.2 and described below:

- Long Short-Term Memory (LSTM) [41]: LSTM is designed to overcome the vanishing gradient problem in traditional RNNs, which hampers their ability to capture long-term dependencies. LSTM introduces gating mechanisms that enable the model to retain important information over extended time periods, making it well-suited for tasks requiring memory of distant events.
- Gated Recurrent Unit (GRU) [42]: GRU is a simplified version of LSTM that also employs gating mechanisms but with fewer parameters. It strikes a balance between LSTM and traditional RNNs, providing efficient memory management while being computationally lighter.
- Bidirectional RNN (BiRNN) [43]: BiRNN processes the input sequence in both forward and backward directions, allowing the model to access future as well as past context. This bidirectional context incorporation can enhance the understanding of the entire sequence.

In conclusion, Recurrent Neural Networks are powerful models for handling sequential data due to their recurrent architecture, which allows them to capture temporal dependencies. The introduction of variants like LSTM, GRU, and BiRNN has further improved the ability of RNNs to model long-term dependencies and complex sequential patterns. These models have revolutionized various applications in natural language processing, time series analysis, and sequential data processing, and continue to be a focus of research and development in the deep learning community.

2.4.3 Transformer

The Transformer is a groundbreaking deep learning architecture that has had a profound impact on the field of NLP. Introduced in the paper "Attention is All You Need" by Vaswani et al. [13], the Transformer has rapidly become the de facto standard for various NLP tasks, including machine translation, text generation, and language understanding. Before the advent of the Transformer, traditional sequence-to-sequence models, such as RNNs and LSTM networks, were widely used for NLP tasks. However, these models faced challenges in capturing long-range dependencies and suffered from computational inefficiencies due to sequential processing. The Transformer architecture addresses these limitations by introducing the attention mechanism, which allows the model to focus on relevant parts of the input sequence while processing each element. This parallelization capability significantly speeds up training and enables the model to capture dependencies between distant elements in the sequence, revolutionizing the way sequential data is processed. Figure 2.3 visualizes the Transformer architecture.

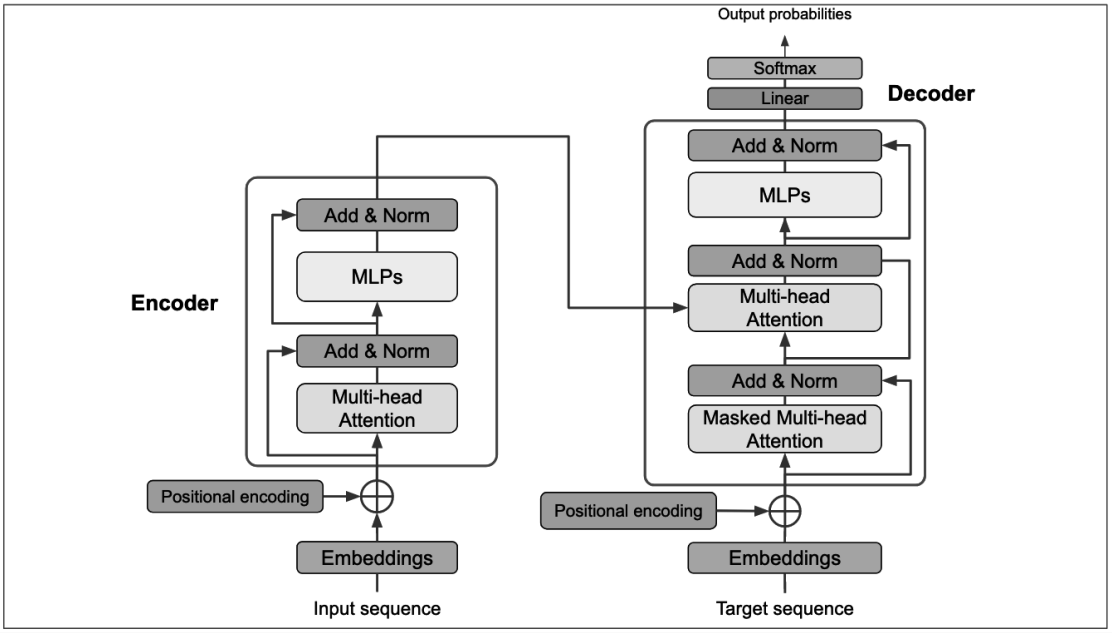

Figure 2.3: Transformer architecture.

The Transformer follows an encoder-decoder architecture, which is common in sequence-to-sequence tasks like machine translation. The encoder processes the input sequence and produces context-aware representations, while the decoder generates the output sequence based on the encoder's context. The self-attention mechanism in the encoder allows the model to understand the relationships among different elements in the input sequence, while the decoder's attention mechanism helps focus on the relevant parts of the encoder's representations during generation.

The key innovation in the Transformer is the self-attention mechanism. Unlike traditional models that process sequences sequentially, the Transformer computes the importance of each input element relative to the others. This is achieved by calculating attention scores between all pairs of elements in the input sequence, producing an attention matrix. The attention matrix is used to compute weighted sums of the input elements, where the weights represent the importance of each element based on its relevance to others. This allows the model to capture long-range dependencies efficiently and effectively.

Technically speaking, attention measures the relevance of each element in the input sequence to the current element being processed. This is done by computing an attention score between the current element and every other element in the sequence, producing an attention vector. The attention vector is then used to compute a weighted sum of the input elements, where the weights represent the importance of each element based on its relevance to the current element. This

allows the model to capture long-range dependencies efficiently and effectively. The mathematical formulation of the attention mechanism is as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{2.1}$$

Where Q is the query matrix, K is the key matrix, V is the value matrix, and d_k is the dimension of the key vectors. The query matrix is used to compute the attention scores between the current element and every other element in the sequence, while the key matrix is used to compute the attention scores between the current element and every other element in the sequence. The value matrix is used to compute the weighted sum of the input elements, where the weights represent the importance of each element based on its relevance to the current element. The attention scores are computed by taking the dot product of the query and key vectors, and then dividing by the square root of the dimension of the feature vectors. The attention scores are then normalized using the softmax function, which ensures that the weights sum to one. The weighted sum of the input elements is then computed by multiplying the attention scores with the value vectors. This kind of attention mechanism is called scaled dot-product attention, and could be visualized in Figure 2.4.

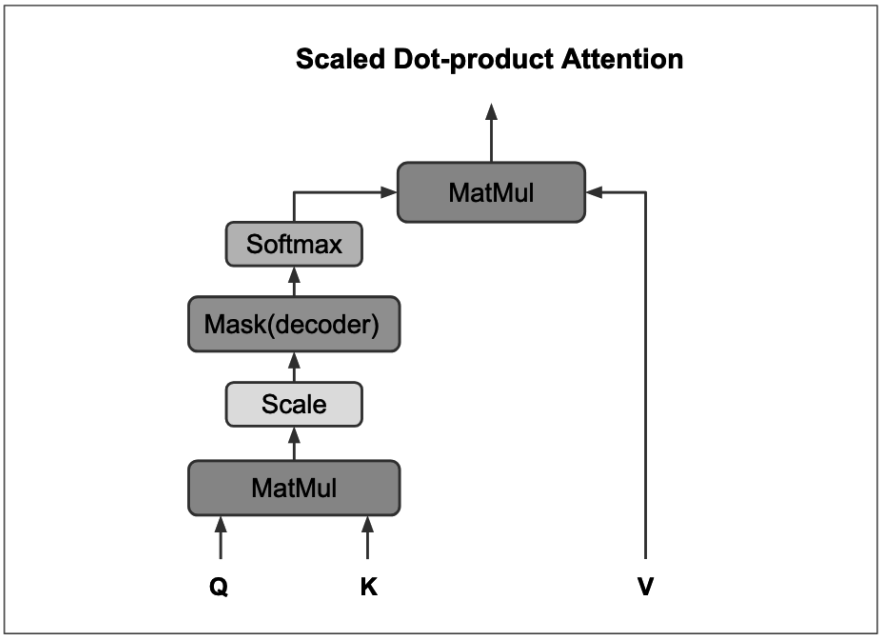

Figure 2.4: Graphical representation of scaled dot-product attention.

Transformer uses multi-head attention mechanism, which allows the model to jointly attend to information from different representation subspaces at different

positions. This enables the model to learn multiple representations of the input sequence and capture different types of dependencies. The multi-head attention mechanism is also used in the decoder to combine information from different parts of the encoder's representations.

Since its introduction, several variants of the Transformer have been proposed to enhance its capabilities further. Notably, models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) have achieved remarkable success in a wide range of NLP tasks. The Transformer has also been applied to other domains, including computer vision and speech recognition, with promising results. The Transformer architecture has revolutionized the field of NLP and continues to be a focus of research and development in the deep learning community.

Chapter 3

Methodology

Our goal is to employ a sequence of utterances to capture more context and physiological changes during conversation. Therefore, we proposed a many-to-many recurrent neural network that utilizes time-series physiological and linguistic data from multiple exchanges to predict user sentiment. An illustration of our proposed approach is shown in Figure 3.1. In our work, we consider physiological signals retrieved from a wearable device as well as linguistic data. In Section 3.1, we describe the techniques used to preprocess time-series physiological data. Then, in Section 3.2, we present the architecture of convolutional neural network, which will be trained end-to-end to learn representations of physiological signals. The linguistic data is processed using a pre-trained BERT model, which is described in Section 3.3. Finally, we introduce our proposed framework for utilizing multimodal data from multiple exchanges in Section 3.4. We expect this framework to be able to detect sudden shifts from positive to negative sentiment by analyzing the sequence of utterances, given that sentiment can change abruptly during conversation.

3.1 Physiological Signal Preprocessing

The physiological data used in this study were collected using a wristband device with multiple sensors. Specifically, we used four signals, including electrodermal activity (EDA), blood volume pulse (BVP), heart rate (HR), and skin temperature (TEMP). The EDA measures the electrical activity of human skin, which reflects the activity of sweat glands. Using spectral analyses of the blood vessels, the BVP detects physiological changes in cardiovascular activity.

Following the preprocessing steps outlined in [22], we first normalized the raw signals of each subject individually to account for differences in physiological responses due to age, gender, and personality, etc. In addition, the EDA signal could

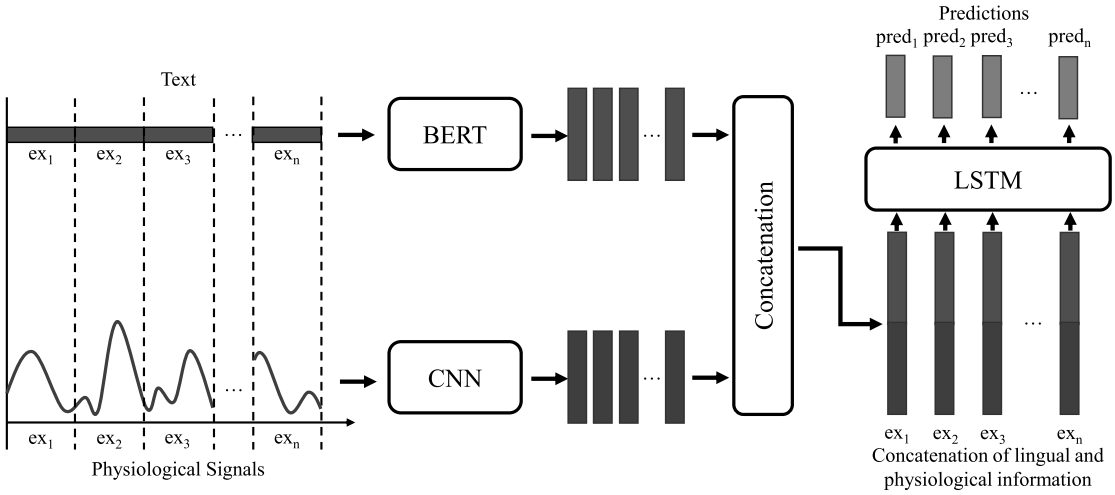

Figure 3.1: Our proposed overall architecture. In this architecture, exchange-level physiological signals and linguistic information are fed into two modality-specific encoder: Bert for text data and CNN for raw signals data. After retrieving the exchange-level representations, we concatenated them to get one feature vector for each exchange. The LSTM will be used as conversational-level predictor, which will be trained to learn the temporal relationship between exchanges.

be decomposed into two components: skin conductance level (SCL), also known as tonic, and skin conductance response (SCR), also known as phasic driver. As suggested in [44], we retained the SCR because it is considered the most reliable signal for determining an individual's response to a stimulus.

In addition, because the four signals were collected by sensors with different sampling frequencies, in order to synchronize various physiological signals, we re-sampled them all to the same frequency. Concretely, we interpolated the lower frequency signals based on the highest sampling frequency by using the nearest-neighbor method. Formally, the user physiological signals for exchange i can be represented by a vector $\mathbf{x}_i^{\alpha} \in R^{L_{raw}}$, where α is type of signal (e.g. EDA) and $L_{raw} = t_i \times f$, with t_i is the length of utterance and f is the selected sampling frequency. In contrast to other works such as [5, 6], we kept the preprocessing minimal and straightforward so as to better comprehend the effect of the proposed model on learning representations.

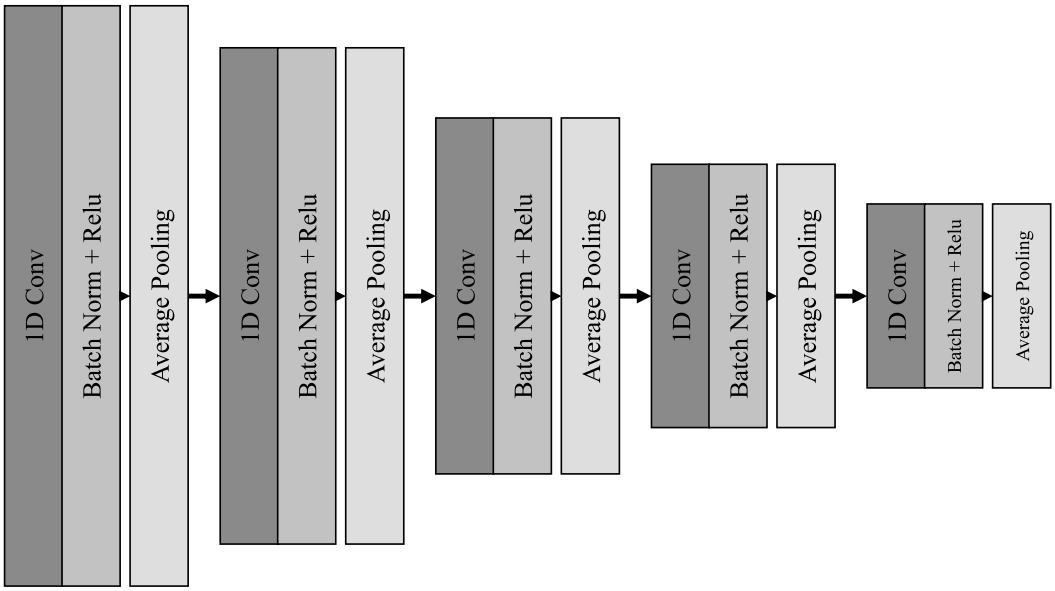

Figure 3.2: Our proposed CNNs architecture. The module contains sequence of 1D Convolution-BN-ReLU and Average Pooling

3.2 Convolutional Neural Networks for Physiological Signals

(1) Convolutional front-end

Our goal is to learn the end-to-end characteristics of physiological signals without any manually designed feature processing. Previous works [1][2][3] have demonstrated that CNNs are able to extract fine-grained representations from this type of low-level data, and perform well for many prediction tasks. Therefore, CNNs were used to extract features from physiological inputs. Specifically, we implemented a distinct CNNs block for each signal type. Inspired by [21], the convolutional front-end module contains sequence 1D convolution block, then average pooling (shown in Figure 3.2). Each block consists of 1D convolution layer, followed by a normalization layer and activation layer (e.g. ReLU). We set the out-channels produced by this module so that it will align with the linguistic embedding dimensions.

For every exchanges, we have four signals of length L_e. Since the conversation length of each exchange differs, zero padding the short signals and truncating the long ones was performed after the preprocessing pipeline. Next, CNNs modules transformed the each raw signals into latent representations of dimensions $L_t \times E$. Where L_t could be regarded as the number of tokens in one sentence, and E can be viewed as the size of the token's embedding.

(2) Multi-signal fusion
As emotions are subjective feelings produced by the complex coordination of multiple neurophysiological systems, [45] It is recommended that using multiple signals is better than using only one signal [21]. Thus, after transforming the signals using CNNs module in Section 3.2, we combined the 4 type of signals together and treated them as an entire physiological embedding that represent the hidden affective information. The embedding can be viewed as a vector:

$$P = [P_{bvp}, P_{eda_{SCR}}, P_{hr}, P_{temp}] \in R^{L_t \times (E*4)} \tag{3.1}$$

where $P_{bvp}, P_{eda_{SCR}}, P_{hr}$ and P_{temp} are feature representations of BVP, EDA, HR and TEMP extracted from the previous CNNs. Each feature embedding has dimension of $L_t \times E$, and we concatenated them over the embedding dimensions. In our study, the final dimensions are 64×768, as it corresponds to the linguistic embedding dimensions.

3.3 BERT Representations

In recent years, Bidirectional Encoder Representations from Transformers (BERT) has emerged as a state-of-the-art language representation model, demonstrating exceptional performance across various NLP tasks [14]. Language model pretraining plays a pivotal role in enhancing model performance, and Tohoku University has recently developed a pretrained Japanese BERT model, which has exhibited remarkable capabilities in sentiment analysis, particularly in tweet emotion recognition [46]. Leveraging this pretrained Tohoku BERT model, we adopted it as a fundamental component in our study.

For each dialogue exchange, the participant's and system's utterances were separated using a special token ([SEP]). Prior to BERT processing, utterance sequences underwent tokenization through MeCab and were further split into subwords using the WordPiece algorithm. This process enabled the representation of text in a format compatible with the BERT model.

We then utilized the activations from the second-to-last hidden layer of the BERT model, extracting valuable contextual information for each token in the sequences. By employing average pooling over these activations, we obtained a single vector of length 768, effectively summarizing the contextual information of the entire sequence [14]. This vector was designated as the input feature vector for each of the subsequent models employed in the study.

One significant advantage of utilizing BERT lies in its ability to eliminate the need for complex handcrafted feature extraction. By leveraging the rich contextual information encoded within the BERT model, we effectively bypassed the labor-intensive process of feature engineering, streamlining the overall pipeline. Fur-

thermore, the adoption of BERT facilitates seamless fusion with other modalities. This multimodal fusion capability holds great potential for enriching the understanding of emotions by incorporating information from various sources, such as audio, images, and video, in conjunction with textual data.

3.4 Inter-exchange multimodal feature modeling

We used LSTM [41] to learn the temporal relationships between multiple exchanges during conversation. Our overall architecture is depicted in the Figure 3.1. It consists o three phases: pre-processing, encoding, and time-fusion. Specifically, we represented each exchange in the dialogue as a sequence of input tokens. These tokens can include textual utterances, physiological signals, or a combination of both. Next, we used pretrained Bert model [14] to extract the embedding of text data and use CNNs to encode the physiological signals, as described in Section 3.2. Thus, for each exchange t, we got an output vector $e_t = [e_L, e_P]$, where we concatenated the embedding e_L of linguistic data and e_P of physiological data at exchange i over the embedding dimensions. Note that the Bert module is freezed during training as in other works. Therefore, a LSTM model based on sequence of multimodal representations $\boldsymbol{e}_t$ can be represented as

$$\begin{aligned} \begin{pmatrix} \boldsymbol{f}_t \\ \boldsymbol{g}_t \\ \boldsymbol{\iota}_t \\ \boldsymbol{o}_t \end{pmatrix} &= \begin{pmatrix} \sigma \\ \tanh \\ \sigma \\ \sigma \end{pmatrix} W \begin{pmatrix} \boldsymbol{e}_t \\ \boldsymbol{h}_{t-1} \end{pmatrix} \\ \boldsymbol{c}_t &= \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{g}_t \odot \boldsymbol{\iota}_t \\ \boldsymbol{h}_t &= \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \end{aligned} \tag{3.2}$$

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. In our settings, the time t corresponds to the number of exchanges used, could be regarded as context length. $\boldsymbol{e}_t$ is a latent vector represents the linguistic and physiological information of each exchange.

Finally, the output of the LSTM model is fed into a fully connected layer to predict the sentiment of the user at each exchange. The output of the fully connected layer is a vector of continuous values, which are then mapped to a range of $[1, 7]$ using the scaled sigmoid function. The output of this function is the predicted sentiment score of the user at each exchange. The loss function is the mean squared error between the predicted sentiment score and the ground truth

sentiment score. The overall loss function is defined as follows:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{3.3}$$

where N is the number of exchanges, $\hat{y}_i$ is the predicted sentiment score of the user at exchange i, and y_i is the ground truth sentiment score of the user at exchange i. Note that in our implementation, we used many-to-many LSTM model, which means that the LSTM model takes a sequence of multimodal representations as input and outputs a sequence of predicted sentiment scores. Therefore, when training, we calculated the loss function for every exchange in the input sequence by using all hidden states corresponding to the each exchange. This setting improves the training efficiency and allows the LSTM model to learn the underlying patterns and dependencies between the exchanges. However, when doing prediction, we only used the last hidden state to calculate the predicted sentiment score of the user at the last exchange. This is because the last hidden state contains the information of all previous exchanges, which is sufficient to predict the sentiment score of the user at the current time step.

In summary, our proposed approach processes and encodes the physiological data, then combines them with text representations, which will be represented as exchange feature embedding. These representations are fed into the LSTM model, allowing user sentiment changes during conversation to be captured.

Chapter 4

Experiments and Results

This chapter describes the experimental settings for the evaluation of our proposed model. We use Hazumi1911 dataset [40] to evaluate our proposed methods, and Section 4.1 summarizes the dataset. The evaluation procedure is described in Section 4.2 and section 4.3 describes the models used as baselines for comparison. Section 4.4 shows the implementation details. Finally, Section 4.5 presents the results of our experiments.

4.1 Dataset

In this study, we utilized the Hazumi1911 dataset [40], a multimodal human-agent dialog corpus, which is a publicity available for research purpose. The data collection process involved participants engaging in conversations with an agent, which operated using the Wizard of Oz method. For our experiment, we focused on data from 26 participants, resulting in a total of 2468 exchanges. The data annotations were obtained by having participants watch videos of themselves after the conversations and assign sentiment scores to each exchange. The sentiment scores ranged from 1 (indicating no enjoyment of the dialog) to 7 (indicating enjoyment of the dialog), and these scores were used as targets in our regression tasks.

Within the Hazumi1911 dataset, the participants' utterances were transcribed manually, providing us with textual data. To extract language representations, we employed BERT, following the approach described in Chapter 3. In addition to textual data, physiological signals were recorded using an Empatica E4 wristband developed by Empatica Inc. This wristband is non-intrusive and comfortable to wear, making it suitable for affective computing research. It has been widely used in previous studies, such as [47, 48, 49]. The E4 device recorded Electrodermal Activity (EDA), Blood Volume Pulse (BVP), Heart Rate (HR), and Skin Tem-

perature (TEMP) signals at respective frequencies of 4 Hz, 64 Hz, 1 Hz, and 4 Hz. To preprocess the time-series physiological signals, we followed the procedures outlined in Section 3.1.

4.2 Evaluation Procedure

For evaluation, we employed a leave-one-person-out cross-validation (LOPOCV) approach. In this method, the test data consisted of samples corresponding to each exchange between a participant and the dialog system, while the training data comprised the remaining samples from the other 25 participants. By excluding the test data of one participant from the training dataset, we ensured the prevention of leakage and overestimation. To assess the performance of our proposed approach, we calculated the mean absolute error (MAE) and Pearson correlation coefficient (Corr) for each evaluation. Specifically, we computed the MAE and Corr values for each participant using the LOPOCV method and reported the average values across all participants. To account for variability, we conducted the experiments three times with random initializations and averaged the evaluation values obtained from the three repetitions. By comparing the evaluation values among different models, we could assess the effectiveness of our proposed approach relative to other methods.

4.3 Baselines

This section describes the baseline models that were used for comparisons with our proposed method. The baselines include both single exchange and multiple exchange models, as we wish to demonstrate the efficacy of multiple exchanges.
Single exchange: We describe the models trained using only one exchange information, as first to show the effectiveness of the CNNs module.

(1) Text based Transformer (TR_T): As a baseline model, we utilized a conventional Transformer encoder from Bert model [13] for sentiment estimation, which solely relied on linguistic information. The model consists of a Bert encoder and using feed forward neural networks as regression head.

(2) Physiological signals CNNs (CNN_P): The CNNs architecture described in Section 3.2, which trained using only physiological data was used as one of our baselines to validate the effectiveness of end-to-end model.

(3) Time-series Physiological Transformer (TPTr): The TPTr was proposed in [5]. This model captures multimodal signals using crossmodal attention and achieves SOTA results in multimodal sentiment estimation. In contrast with normal cross attention model, TPTr model applies attention with linguistic and

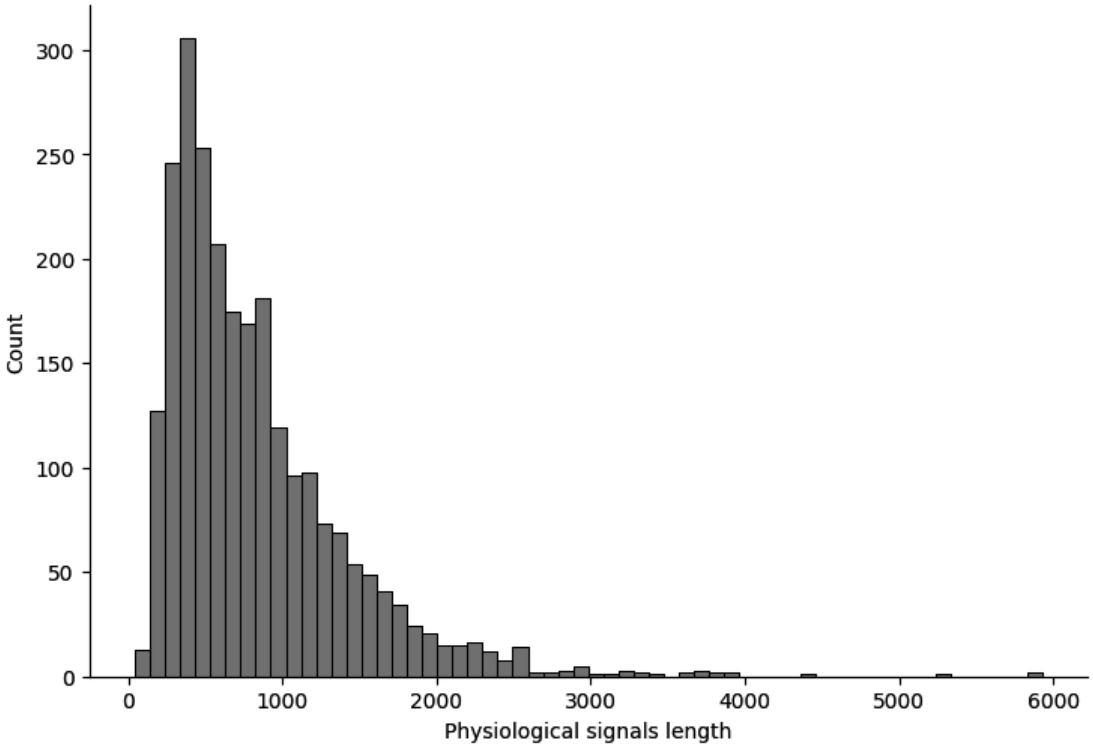

Figure 4.1: Histogram of physiological lengths. Most exchanges have signal length less than 2048.

physiological modalities and has a fixed attention direction, with the assumption that physiological will comprehend linguistic information. In this study, authors split the physiological signals into number of tokens, then take the average as the token physiological information.

(4) End-to-end TPTr (ETPTr - ours): The ETPTr and Transformer models have the same parameter settings as the TPTr. However, we applied CNNs as feature extractor for the physiological signals. As we believed that CNNs allow learning feature representation more effectively.

Multi exchange: This part describes baselines models for multi-exchange learning. All of the models listed below will accept sequence of exchanges information and their labels as training input, we referred number of exchanges as context length.

(1) Multi-exchanges Feed Forward Neural network (M-FFN) This model using a feed forward neural networks as regression head. After encoding each exchanges information, we concatenated the exchange embedding and fed to the FFN to predict the corresponding labels. This model settings are very minimal and we set it as our baseline for the multi-exchanges paradigm.

(2) Multi-exchanges TPTr (M-TPTr): In this model, we made the multimodal Transformer described previously applicable to multiple exchanges. Specifically, the embedding dimensions of multiple exchanges were concatenated and fed to the cross-attention modules.

(3) Multi-exchanges LSTM (M-LSTM): This is our proposed model, as described in the Section 3. We validated the model performance with context length of 3, 5 and 7, correspond to number of consecutive utterances in a conversation. However, we keep the training parameters the same in these studies, which caused some decrease in the performances, as described in the Section 4.5.3.

4.4 Implementation Details

We implemented our proposed system using the Pytorch framework. We used learning rate of 0.0001 and used Adam optimizer to optimize our models. We set dropout parameters to 0.1 to overcome overfitting. As described in Section 3.1, we interpolated all signals to 64Hz, in order to synchronize them. In addition, due to the varying lengths of utterances, we fixed the size of raw physiological to 2048 and zero-padded the short utterances. As shown in Figure 4.1, 2048 falls within 95 quantiles, so it will cover nearly all cases. The CNN modules include sequences of 5 1D convolution blocks with kernel size varies from 11 to 3, after each blocks, an average pooling layer will reduce the signal lengths by half. We set our final physiological embedding dimension is 768, as the same as text embedding dimension. Thus, the embedding dimension for each signals are $768/n_{signals}$. For example, if we use four signals, the dimension for each should be 192.

4.5 Results and Discussion

First, we demonstrate the efficacy of models based on our proposed CNNs module for extracting physiological features in Section 4.5.1. The models that used CNNs as feature extractors performed better than the previous ones, which used handcrafted feature engineering. Second, we studied the model performances under different sub-modalities in Section 4.5.2. Finally, to explore the effectiveness of the interplay between exchanges, various long-context models were evaluated. Their performances are shown in Section 4.5.3.

4.5.1 Performance of CNNs module

The results of the relf-reported sentiment prediction using single exchange are shown in Table 4.1. Our proposed architecture achieved better performance compared with all baseline approaches. Based on the results presented in the table,

we observe the performance of different models in terms of mean absolute error (MAE) and Pearson correlation coefficient (Corr) for sentiment estimation. The TRT model achieved an MAE of 1.067 and a Corr of 0.212, while our proposed CNNP model obtained an MAE of 1.094 and a Corr of 0.189. The TPTrT + P model demonstrated improved performance with an MAE of 1.056 and a Corr of 0.243. However, our proposed ETPTrT + P model outperformed all other models, achieving the lowest MAE of 1.049 and the highest Corr of 0.270. These results indicate the efficacy of our proposed approach in incorporating both linguistic information and physiological signals, resulting in improved sentiment estimation accuracy. And the utilization of convolutional neural networks (CNNs) for processing physiological signals demonstrates its effectiveness in capturing relevant features and enhancing the utilization of these signals for sentiment recognition. Overall, our proposed $\text{ETPTr}_{\text{T+P}}$ model showcases its superiority in accurately estimating sentiment in single exchange prediction.

Table 4.1: Result of sing-exchanges models, using EDA signals, to validate the effectiveness of the CNNs modules.

Model	MAE↓	Corr↑
TR_{T}	1.067	0.212
CNN_{P}	1.094	0.189
$\text{TPTr}_{\text{T+P}}$	1.056	0.243
$\text{ETPTr}_{\text{T+P}}$ (ours)	**1.049**	**0.270**

4.5.2 CNNs Based on Other Submodalities

In this study, we investigated the impact of different signal configurations on sentiment estimation. Specifically, we evaluated the models, which use CNNs as features extractor, with varying numbers of signals: 1 signal using BVP (Blood Volume Pulse), 2 signals combining BVP and EDA (Electrodermal Activity), and 3 signals incorporating BVP, EDA, and HR (Heart Rate) and all four signals, including Skin Temperature. The Table 4.2 presents the results obtained with different numbers of signals for two models: CNN_P and ETPTr_{L+P}. The CNN_P model, which contains only convolutional neural networks, was firstly tested. Among these configurations, the model achieved the lowest MAE and the highest Corr when using all four signals, with an MAE of 1.065 and a Corr of 0.154. On the other hand, the ETPTr_{L+P} model, which combines linguistic and physiological signals, also exhibited different performance across varying signal numbers. Notably, the model achieved the best results with a single signal, with MAE of 1.049

and Corr of 0.2697. However, as the number of signals increased, the performance of the model showed slight variations. These results highlight the importance of considering the number of signals in exchange-level sentiment estimation.

Table 4.2: The evaluation on submodalities. For 1 signal, we use BVP; for 2 signals, we add EDA; we add Hr as 3rd signal and all four signals in the final.

Model	Number of signals	MAE↓	Corr↑
CNN_P	BVP	1.083	0.130
	BVP+EDA	1.082	0.147
	BVP+EDA+Hr	1.085	0.147
	BVP+EDA+Hr+Temp	**1.065**	**0.154**
$ETPTr_{L+P}$	BVP	**1.049**	**0.2697**
	BVP+EDA	1.050	0.2064
	BVP+EDA+Hr	1.052	0.2454
	BVP+EDA+Hr+Temp	1.107	0.2244

4.5.3 Analysis of the Context Length

In this section, we evaluate the performance of architectures with CNNs as features extractor using text with four physiological signals: EDA, BVP, HR, and TEMP, with varying context lengths. The results are shown in Table 4.3. For the baseline M-FFN model, which utilizes a feedforward neural network architecture, we got MAE of 1.064 and Corr of 0.268. In contrast, the M-TPTr model, based on the Transformer architecture, showed a decrease in performance. With a same context length of 3, the model obtained a MAE of 1.137 and Corr of 0.216. This suggests that adapting the Transformer for longer context length make it more challenging for the model to capture the temporal dependencies effectively.

The M-LSTM models, which employ Long Short-Term Memory networks, demonstrated a different trend. As the context length increased from 3 to 5, there was minor decrease in the MAE but increase in the correlation score. The M-LSTM3 achieved a MAE of 1.019 and a correlation coefficient of 0.285, while the M-LSTM5 achieved a slightly higher MAE of 1.028 but a higher correlation coefficient of 0.295. However, further increasing the context length to 7 led to a slight decrease in performance. We suspect that this phenomenon is due to the training hyper-parameters since we only optimized the hyper-parameters for the context length of 3.

In summary, the choice of sequence length had varying effects on the performance of different sentiment estimation models. But our proposed M-LSTM

Table 4.3: Result of models with varying context length. We evaluated the context lengths of 3, 5 and 7.

Model	MAE↓	Corr↑
M-FFN$_3$	1.064	0.268
M-TPTr$_3$	1.137	0.216
M-LSTM$_3$	**1.019**	0.285
M-LSTM$_5$	1.028	**0.295**
M-LSTM$_7$	1.070	0.256

model got the best performance, which is close to the human performance (MAE of 1.008), indicating the importance of capturing additional context for sentiment estimation.

Chapter 5

Conclusion and Future Works

5.1 Conclusion

In conclusion, this study addresses the challenges of accurately recognizing and adapting to a user's real-time emotional state in dialog systems. By incorporating physiological signals and inter-exchange information, our proposed approach enhances sentiment estimation and adaptation capabilities. We highlight the limitations of relying solely on linguistic cues, as emotional masking can hinder the accurate identification of a user's emotional state. To overcome this challenge, we leverage alternative sources of information, such as physiological signals and facial expressions, to gain deeper insights into hidden emotional nuances. By integrating inter-exchange information, our approach captures the dynamics and evolving sentiment across multiple exchanges, enabling more precise and adaptive interactions. We introduce the use of convolutional neural networks (CNNs) to learn robust representations directly from physiological signals, eliminating the need for hand-designed features and enhancing the utilization of these signals for sentiment adaptation. Overall, our contributions lie in the effective integration of physiological signals and inter-exchange information, paving the way for more accurate and comprehensive sentiment estimation in dialog systems.

5.2 Limitations and Future Works

One notable limitation of our study is the absence of publicly available datasets that combine exchange-level self-sentiment labels, linguistic information, and physiological signals, except for the Hazumi dataset used in our research. As a result, we were unable to evaluate our proposed model using different datasets, which is an avenue for future exploration. Moving forward, it would be beneficial to investigate effective strategies for integrating time-series audiovisual signals into

our proposed approach, encompassing all four modalities. Moreover, to address the limited training data for the convolutional neural networks (CNNs) employed in physiological signal processing, future work could involve pretraining the CNNs on a larger dataset to acquire more comprehensive and generalizable physiological representations. Additionally, exploring self-supervised learning frameworks [50] could enhance the robustness of exchange embeddings across diverse participants, thereby bolstering the overall performance of our proposed model. These avenues of investigation will contribute to advancing multimodal sentiment analysis and expanding its applicability in real-world scenarios.

Bibliography

[1] John D. Mayer and Peter Salovey. The intelligence of emotional intelligence. *Intelligence*, 17(4):433–442, 1993.

[2] Tin Lay Nwe, Say Wei Foo, and Liyanage C De Silva. Speech emotion recognition using hidden markov models, Nov 2003.

[3] Zhanna Sarsenbayeva, Gabriele Marini, Niels van Berkel, Chu Luo, Weiwei Jiang, Kangning Yang, Greg Wadley, Tilman Dingler, Vassilis Kostakos, and Jorge Goncalves. Does smartphone use drive our emotions or vice versa? a causal analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–15, New York, NY, USA, 2020. Association for Computing Machinery.

[4] Kangning Yang, Chaofan Wang, Zhanna Sarsenbayeva, Benjamin Tag, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Benchmarking commercial emotion detection systems using realistic distortions of facial image datasets, Jun 2020.

[5] Shun Katada, Shogo Okada, and Kazunori Komatani. Transformer-based physiological feature learning for multimodal analysis of self-reported sentiment. In *Proceedings of the 2022 International Conference on Multimodal Interaction*, ICMI '22, page 349–358, New York, NY, USA, 2022. Association for Computing Machinery.

[6] Shun Katada, Shogo Okada, Yuki Hirano, and Kazunori Komatani. Is she truly enjoying the conversation? analysis of physiological signals toward adaptive dialogue systems. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, page 315–323, New York, NY, USA, 2020. Association for Computing Machinery.

[7] Fanny Larradet, Radoslaw Niewiadomski, Giacinto Barresi, Darwin G. Caldwell, and Leonardo S. Mattos. Toward emotion recognition from physiological signals in the wild: Approaching the methodological issues in real-life data collection, Jul 2020.

[8] Lin Shu, Jinyan Xie, Mingyue Yang, Ziyi Li, Zhenqi Li, Dan Liao, Xiangmin Xu, and Xinyi Yang. A review of emotion recognition using physiological signals, Jun 2018.

[9] Shun Katada, Shogo Okada, and Kazunori Komatani. Effects of physiological signals in different types of multimodal sentiment estimation. In *IEEE Transactions on Affective Computing*, Los Alamitos, CA, USA, 2022. IEEE Computer Society.

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

[11] Yoon Kim. Convolutional neural networks for sentence classification, 2014.

[12] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

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

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[15] Ye-Yi Wang, Li Deng, and Alex Acero. Spoken language understanding. *IEEE Signal Processing Magazine*, 22(5):16–31, 2005.

[16] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions, Mar 2023.

[17] Bobo Zhao, Zhu Wang, Zhiwen Yu, and Bin Guo. Emotionsense: Emotion recognition based on wearable wristband. In *2018 IEEE*

SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI), pages 346–355, 2018.

[18] Tianyuan Xu, Ruixiang Yin, Lin Shu, and Xiangmin Xu. Emotion recognition using frontal eeg in vr affective scenes, May 2019.

[19] Elaine Sedenberg and John Chuang. Smile for the camera: Privacy and policy implications of emotion ai, 2017.

[20] Yi Wang, Zhiyi Huang, Brendan McCane, and Phoebe Neo. Emotionet: A 3-d convolutional neural network for eeg-based emotion recognition. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7, 2018.

[21] Kangning Yang, Benjamin Tag, Yue Gu, Chaofan Wang, Tilman Dingler, Greg Wadley, and Jorge Goncalves. Mobile emotion recognition via multiple physiological signals using convolution-augmented transformer. In *Proceedings of the 2022 International Conference on Multimedia Retrieval*, ICMR '22, page 562–570, New York, NY, USA, 2022. Association for Computing Machinery.

[22] M. Sami Zitouni, Cheul Young Park, Uichin Lee, Leontios Hadjileontiadis, and Ahsan Khandoker. Arousal-valence classification from peripheral physiological signals using long short-term memory networks. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 686–689, 2021.

[23] Emotions revealed.

[24] Jeffrey F. Cohn. Foundations of human computing. In *Proceedings of the 8th international conference on Multimodal interfaces*. ACM, November 2006.

[25] K Scherer. Vocal communication of emotion: A review of research paradigms. *Speech Communication*, 40(1-2):227–256, April 2003.

[26] Maja Pantic, Anton Nijholt, Alex Pentland, and Thomas S. Huanag. Human-centred intelligent human computer interaction: how far are we from attaining it? *International Journal of Autonomous and Adaptive Communications Systems*, 1(2):168, 2008.

[27] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, November 2009.

[28] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias. Emotion analysis in man-machine interaction systems. In *Machine Learning for Multimodal Interaction*, pages 318–328. Springer Berlin Heidelberg, 2005.

[29] Koji Inoue, Divesh Lala, Katsuya Takanashi, and Tatsuya Kawahara. Latent character model for engagement recognition based on multimodal behaviors. In *Lecture Notes in Electrical Engineering*, pages 119–130. Springer Singapore, 2019.

[30] Chloé Clavel and Zoraida Callejas. Sentiment analysis: From opinion mining to human-agent interaction. *IEEE Transactions on Affective Computing*, 7(1):74–93, 2016.

[31] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[32] Haifeng Chen, Dongmei Jiang, and Hichem Sahli. Transformer encoder with multi-modal multi-head attention for continuous affect recognition. *IEEE Transactions on Multimedia*, 23:4171–4183, 2021.

[33] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, AmirAli Bagher Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pretrained transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2359–2369, Online, July 2020. Association for Computational Linguistics.

[34] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM International Conference on Multimedia*, page 1122–1131, New York, NY, USA, 2020. Association for Computing Machinery.

[35] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy, July 2019. Association for Computational Linguistics.

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

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

[38] Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. AMIGOS: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12(2):479–493, 2021.

[39] Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. 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)*, pages 1–8, Los Alamitos, CA, USA, 2013. IEEE, IEEE Computer Society.

[40] Kazunori Komatani and Shogo Okada. Multimodal human-agent dialogue corpus with annotations at utterance and dialogue levels. In *2021 9th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8, Los Alamitos, CA, USA, 2021. IEEE Computer Society.

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

[42] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014.

[43] M. Schuster and K.K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[44] Roberto Sánchez-Reolid, Francisco López de la Rosa, María T. López, and Antonio Fernández-Caballero. One-dimensional convolutional neural networks for low/high arousal classification from electrodermal activity, Jan 2022.

[45] JONATHAN POSNER, JAMES A. RUSSELL, and BRADLEY S. PETERSON. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology, Sep 2005.

[46] Tatsuki Akahori, Kohji Dohsaka, Masaki Ishii, and Hidekatsu Ito. Efficient creation of japanese tweet emotion dataset using sentence-final expressions. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*. IEEE, March 2021.

[47] Heath Yates, Brent Chamberlain, Greg Norman, and William H. Hsu. Arousal detection for biometric data in built environments using machine learning. In Neil Lawrence and Mark Reid, editors, *Proceedings of IJCAI 2017 Workshop on Artificial Intelligence in Affective Computing*, volume 66 of *Proceedings of Machine Learning Research*, pages 58–72, Cambridge, MA, USA, 20 Aug 2017. PMLR.

[48] Marco Maier, Daniel Elsner, Chadly Marouane, Meike Zehnle, and Christoph Fuchs. 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*, pages 1415–1421, Palo Alto, CA, USA, 7 2019. International Joint Conferences on Artificial Intelligence Organization.

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

[50] Pritam Sarkar and Ali Etemad. Self-supervised ECG representation learning for emotion recognition. *IEEE Transactions on Affective Computing*, 13(3):1541–1554, jul 2022.