國立臺灣大學管理學院資訊管理學研究所

碩士論文

Department of Information Management
College of Management
National Taiwan University
Master Thesis

基於多任務與遷移學習的對話情感預測

Predict Before You Speak: Sentiment Forecasting in Dialogue with Multi-task and Transfer Learning

郭宇雋

Yu-Jun Kuo

指導教授：魏志平 博士

Advisor: Chih-Ping Wei, Ph.D.

中華民國 112 年 7 月

July 2023

doi:10.6342/NTU202304293

誌謝

兩年的研究所生涯匆匆而逝，此刻臨近旅程尾聲，在這裡我想要由衷地表達我的感謝。首先，我想對我的指導老師魏志平教授致上最衷心的感謝。您淵博的學識與清晰的口條，總能激發我在學術研究中找到更多靈感，並培養我以不同的視角看待事物的能力。而您不僅耐心地給予我學術指導，您分享的人生經驗更是深深影響著我。再次感謝您給予我的無私支持與鼓勵，再多的言語都道不盡我對您的感謝。

而 BI Lab 的小夥伴們，我對你們的感謝大概就像我們一起去唱歌的次數一樣，多到數不清了！在研究所這兩年的時光裡，能有一群這麼厲害又興趣相投的小夥伴們，我真的感到很幸福。從碩一時一起替實習奮鬥，到碩二時一起埋首跑實驗，最後一起寫論文準備口試，當然還有飲料小隊與唱歌軍團，這些回憶將永遠在我生命中散發歡樂的光芒。我也想感謝 OP Lab 麻吉們，在最後準備口試水深火熱時，你們的陪伴替我的生活增添了許多色彩，你們是一群可愛又有趣的人，能夠認識你們我真的很開心！

此外，我也要感謝我的家人。儘管你們並不熟悉我的研究領域，但每次仍然都很認真地聽我分享我的研究，努力關心我的近況並給予我建議，你們的陪伴與關心是我研究所生活中不可或缺的養分，感謝你們一直在背後支持著我，這份情感我永遠銘記在心。我還想感謝身邊所有一路陪伴我成長的好朋友們，以及在我研究所生涯中，所有給予過我幫助、關心與支持的同學朋友、學長姐、學弟妹以及研究助理。是你們的存在，才讓我能夠走到今天這一刻。

最後，我也想感謝這兩年來的自己。感謝自己努力克服所有困難與挑戰，勇敢面對所有失敗與恐懼。研究所充滿成長的時光對我來説意義非凡，也將成爲我人生中很珍貴的回憶。謝謝大家，我愛你們！

郭宇雋 謹誌
于國立臺灣大學資訊管理研究所
中華民國一一二年七月

摘要

情感狀態對於人類的行爲、動機和決策具有重要影響，因此旨在模擬或預測人類反應的對話系統必須仔細考慮情感這個因素。爲了優化對話系統的對話體驗與使用者滿意度，進行使用者情感狀態的預測至關重要。目前的研究主要集中於識別現有對話中的情感狀態，忽略了對即將到來的情感狀態進行主動預測的重要性。然而，若能主動預測對話中即將到來的情感狀態，就能使對話系統有能力事前主動調整將給予使用者的回覆。

因此在本研究中，我們專注於探討對話情感預測這個任務，並提出了一個多任務學習模型，將歷史對話情感識別、歷史對話行爲識別，以及未來對話行爲預測作爲輔助任務，並發展一個新穎的機制讓模型在訓練過程中動態調整不同任務之間的重要性。實驗證實了我們的多任務學習模型能有效地捕獲更多面向的情感相關資訊，並讓模型能夠學習到更好的情感特徵表示，從而提高了情感預測任務的表現，並在整體的準確率上優於當今表現最好的方法。

此外，爲了能更貼近現實應用，我們也創建了一個基於常見對話系統場景的全新對話資料集，並在此資料集上進行了領域遷移實驗，最後也驗證了我們提出的領域遷移方法的有效性。我們的研究強調了多任務學習和領域遷移學習在情感預測任務中的重要性，也爲開發更複雜的情感分析技術提供了基礎，以提升對話系統中的情感理解能力並改善使用者體驗。

關鍵字: 對話系統、情感分析、對話情感預測、多任務學習、遷移學習、領域自適應

doi:10.6342/NTU202304293

Abstract

Affective states profoundly influence human behaviors, motivations, and decisions, making them a crucial factor to consider in dialogue systems aimed at simulating or predicting human reactions. To improve the conversational experience and user satisfaction in dialogue systems, prediction of users' affective states is essential. Existing research primarily focuses on recognizing affective states within dialogue history, neglecting the proactive forecasting of upcoming affective states. However, the ability to forecast upcoming affective states proactively can enable dialogue systems to adjust responses in advance.

Therefore, in this research, we concentrate on the task of Sentiment Forecasting in Dialogue and propose a multi-task learning model by incorporating sentiment recognition and dialogue act recognition within dialogue history sequence and upcoming dialogue act forecasting as auxiliary tasks. We also develop a novel mechanism to dynamically adjust the importance of each task during training. Experimental results demonstrate the effectiveness of our model in capturing diverse sentiment-related information and learning better sentiment representations, leading to improved sentiment forecasting performance, surpassing existing state-of-the-art methods.

doi:10.6342/NTU202304293

Additionally, to enhance real-world applicability, we collect a new dialogue dataset simulating common dialogue scenarios and conduct domain transfer experiments, further validating the efficacy of our proposed domain transfer methods. Our research emphasizes the significance of multi-task learning and domain transfer in sentiment forecasting tasks, providing a foundation for developing more sophisticated sentiment analysis techniques, improving sentiment understanding in dialogue systems, and enhancing user experiences.

Keywords: Dialogue System, Sentiment Analysis, Sentiment Forecasting in Dialogue, Multi-task Learning, Transfer Learning, Domain Adaptation

doi:10.6342/NTU202304293

Table of Contents

doi:10.6342/NTU202304293

doi:10.6342/NTU202304293

List of Figures

doi:10.6342/NTU202304293

List of Tables

doi:10.6342/NTU202304293

Chapter 1

Introduction

1.1 Background

Affective states have a crucial impact on humans' behaviors, motivations, and decisions. Therefore, when a system aims to simulate or predict human reactions, this factor needs to be carefully considered (Salmeron, 2012). In essence, in the process of human-machine interaction, having affective intelligence is also considered an important component to improve the process (Wang et al., 2021). Given the significance of affective states in the realm of human-machine interaction, and considering the growing popularity of dialogue systems that aid users in various daily activities (Peng et al., 2020), numerous researchers in the field of Natural Language Processing (NLP) have embarked on investigations into the presence of affective states within dialogues.

We can delve deeper into examining the impact of affective states on the functionality of dialogue systems in the context of human-computer interaction. To illustrate this, let us consider a ubiquitous scenario of real-time chat customer service. In this particular setting, our aim is to enhance the conversational experiences of customers. Notably, this scenario aligns with the one described in Guibon et al. (2021).

doi:10.6342/NTU202304293

In such situations, the primary concern in numerous studies revolves around comprehending user intent more effectively and generating responses that are both fluent and semantically rich. As a result, extensive research efforts have focused on enhancing dialogue understanding and generation tasks. A prevalent approach involves leveraging pre-trained conversation models (Peng et al., 2020; He et al., 2022) or even employing few-shot learning methods to deal with real-world data (Peng et al., 2020).

However, even if the response effectively conveys information and is linguistically fluent enough to capture users' attention, disregarding the users' affective feelings towards the response can result in a subpar user experience (Figure 1.1). Therefore, early prediction of the user's affective state and its evolution within the dialogue flow can assist the system in delivering more suitable responses in advance, thereby enhancing user satisfaction (Shahriar and Kim, 2019).

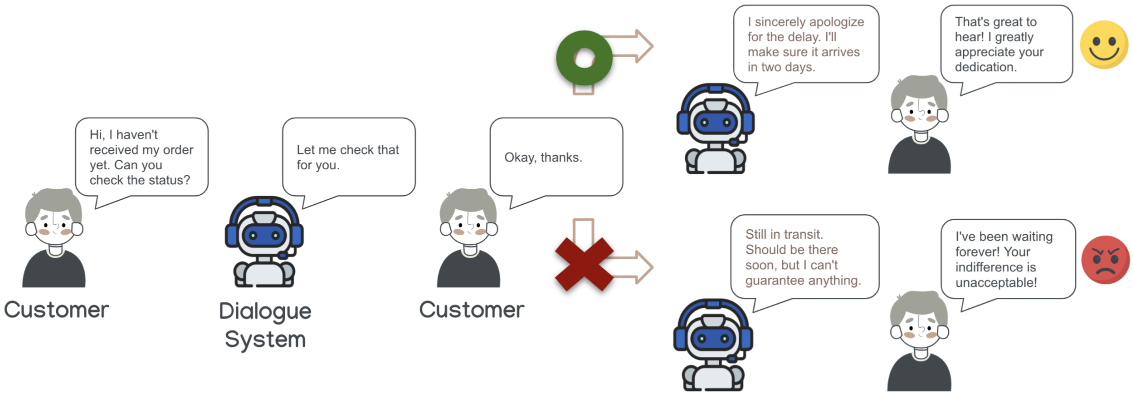

Hi, I haven't received my order yet. Can you check the status?
Customer
Let me check that for you.
Dialogue System
Okay, thanks.
Customer
I sincerely apologize for the delay. I'll make sure it arrives in two days.
That's great to hear! I greatly appreciate your dedication.
Still in transit. Should be there soon, but I can't guarantee anything.
I've been waiting forever! Your indifference is unacceptable!

Figure 1.1: *Scenario of A Dialogue System*

In order to facilitate advancements in research on affective dialogue, several studies have introduced new benchmark datasets. One notable example is the **Interactive Emotional Dyadic Motion Capture (IEMOCAP)** dataset introduced by Busso et al. (2008), recognizing that affective states can be conveyed through verbal and non-verbal channels, such as facial expressions and hand gestures. This dataset comprises

doi:10.6342/NTU202304293

approximately 10+ hours of dyadic dialogues, including videos, transcriptions, and specific labels of affective states. Its primary objective is to promote research in the fields of multi-modal communication and human expressive interaction, making it a commonly employed resource for affective-related tasks.

Recently, Chen et al. (2022) constructed a comprehensive benchmark dataset called **Chinese Personalized and Emotional Dialogue dataset (CPED)** for conversational AI. This dataset comprises over 12,000 dialogues extracted from around 400 speakers and 40 TV shows, incorporating textual, audio, and video features. Notably, CPED takes into account not only speakers' personalities and affective states, but also factors such as age, gender, dialogue acts (i.e., intent in the utterance, such as greeting or question), and scenes. By considering these comprehensive factors, CPED presents new avenues for research related to affective dialogue.

1.2 Research Motivation

Due to the significant importance of affective states in dialogue systems, a considerable amount of research has been conducted in this area. Early studies primarily centered around the task of recognizing affective states within dialogues, commonly referred to as **Emotion Recognition in Conversation (ERC)** (Kim and Kim, 2018; Wang et al., 2021; Guibon et al., 2021; Saha et al., 2021). Specifically, emotion recognition involves recognizing the potential current affective state upon encountering a specific utterance from a speaker. In essence, the question at hand is: *"What do you believe is the affective state of this utterance?"* However, given their access to the content of the target utterance, this task is relatively straightforward and has limited practicality. In

doi:10.6342/NTU202304293

many scenarios, it becomes necessary for the dialogue system to predict or forecast in advance the type of affective state a particular machine-generated utterance may evoke in the user, in order to proactively adjust and rectify the candidate machine-generated utterance beforehand.

Therefore, some prior studies have proposed a new task known as **Emotion Forecasting in Dialogue** (Shahriar and Kim, 2019; Shi et al., 2020; Abouzeid et al., 2021; Zou et al., 2022), which is regarded as an emerging and promising field of research. According to Shahriar and Kim (2019), this task involves unique problem formulations that differ from the traditional emotion recognition task. The distinction between these two tasks is illustrated in Figure 1.2. Emotion forecasting aims to forecast the speaker's future affective state based on past cues. In other words, it requires forecasting the affective state of an upcoming utterance in the dialogue before the content of the utterance is revealed. In essence, the question at hand is: *"What do you believe will be the affective state present in the upcoming utterance?"* Due to the absence of information about the content of the upcoming utterance, this task is relatively challenging, but has shown potential applications in various fields in recent years (Salmeron, 2012; Shahriar and Kim, 2019). By having this capability, the dialogue system can select more empathetic responses, and can also steer the dialogue towards a desired affective state, resulting in the user's affective state aligning more closely with expectations.

However, there has still been limited research conducted on this particular task. Additionally, the existing studies tend to solely focus on utilizing dialogue context as features, without delving into further feature exploration, more effective auxiliary tasks, or model architecture design. To the best of our knowledge, no research has specifically

doi:10.6342/NTU202304293

addressed the incorporation of affective state and dialogue act features within the context sequence, nor considered the potential dialogue act feature of the upcoming utterance. Nevertheless, it is evident that these features have a substantial impact on the affective dynamics of speakers within a dialogue (Cao et al., 2021; Chen et al., 2022). For instance, certain dialogue acts may inherently exhibit specific affective states, and there may exist continuity or causal relationships among multiple dialogue acts. Additionally, the flow of affective states and dialogue acts within a dialogue may adhere to specific patterns, which are intuitively evident. Therefore, if we can model the information of dialogue act sequence or affective state sequence, it can lead to a more profound understanding of the dynamics and progression of the whole dialogue. Furthermore, the ability to forecast upcoming dialogue act may be intricately linked to forecasting upcoming affective state. Hence, the modeling of dialogue act and affective state sequence is considered crucial for effectively forecasting the affective state of an upcoming utterance.

Additionally, previous studies have predominantly relied on RNN-based modules for encoding contextual features, disregarding the potential advantages provided by attention mechanism (Vaswani et al., 2017). Moreover, the benchmark datasets commonly used in these studies are primarily derived from TV dramas or movies, where dialogues are constrained by predefined flows and speakers' roles, and the scenarios also differ from practical applications such as customer service, resulting in limited real-world applicability.

doi:10.6342/NTU202304293

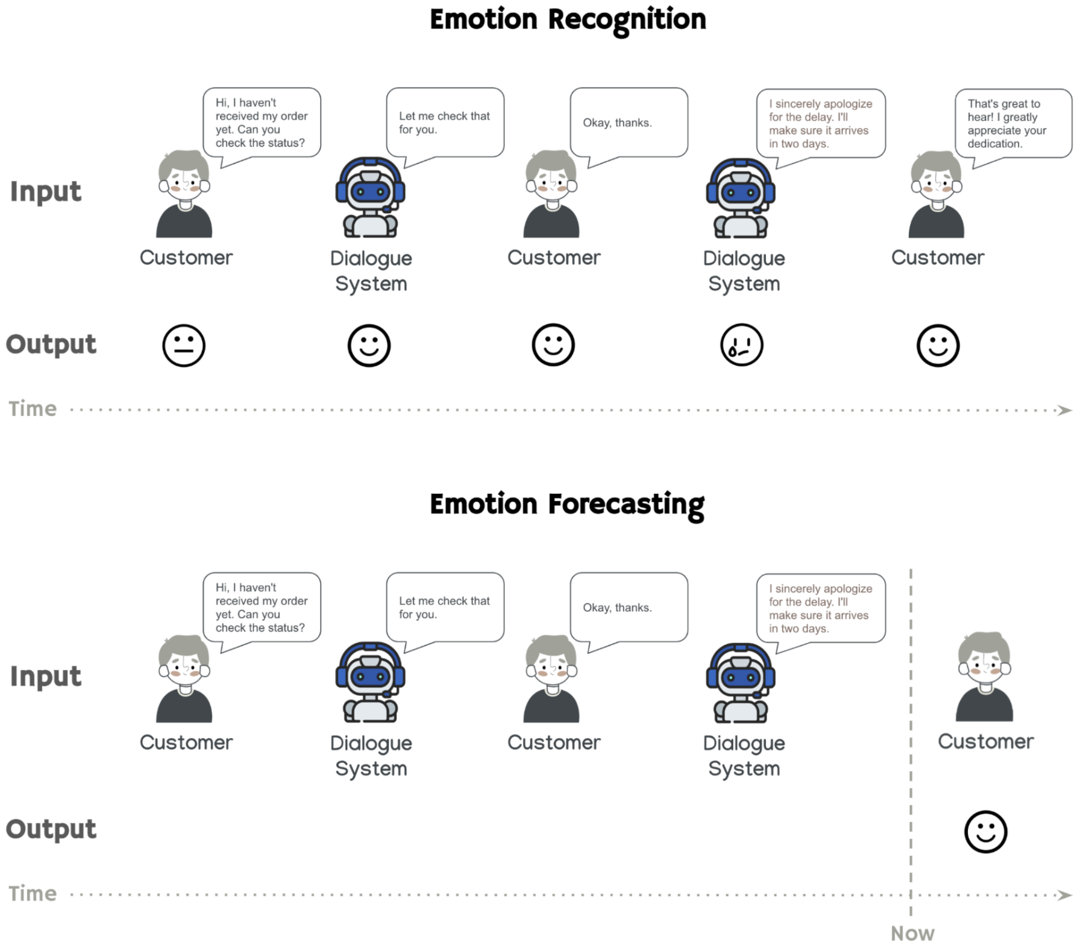

Emotion Recognition
Input
Hi, I haven't received my order yet. Can you check the status?
Let me check that for you.
Okay, thanks.
I sincerely apologize for the delay. I'll make sure it arrives in two days.
That's great to hear! I greatly appreciate your dedication.
Customer
Dialogue System
Customer
Dialogue System
Customer
Output
Time
Emotion Forecasting
Input
Hi, I haven't received my order yet. Can you check the status?
Let me check that for you.
Okay, thanks.
I sincerely apologize for the delay. I'll make sure it arrives in two days.
Customer
Dialogue System
Customer
Dialogue System
Customer
Output
Time
Now

Figure 1.2: *Illustrations of Emotion Recognition Task vs. Emotion Forecasting Task*

doi:10.6342/NTU202304293

1.3 Research Objective

Considering the limitations of existing emotion forecasting studies, our primary objective is to improve the forecasting effectiveness of this challenging task. Moreover, we intend to introduce a new dataset that closely simulates real-world applications. Finally, we will leverage transfer learning techniques to effectively transfer the knowledge acquired by the model from the public dataset to our dataset's new domain. This approach will enable the model to be applicable in practical real-world scenarios.

However, how can we improve the performance of this challenging task? First, it is necessary to investigate additional features that may impact the affective state of the upcoming utterance within a dialogue. Wen et al. (2021) discovered that emotion transitions are influenced by both the dialogue context and specific personality traits. Furthermore, Chen et al. (2022) stated that in addition to personality traits, factors such as gender, age, dialogue act (DA), and scene also exert influence on dialogues. This causal relationship between emotions and dialogue acts (DA) has been further confirmed by Cao et al. (2021) through qualitative and quantitative research. Considering the potential value of these features, it is worthwhile to explore how sentiment and dialogue act (DA) in the dialogue context impact the affective state of the upcoming utterance within a dialogue.

Next, an appropriate model architecture design is essential to incorporate these significant features effectively. In the past, many studies have applied the Multi-Task Learning (MTL) architecture to tackle such situations. This approach enables the model to learn multiple tasks simultaneously and allows certain parameters or representations to be shared among related tasks. This design fosters a complementary effect and facilitates knowledge sharing between tasks, as demonstrated by its superior performance in various

doi:10.6342/NTU202304293

dialogue-related experiments (Kim and Kim, 2018; Saha et al., 2021; Zou et al., 2022) compared to learning a single task. Therefore, we decide to adopt and enhance this multi-task learning framework to effectively incorporate the identified features. We firmly believe that addressing these issues will significantly improve the overall effectiveness of the emotion forecasting task.

doi:10.6342/NTU202304293

Chapter 2

Literature Review

2.1 Affective State Definition

How should the affective state in dialogue be formally defined? According to Naar (2018), there is a distinction in everyday language between affective states predicated for a relatively short time and relatively long time. When an affective state lasts for a short time, it is called ***emotion***. If it lasts for a long time, it is called ***sentiment***. Therefore, affective states in dialogue can be divided into two types, emotion and sentiment, which differ in terms of their duration.

Despite the difference between emotion and sentiment, there is still an internal connection. Liu (2020) has stated that sentiment can be understood as the underlying positive or negative feeling, attitude, or evaluation associated with an opinion. Based on the definition, the terms emotion and sentiment are actually interconnected, as the sentiment seems to be a general state or underlying feeling of the emotion (Adam, 2019). In other words, sentiment can be described as a coarse-grained state of emotion. Figure 2.1 illustrates the difference.

In the realm of affective dialogue research, diverse approaches have been employed

doi:10.6342/NTU202304293

to label ***emotion*** across different studies. For instance, Busso et al. (2008) used two different types of label systems: a continuous label in the valence-arousal dimension and a 9-category label scheme, while Chen et al. (2022) labeled emotion with 13 categories. However, when it comes to ***sentiment*** labeling, there is a notable consensus among researchers. Almost all studies adhere to a consistent definition of sentiment, categorizing it into three distinct polarities: negative, neutral, and positive. Consistent with this prevailing trend, our research will use ***sentiment*** as our target label, and also adopt the same three types of sentimental polarity as the basis for our target labeling.

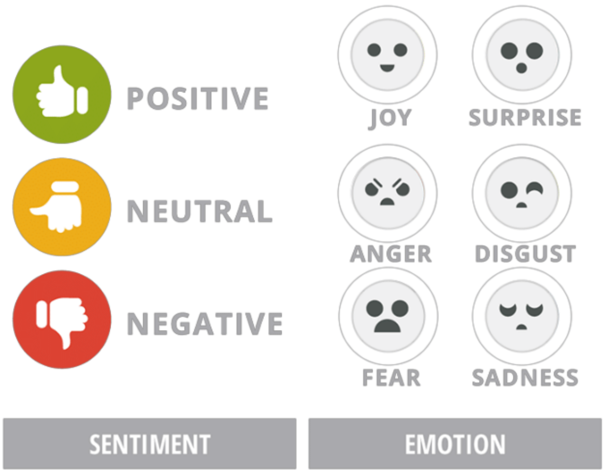


Figure 2.1: *Difference between Emotion and Sentiment (Adopted from Adam (2019))*

2.2 Sentiment Forecasting in Dialogue

Figure 2.2 illustrates the overall architecture employed in prior research on Sentiment Forecasting in Dialogue, as well as Emotion Forecasting in Dialogue. This architecture encompasses various models utilized in prior studies.

To begin, the **Utterance Encoder**, alternatively referred to as the **Sentence Encoder**, is employed to acquire the representation of each individual utterance present in the dialogue. Previous investigations commonly employed Convolutional Neural Networks

doi:10.6342/NTU202304293

(CNNs) (LeCun et al., 1998) or Recurrent Neural Networks (RNNs) and their variants, such as Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) or Gated Recurrent Unit (GRU) (Chung et al., 2014), as the Utterance Encoder to obtain the fundamental utterance representation. Recent studies have also incorporated pre-trained language models such as BERT (Devlin et al., 2018) to enhance the quality of the utterance representation.

Once the utterance representations are obtained, the next step involves employing a **Context Encoder** to capture the interdependent contextual information among utterances in the dialogue. Previous studies commonly implement the Context Encoder as two components: the **Pair-wise Encoder** and the **Sequence-wise Encoder**. The Pair-wise Encoder focuses on extracting information from utterances spoken by the same speaker. Its role is to capture the intra-speaker dependencies within the dialogue. On the other hand, the Sequence-wise Encoder is responsible for extracting information from the sequence of all utterances spoken by both speakers in the dialogue. Its role is to capture the inter-speaker dependencies within the dialogue. By incorporating both the Pair-wise and Sequence-wise encoders, this architecture can effectively model both the local internal influence of each speaker and the global external influence of all speakers involved in the dialogue. To construct the **Context Encoder**, these studies typically employ the attention mechanism. The attention mechanism has demonstrated remarkable performance in modeling contextual information across various dialogue-related scenarios. By leveraging attention, the Context Encoder can effectively capture the relevant dependencies and relationships between utterances, enabling a comprehensive representation of the dialogue context.

Once the utterance representations are passed through the Context Encoder, we obtain two essential representations: the local contextual representation and the global contextual representation. Previous studies have commonly employed an attention-based **Fusion Module** to merge these representations effectively. The purpose of this module is to combine the local and global contextual information in a meaningful way. Following the fusion of contextual representations, the next step involves utilizing a **Classification Module** comprised of fully-connected layers. This module is responsible for forecasting the sentiment polarity of the upcoming utterance. By leveraging the merged representations, the classification module can make informed predictions regarding the sentiment in the upcoming utterance.

This baseline architecture serves as a foundation for integrating contextual information, merging representations, and forecasting the affective state of the upcoming utterance within dialogues.

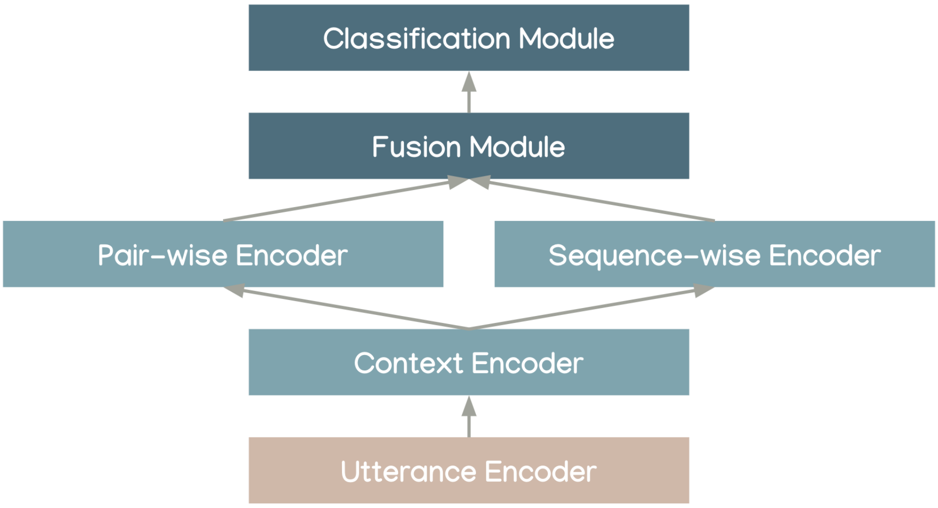


Figure 2.2: *General Architecture of Sentiment/Emotion Forecasting in Dialogue*

Several studies have focused on the task of **Emotion Forecasting in Dialogue**. Shahriar and Kim (2019) introduced the first research that aimed to forecast the future

doi:10.6342/NTU202304293

emotion of a speaker using audio-visual data. Their study highlighted the significance of considering both present and previous utterances when dealing with emotion forecasting. However, this study incurs several limitations. First, it only models the emotional flow of a single speaker and does not consider the other speaker's information. Second, the emotional flow may vary depending on the dialogue scene. To address these limitations, Shi et al. (2020) proposed an emotion forecasting model that incorporates the multi-modal context, including text and audio modalities, and takes into account the interactive information from both speakers. They also analyzed the impact of different conversation scenarios, such as comforting or convincing, on emotion changes. Wen et al. (2021) offered a different perspective on emotion forecasting, considering it as the process of "emotion transition" in the valence-arousal-dominance (VAD) space between the previous and upcoming emotions.

In terms of **Sentiment Forecasting in Dialogue**, which is the focus of our study, Wang et al. (2020) introduced this novel task and proposed the **Neural Sentiment Forecasting (NSF)** model. They focus on simulating the next upcoming utterance and learning the influence of context on the upcoming utterance. Zou et al. (2022) proposed the **Emotion-Assisted Sentiment Forecasting (EASF)** model, which incorporates sentiment and emotion features. Similar to the architecture mentioned above, they forecast sentiment by considering both internal and external influences. Additionally, their research explored the impact of each utterance using attention mechanisms and suggested applying this architecture to dialogue generation tasks in chatbots in future work.

These studies on dialogue emotion and sentiment forecasting have laid the foundation for future research. In contrast to existing methods, our study aims to explore other

doi:10.6342/NTU202304293

potential auxiliary tasks and features that can effectively extract useful information from the dialogue context, to help improve the performance of the sentiment forecasting task.

doi:10.6342/NTU202304293

Chapter 3

Methodology

3.1 Problem Formulation

In our research, we define the problem of sentiment forecasting in dialogue as follows: Given a dyadic dialogue history D, consisting of n utterances, denoted as $D = [u_1, u_2, \ldots, u_n]$, where each u_i represents the i-th utterance spoken within the dialogue. The objective of this task is to predict the sentiment polarity (positive, negative, or neutral) of the upcoming utterance u_{n+1}. It is important to note that the upcoming utterance, u_{n+1}, does not currently exist in the dialogue history. Also, during the testing phase, we encounter the lack of information regarding the sentiment polarity of the existing utterances within the dialogue.

3.2 Overview of Our Proposed Architecture

We propose a novel model called **Multi-task Sentiment Forecasting (MTSF)** to tackle the challenge of sentiment forecasting in dialogue. Our model treats sentiment forecasting as a multi-task problem, necessitating the handling of one main task along with three auxiliary tasks. The main task is the sentiment forecasting (upcoming), while the auxiliary tasks involve dialogue act forecasting (upcoming), sentiment recognition (sequence), and dialogue act recognition (sequence). The objective of this architecture is to improve the effectiveness of the main task, which is the sentiment forecasting (upcoming) task, by sharing the learned parameters or representations across multiple related auxiliary tasks. The complete architecture of the MTSF model is illustrated in Figure 3.1.

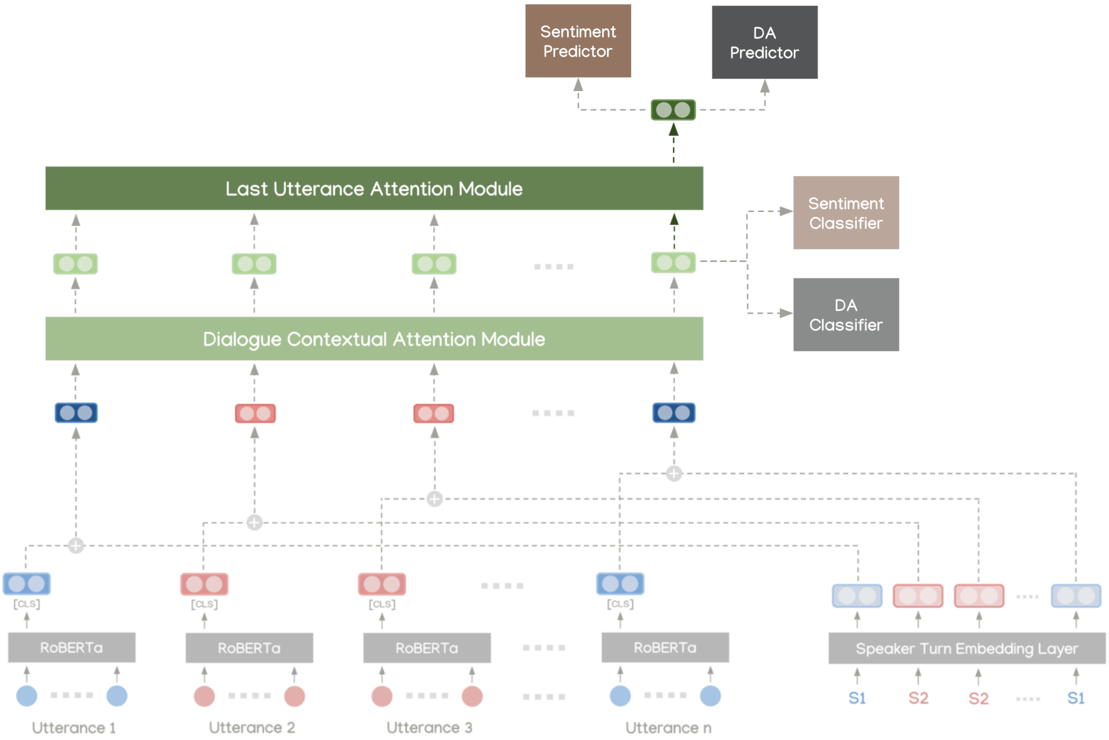


Figure 3.1: *Architecture of Our Proposed MTSF Model*

The **sentiment forecasting task (upcoming)** aims to forecast the sentiment polarity of the upcoming utterance u_{n+1}. Similarly, the **dialogue act forecasting task (upcoming)** aims to forecast the dialogue act label of the upcoming utterance u_{n+1}. The **sentiment recognition task (sequence)** focuses on recognizing the sentiment polarities of all the utterances in the dialogue history $D = [u_1, u_2, \ldots, u_n]$. The **dialogue act recognition task (sequence)** focuses on recognizing the dialogue act labels of all the utterances in the dialogue history D. It is important to note that the main task is the upcoming sentiment forecasting, while the other three tasks serve as auxiliary tasks, contributing to the model's learning of better representations.

Our proposed MTSF model consists of three fundamental feature extraction components. The first component is the **Utterance Encoder**, which utilizes the pretrained RoBERTa model (Liu et al., 2019b) to encode tokens in each utterance. The hidden state of the [CLS] token is then obtained as the representation of the utterance. The second component is the **Speaker Turn Embedding Layer**, which aims to learn representations for each speaker turn. In our dialogue scenario involving two speakers, we obtain two representations to indicate who is speaking in each utterance. These speaker representations are concatenated with the corresponding utterance representations based on their positions in the dialogue.

The third component is the **Dialogue Contextual Attention Module**, which employs multi-head attention to effectively capture interdependent contextual information among utterances in the dialogue, and utilizes a simple feed-forward network to reduce the dimensionality of the representations. This module produces contextual representations for each utterance in the dialogue, represented as $H' = [h'_1, h'_2, ..., h'_n]$.

After obtaining the contextual representation h'_i for each utterance u_i, two classifiers are employed: the **Sentiment Classifier** and the **Dialogue Act Classifier**. These classifiers use linear transformation and softmax function to utilize the contextual representation h'_i for recognizing the sentiment polarity and dialogue act label of each utterance u_i within the dialogue, respectively. Moreover, an additional attention module called the **Last Utterance Attention Module** is introduced. This module uses the contextual representation of the last utterance, h'_n, as a query, and applies multi-head attention to the contextual representations of the entire dialogue history sequence, H',

doi:10.6342/NTU202304293

in order to obtain a new representation for the last utterance, which we denote as h''_n. This representation h''_n is believed to capture the maximum potential information for the upcoming utterance, u_{n+1}.

Finally, with the representation h''_n of the last utterance u_n, two predictors are employed: the **Sentiment Predictor** and the **Dialogue Act Predictor**. These predictors use linear transformation and softmax function to utilize the representation h''_n for forecasting the sentiment polarity and dialogue act label of the upcoming utterance u_{n+1}, respectively.

3.3 Utterance Encoder

We employ the large pretrained language model RoBERTa to obtain contextual representations for each token in the utterances. Then, we rely on the hidden state of the [CLS] token from the last layer of RoBERTa to serve as the representation of each utterance u_i, denoted as $e(u_i)$. Figure 3.2 illustrates the workflow of utterance encoder.

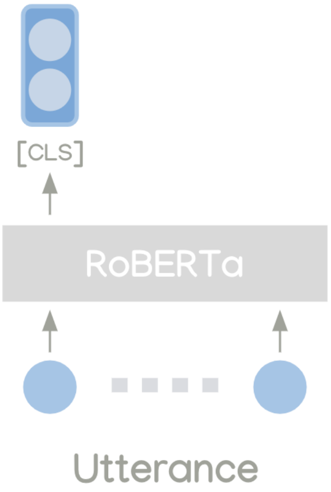


Figure 3.2: *Structure of Utterance Encoder*

Given that our experimental data consists of Chinese dialogues, we opt to employ the

doi:10.6342/NTU202304293

Erlangshen-Roberta-110M-Sentiment model (Zhang et al., 2022) as our pretrained model. This model is a sentiment analysis version of the Chinese RoBERTa-wwm-ext-base model, which has been fine-tuned on eight Chinese sentiment analysis datasets, comprising a total of 227,347 samples. We select this model as it offers a solid foundation for extracting valuable sentimental information from each utterance. In addition, it is worth noting that we do not fine-tune the parameters of the pretrained RoBERTa model during the training process.

3.4 Speaker Turn Embedding Layer

In order to achieve a comprehensive understanding of dialogues that arise in interactive environments, it is crucial to model **speaker turn-taking behavior** and account for the temporal dynamics of dialogues. The majority of previous research on this task has employed individual recurrent modules to model the information associated with each speaker role (Wang et al., 2020; Zou et al., 2022). However, this approach inevitably introduces a significant number of parameters that need to be trained.

Therefore, He et al. (2021) proposed a novel approach to enhance dialogue modeling by integrating speaker turns into the encoding process of utterances. Their method involves the introduction of **dialogue-invariant speaker turn embeddings**, which are independent of any given dialogue or speaker pair and then combined with utterance embeddings to capture the dynamics of turn changes among speakers within dialogues. This integration enables the effective representation of semantic information within dialogue content while considering the distinct contributions of different speakers. Moreover, their method introduces only two global additive embedding vectors, requiring

doi:10.6342/NTU202304293

minimal modifications to the model and introducing only O(1) space complexity.

In line with the work of He et al. (2021), our model incorporates a **Speaker Turn Embedding Layer**, to generate speaker turn representations based on the corresponding speaker labels. These representations are then combined with the utterance representations, yielding **speaker turn aware utterance representations**. To acquire the speaker turn aware utterance representation, denoted as $e'(u_i)$, for a given utterance u_i and its binary speaker turn label s_i, we add the speaker turn representation $f(s_i)$ to the utterance representation $e(u_i)$. This results in the formula $e'(u_i) = e(u_i) + f(s_i)$, where $s_i \in \{0,1\}$, and the symbol "+" denotes element-wise addition operator. It is worth noting that the speaker turn representations $f(s_i)$ are learnable parameters during optimization and have the same size as the utterance representations $e(u_i)$. Figure 3.3 displays the content of this module.

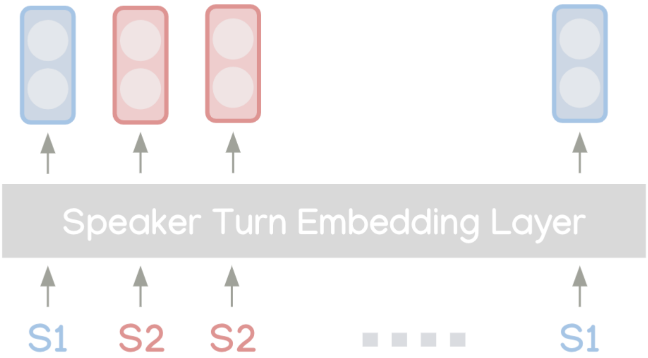


Figure 3.3: *Structure of Speaker Turn Embedding Layer*

By effectively combining the utterance and speaker turn information, this approach enhances the representation of each utterance with respect to its corresponding speaker turn. Leveraging these representations offers a simple yet effective approach to obtain more robust and informative utterance representations.

doi:10.6342/NTU202304293

3.5 Dialogue Contextual Attention Module

The Dialogue Contextual Attention Module employs a multi-head attention mechanism to effectively capture interdependent contextual information among utterances in the dialogue. It is followed by a fully connected feed-forward network, which applies a linear transformation with non-linear ReLU activation function at each position to effectively reduce the dimensionality of the representations. This module produces contextual representations for each utterance in the dialogue, represented as $H' = [h'_1, h'_2, \ldots, h'_n]$. The input and output of this module are illustrated in Figure 3.4, and the contextual representation h'_i of the i-th utterance u_i can be defined as follows:

$$h_i = Attention(e'(u_i),\ e'(u_j)_{j=1}^{n}) = \sum_{j=1}^{n} \alpha_{ij}\, e'(u_j) \tag{3.1}$$

$$h'_i = Feedforward(h_i) = max(0,\ Wh_i + b) \tag{3.2}$$

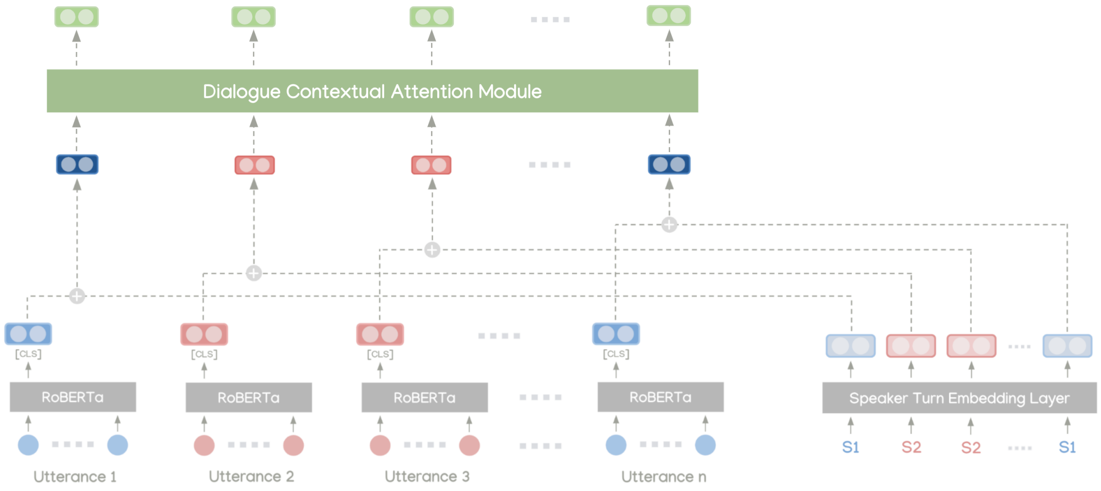


Figure 3.4: *Input/Output of Dialogue Contextual Attention Module*

3.6 Last Utterance Attention Module

The Last Utterance Attention Module plays a critical role in our model structure, as it generates the representation that will be fed into the main task module. By utilizing the contextual representation of the last utterance, denoted as h'_n, as a query, the module employs multi-head attention on the contextual representations of the entire dialogue history sequence, H'. This process yields a new representation for the last utterance, referred to as h''_n. **This representation $\mathbf{h''_n}$ is believed to capture the maximum potential information for the upcoming utterance, $\mathbf{u_{n+1}}$,** for the study conducted by Zou et al. (2022) has provided compelling evidence that distance plays a crucial role in dialogue sentiment forecasting. Specifically, their findings indicate that the relative impact increases as the utterance gets closer to the upcoming utterance. Additionally, it is worth noting that the design of this module also draws inspiration from the work of Shi et al. (2020) in their study on emotion forecasting task. The representation h''_n can be defined as follows:

$$h''_n = Attention(h'_n, (h'_j)_{j=1}^{n}) = \sum_{j=1}^{n} \alpha_{nj} h'_j \tag{3.3}$$

$$h''_n = ReLU(h''_n) = max(0, h''_n) \tag{3.4}$$

3.7 Multi-task Classification and Prediction

After obtaining the representations from various stages (e.g., h' or h''), we employ two classifiers and two predictors to perform four distinct tasks. These tasks collaboratively refine the underlying representations of the model with the aim of improving the effectiveness of the main task, which refers to the sentiment forecasting task.

Through the implementation of this multi-task approach, our proposed model achieves the simultaneous learning of multiple tasks, while also facilitating the sharing of certain parameters and representations among related tasks. This design not only fosters a complementary effect but also facilitates knowledge sharing between different tasks.

3.7.1 Sentiment Classifier and Dialogue Act Classifier (Sequence)

Once the contextual representation h_i' is obtained for each utterance u_i within the dialogue history sequence from the Dialogue Contextual Attention Module, we employ the Sentiment Classifier and the Dialogue Act Classifier. Both classifiers leverage the representation h_i' and apply a linear transformation and softmax function to obtain the predicted sentiment label distribution and the predicted dialogue act label distribution, respectively. These distributions are subsequently utilized to recognize the sentiment polarity and the type of dialogue act for each utterance u_i in the dialogue history sequence. The predicted distributions for sentiment polarity $\hat{y}_i^s$ and the type of dialogue act $\hat{y}_i^d$ for the i-th utterance u_i are exhibited as follows:

$$\hat{y}_i^s = softmax(W^s h_i' + b^s) \tag{3.5}$$

$$\hat{y}_i^d = softmax(W^d h_i' + b^d) \tag{3.6}$$

3.7.2 Sentiment Predictor and Dialogue Act Predictor (Upcoming)

There is another contextual representation h_n'', which is obtained for only the last utterance u_n from the Last Utterance Attention Module. Upon obtaining this critical representation h_n'', we proceed to employ the Sentiment Predictor and the Dialogue Act Predictor. Both predictors leverage the representation h_n'' and apply a linear transformation and softmax function to obtain the predicted sentiment label distribution and the predicted dialogue act label distribution, respectively. These distributions are subsequently utilized to forecast the sentiment polarity and the type of dialogue act for the upcoming utterance u_{n+1}, which has not yet appeared in the dialogue history. The predicted distributions for sentiment polarity $\hat{y}_{n+1}^s$ and the type of dialogue act $\hat{y}_{n+1}^d$ for the upcoming utterance u_{n+1}

doi:10.6342/NTU202304293

are shown as follows:

$$\hat{y}^{s}_{n+1} = softmax(W^{s'} h''_{n} + b^{s'}) \tag{3.7}$$

$$\hat{y}^{d}_{n+1} = softmax(W^{d'} h''_{n} + b^{d'}) \tag{3.8}$$

3.8 Optimization

We employ cross-entropy loss function for all four tasks. The loss function for each task is defined as follows:

$$L(\theta_y) = -\sum_{i=1}^{N}\sum_{j=1}^{C} y_{ij} \log \hat{y}_{ij} \tag{3.9}$$

Here, θ_y represents the set of model parameters. The variables y_{ij} and $\hat{y}_{ij}$ denote the true label and predicted label, respectively, for the i-th sample and j-th class. N refers to the number of training samples and C represents the number of classes in each task.

3.8.1 Dynamic Loss Weighting Strategy

According to Kongyoung et al. (2020), existing multi-task learning models have not explored the dynamic adjustment of the relative importance of different tasks during the learning process, which could lead to the allocation of training resources towards unnecessary tasks and negatively impact the model's performance. To address this gap, they introduced a new hybrid dynamic loss weighting strategy that combines Abridged Linear Schedule (Belharbi et al., 2016) for the main task with Loss-Balanced Task Weighting (Liu et al., 2019a) for the auxiliary tasks. This strategy automatically fine-tunes the task weighting during learning, ensuring that the loss weights of different tasks are adjusted based on their relative importance. They have also demonstrated the effectiveness of this dynamic strategy on the conversational question answering task, where it significantly outperforms static task weighting strategies.

Regarding the loss weighting strategy of the MTSF model, we follow the hybrid approach as Kongyoung et al. (2020) employed, with slight modifications. The specific implementation details are provided below.

First, to prioritize the main task and prevent unnecessary allocation of resources to other tasks, thus avoiding potential underfitting, we adopt the **Evolving Weighting Strategy with Linear Schedule** (Belharbi et al., 2016). This schedule gradually increases the loss weight of the main task during training, while gradually decreasing the loss weights of the auxiliary tasks. By doing so, the model can increasingly devote more attention to learning the main task. Specifically, the loss weights of the auxiliary tasks decrease linearly with each training step, with the auxiliary loss weights λ transitioning from 1 to 0. In contrast, the loss weight of the main task increases linearly as μ equals $1 - \lambda$. Importantly, assuming a predetermined total number of training steps T within each epoch, the loss weight of the main task at step t, denoted as μ_t, can be calculated as $\mu_t = \frac{t}{T}$. The variations in the loss weights of the main task μ and auxiliary tasks λ during the training process are illustrated in Figure 3.5.

Second, we further adjust the loss weights of each auxiliary task using the **Loss-Balanced Task Weighting (LBTW)** method proposed by Liu et al. (2019a). LBTW adjusts the loss weight based on the loss ratio between the current loss and the initial loss, giving higher priority to tasks with ratios closest to one in order to balance the importance of auxiliary tasks. To implement LBTW for all auxiliary tasks, we first employ a hyperparameter α to control the influence of task-specific loss weights. For each training

doi:10.6342/NTU202304293

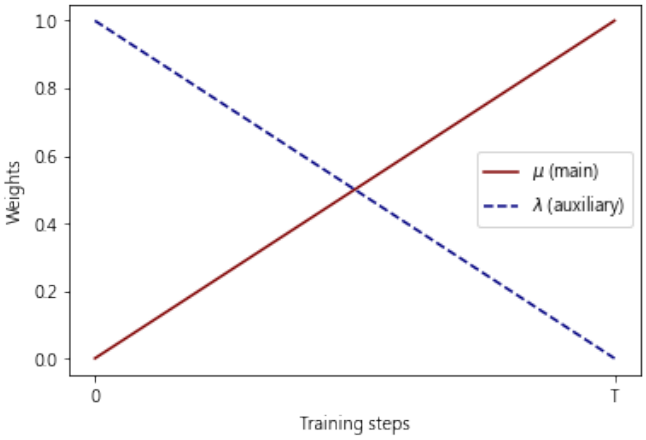


Figure 3.5: *Variations in the Loss Weights*

epoch, the loss weight of an auxiliary task k at step t, denoted as $\delta_{k,t}$, is computed using the loss ratio between the loss at the current step t of the epoch, denoted as $\ell_{k,t}$, and the loss at the first step of the epoch, denoted as $\ell_{k,0}$. This computation is represented as $\delta_{k,t} = (\frac{\ell_{k,t}}{\ell_{k,0}})^{\alpha}$.

Therefore, given an epoch with a total number of training steps T, the Linear Schedule strategy calculates the loss weight of the main task at step t, represented as μ_t, and the loss weight of the auxiliary tasks at step t, denoted as λ_t, can be defined as follows:

$$\mu_t = \frac{t}{T} \tag{3.10}$$

$$\lambda_t = 1 - \frac{t}{T} \tag{3.11}$$

Furthermore, the Loss-Balanced Task Weighting (LBTW) method calculates another version of the loss weight for each auxiliary task. Specifically, the loss weight of a specific auxiliary task k at step t, denoted as $\delta_{k,t}$, can be defined as follows:

$$\delta_{k,t} = (\frac{\ell_{k,t}}{\ell_{k,0}})^{\alpha} \tag{3.12}$$

Finally, the loss weights w_m for the main task and w_k for an auxiliary task, which are

doi:10.6342/NTU202304293

utilized to optimize the model, can be defined as follows:

$$w_m = \mu_t \tag{3.13}$$

$$w_k = \lambda_t \times \delta_{k,t} \tag{3.14}$$

3.8.2 Weighted Loss Aggregation

After obtaining the dynamic loss weight for each task, the final loss is computed as the weighted aggregation of the losses from all tasks. Our training objective is to minimize this aggregated loss:

$$L_{final} = w_m \times L_m + \sum_{k=1}^{K} w_k \times L_k \tag{3.15}$$

Chapter 4

Domain Transfer Strategies

4.1 Domain Transfer Strategies

After training the model using the method described in the previous chapter, the practical deployment of this model in real-world scenarios may encounter situations where annotations are relatively sparse. For instance, there might be a lack of auxiliary task annotations. Therefore, it becomes imperative to establish effective strategies for applying the trained model across various domains, aligning it with practical applications. This chapter aims to explore how to transfer the acquired knowledge learned from a given dataset to new domains, considering different application scenarios, and efficiently utilize data from these domains.

Depending on our understanding of a new application domain, we can choose different methods to transfer the acquired knowledge effectively. In the following sections, we will discuss two common downstream application scenarios and propose two distinct approaches for knowledge transfer.

doi:10.6342/NTU202304293

4.2 Fine-tuning with Limited Labeled Data

In this section, we discuss the first scenario where the target domain contains labeled data; however, the labels are incomplete, and the data quantity is extremely limited. Specifically, in our study, we consider the case where only the main task, namely sentiment forecasting, are labeled, while the auxiliary tasks lack annotations. In this scenario, we can employ a fine-tuning approach on the pre-trained model.

To accomplish this, given a pretrained model, we perform further training on the data from the target domain, focusing solely on the sentiment forecasting task. During this fine-tuning process, we keep the parameters of the lowest layers (i.e., the Utterance Encoder and the Speaker Turn Embedding Layer) fixed and exclude them from the training process. Furthermore, we utilize a smaller learning rate to ensure careful adjustments during the fine-tuning stage. The specific architecture of our method is illustrated in Figure 4.1.

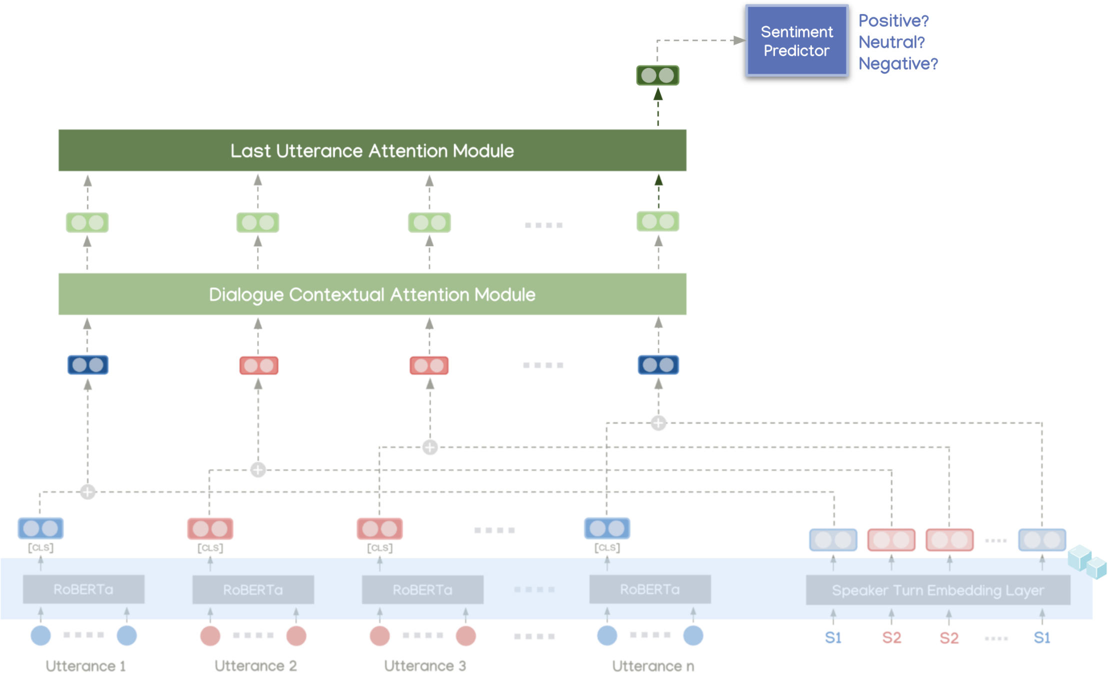


Figure 4.1: *Specific Architecture of the Fine-tuning Approach*

4.3 Domain Adversarial Training with Unlabeled Data

In this section, we discuss the second scenario where the data in the target domain lacks any form of annotation, but the quantity of data is substantial. This scenario better reflects real-world situations where labeled data is unavailable in the target domain, but unlabeled data may be abundant (Ramponi and Plank, 2020).

In the absence of labeled data, directly training a model to learn a specific task on the target domain becomes infeasible. Numerous studies have proposed various approaches to address this challenge, with the most common method being **Domain Adversarial Training** (Ganin and Lempitsky, 2015; Ganin et al., 2016). The objective of Domain Adversarial Training is to improve latent feature representations by simultaneously training the label predictor and domain classifier. The original architecture of this method is shown in Figure 4.2.

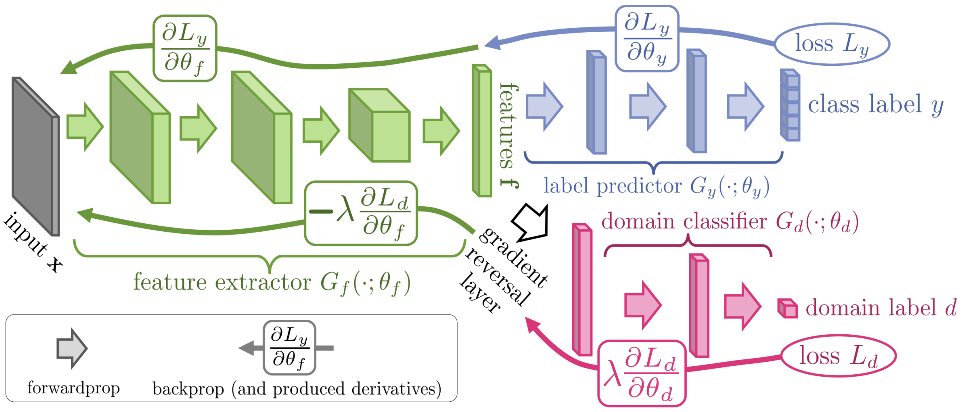

$\frac{\partial L_y}{\partial \theta_f}$
$\frac{\partial L_y}{\partial \theta_y}$
loss L_y
class label y
features $\mathbf{f}$
label predictor $G_y(\cdot;\theta_y)$
$-\lambda \frac{\partial L_d}{\partial \theta_f}$
input x
feature extractor $G_f(\cdot;\theta_f)$
gradient reversal layer
domain classifier $G_d(\cdot;\theta_d)$
domain label d
loss L_d
$\lambda \frac{\partial L_d}{\partial \theta_d}$
forwardprop
backprop (and produced derivatives)

Figure 4.2: *Original Architecture of the Domain Adversarial Training Approach (Adopted from Ganin and Lempitsky (2015))*

In our research, drawing inspiration from the concept of domain adversarial training, we present the specific implementation details as follows: leveraging the pretrained model, we train the main task (i.e., sentiment forecasting) again, using the labeled data

doi:10.6342/NTU202304293

from the source domain with a smaller learning rate. Additionally, we introduce a new task that utilizes data from both the source and target domains. This new task involves predicting whether a given data instance (i.e., a dialogue history sequence) belongs to the source domain or the target domain. The input to this task is the contextual representation h_i' of each utterance u_i in the dialogue history sequence. Ultimately, the losses of these two tasks are aggregated and jointly optimized during model training. The specific architecture of our method is illustrated in Figure 4.3.

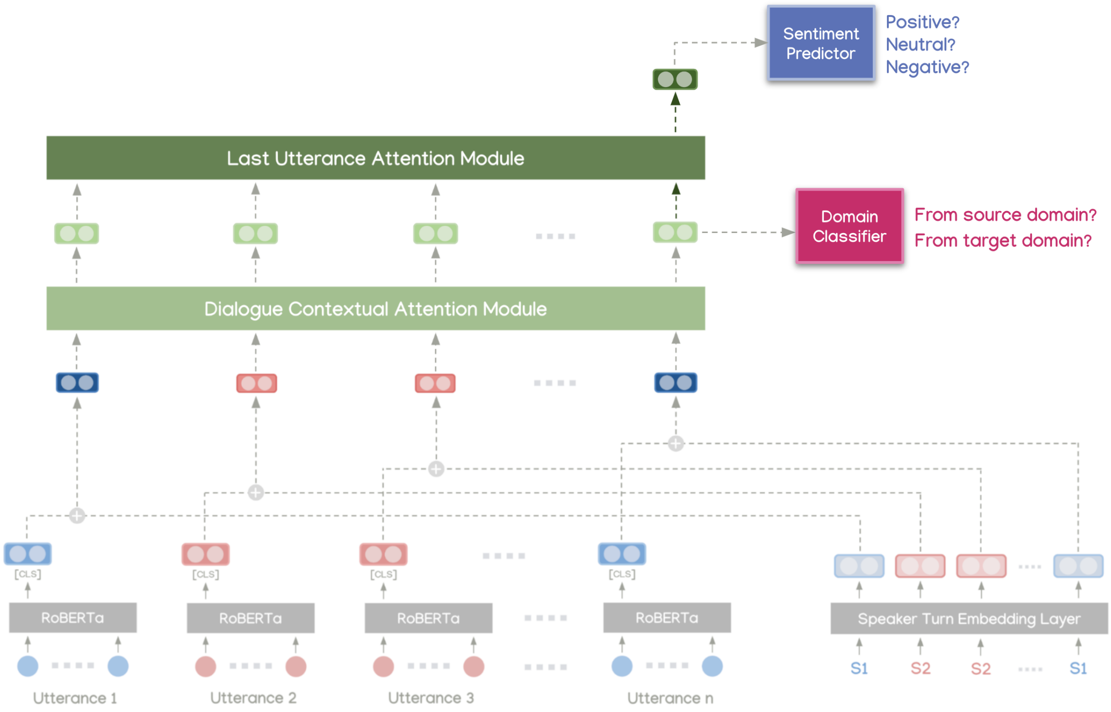


Figure 4.3: *Specific Architecture of the Domain Adversarial Training Approach*

doi:10.6342/NTU202304293

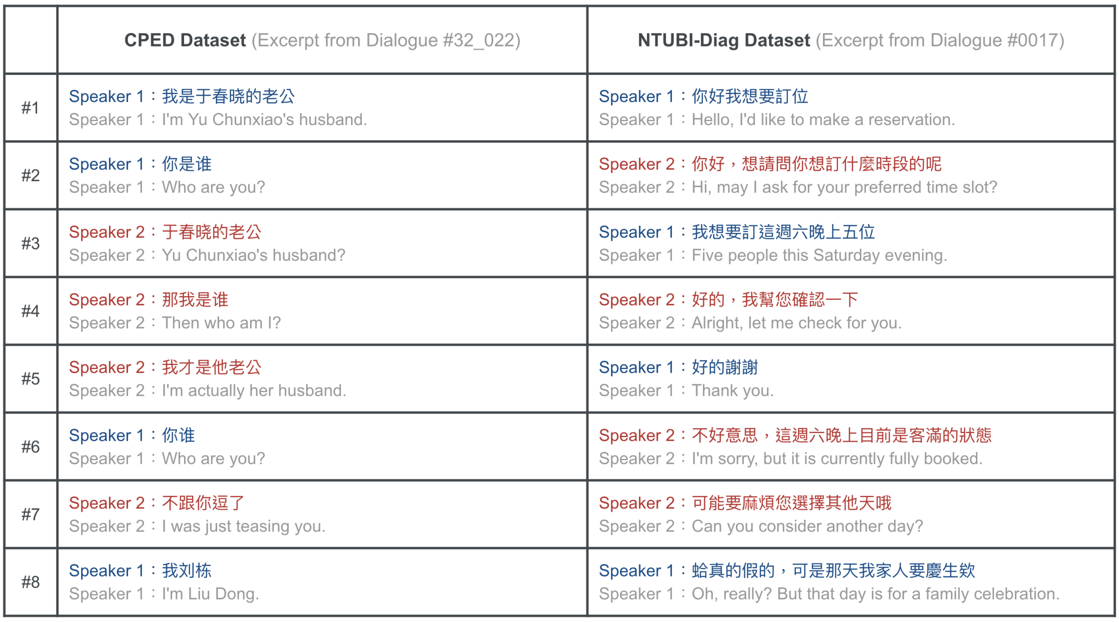
Chapter 5

Empirical Evaluation

5.1 Data Collection

In our research, we conduct experiments using two different datasets. The first dataset is multi-turn Chinese Personalized and Emotional Dialogue dataset, called **CPED** (Chen et al., 2022). CPED is constructed from 40 Chinese TV shows, comprising 12,000 dialogues and 133,000 utterances with multi-modal context. The creators of this dataset claim that CPED is the first Chinese personalized and emotional dialogue dataset. As a result, it can be utilized for complex dialogue understanding tasks.

However, we believe that the data source of CPED, being from TV shows, may not entirely align with real-world dialogue system scenarios. Therefore, in our research, we establish another dialogue dataset called **NTUBI-Diag**, which is collected by our Business Intelligence Lab at National Taiwan University. The inclusion of this dataset is significant as it provides simulations of common real-world dialogue system scenarios, such as customer service dialogues on e-commerce platforms. This new dataset also includes labels for upcoming sentiment, but lacks additional annotations for the auxiliary tasks. We believe that this dataset serves as a more practical and realistic foundation, or starting point, for the sentiment forecasting task in real-world dialogue system applications. The dialogue excerpts from the two datasets are presented in Figure 5.1.



	CPED Dataset (Excerpt from Dialogue #32_022)	**NTUBI-Diag Dataset** (Excerpt from Dialogue #0017)
#1	Speaker 1：我是于春晓的老公 Speaker 1：I'm Yu Chunxiao's husband.	Speaker 1：你好我想要訂位 Speaker 1：Hello, I'd like to make a reservation.
#2	Speaker 1：你是谁 Speaker 1：Who are you?	Speaker 2：你好，想請問你想訂什麼時段的呢 Speaker 2：Hi, may I ask for your preferred time slot?
#3	Speaker 2：于春晓的老公 Speaker 2：Yu Chunxiao's husband?	Speaker 1：我想要訂這週六晚上五位 Speaker 1：Five people this Saturday evening.
#4	Speaker 2：那我是谁 Speaker 2：Then who am I?	Speaker 2：好的，我幫您確認一下 Speaker 2：Alright, let me check for you.
#5	Speaker 2：我才是他老公 Speaker 2：I'm actually her husband.	Speaker 1：好的謝謝 Speaker 1：Thank you.
#6	Speaker 1：你谁 Speaker 1：Who are you?	Speaker 2：不好意思，這週六晚上目前是客滿的狀態 Speaker 2：I'm sorry, but it is currently fully booked.
#7	Speaker 2：不跟你逗了 Speaker 2：I was just teasing you.	Speaker 2：可能要麻煩您選擇其他天哦 Speaker 2：Can you consider another day?
#8	Speaker 1：我刘栋 Speaker 1：I'm Liu Dong.	Speaker 1：蛤真的假的，可是那天我家人要慶生欸 Speaker 1：Oh, really? But that day is for a family celebration.

Figure 5.1: *Dialogue Excerpts from CPED and NTUBI-Diag Datasets*

5.1.1 Chinese Personalized and Emotional Dialogue Dataset (CPED)

CPED is a comprehensive dataset that includes multi-source knowledge, covering 3 sentiments, 13 emotions, 19 dialogue acts, gender, big five personality traits, and other information. Table 5.1 presents a detailed description of the specific categories for sentiment and dialogue act annotations within the CPED dataset. Meanwhile, Table 5.2 provides the detail summary statistics of the original CPED dataset.

Table 5.1: *Description of Annotation Categories within the CPED Dataset*

Annotation	**Categories**
Sentiment	positive, neutral, negative
Dialogue Act	greeting, question, answer, statement-opinion, statement-non-opinion, apology, command, agreement/acceptance, disagreement, acknowledge, appreciation, interjection, conventional-closing, thanking, quotation, reject, irony, comfort, other

Table 5.2: *Statistics of the Original CPED Dataset*

	Training	Dev	Testing
# of TV plays	26	5	9
# of dialogues	8,086	934	2,815
# of utterances	94,187	11,137	27,438
# of speakers	273	38	81
Avg. # utt. per dial.	11.6	11.9	9.7
Max # utt. per dial.	75	31	34
Avg. utt. length	8.3	8.2	8.3
Max utt. length	127	42	45
Avg. # of DAs per dial.	3.6	3.7	3.2

As mentioned, our primary research objective is to forecast users' upcoming sentiment, enabling the system to proactively adjust responses to users. To align with this real-world application, we need to ensure that the "last utterance in the dialogue history sequence (u_n)" and the "upcoming utterance (u_{n+1})" are spoken by different speakers. To achieve this, we further process the original CPED data to obtain new processed data for model training. First, we define "speaker transitions" as situations where two consecutive utterances are spoken by different speakers, and such speaker transitions may occur multiple times within the same dialogue. To represent these transitions, we use $trans_i$ to denote the i-th occurrence of speaker transition within a specific dialogue. Since each speaker transition occurs between two consecutive utterances, we further define $trans_i = (u_j, u_{j+1})$, where u_j and u_{j+1} are the two utterances involved in the i-th speaker transition. Next, we check if the total number of utterances before u_j (inclusive) in the original dialogue data is greater than or equal to n, where n represents the desired dialogue history sequence length. If it meets the condition, we extract the n preceding utterances at u_j (inclusive) as one training data sample, and consider the sentiment at u_{j+1} position as the target sentiment label for that sample. The statistics of the adjusted CPED dataset are shown in Table 5.3. Furthermore, the distribution of sentiment polarities in the training and the testing dataset respectively is depicted in Figure 5.2. Notably, since our experiments are conducted using a 5-fold approach on the training and testing data, we combine the information reported in the figure and table.

Table 5.3: *Statistics of the Adjusted CPED Dataset*

	Training + Testing	Dev
# of dialogues	6,668	600
# of utterances	53,344	4,800
# of utt. per dial.	8	8
Avg. utt. length	8.6	8.4
Max utt. length	54	31
Avg. # of DAs per dial.	3.0	3.1

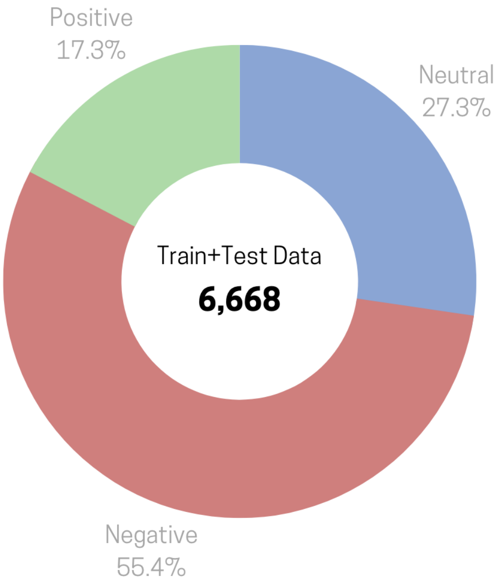


Figure 5.2: *Distribution of Sentiment Polarities in the Training and Testing Data (CPED)*

Two noteworthy aspects are worth mentioning: First, within each dialogue, the same speaker may consecutively deliver multiple utterances, distinguishing CPED from other dialogue datasets. Second, in cases where the original dialogue data contains numerous utterances and speaker transitions, our processing approach may result in multiple new data samples.

5.1.2 Newly Collected Dialogue Dataset (NTUBI-Diag)

In order to better simulate real-world application scenarios, our research develops a novel dataset called NTUBI-Diag, which covers various common dialogue system applications. This dataset spans a wide range of simulated dialogue scenes, from issue resolution in customer service platforms, reservation and compensation matters in the service industry, to sales and bargaining in shopping scenarios. It even includes dialogues portraying expressions of care and casual conversations between friends, making it comprehensive in its coverage. Each scene involves two speakers engaging in the dialogue, with roles and tasks specific to each scene. The dataset covers 14 distinct scenes, carefully designed to cover a wide range of applications and interactions. The specific details of these scenes are summarized in Table 5.4.

Furthermore, in each dialogue, every speaker is randomly assigned a specific personality trait, allowing for a more authentic simulation of real-life conversational dynamics. The detailed personality traits and their occurrence probabilities in the dataset are presented in Table 5.5.

To generate this dataset, we collaborated with numerous individuals who engaged in role-playing and dialogue generation within the specified scenarios. We have collected a total of 480 dialogues, with 8 individuals taking on the role of initiating the dialogue by finding other individuals outside the group to engage in simulated dialogues. Similar to the adjusted CPED dataset, the collected data undergoes the same processing, transforming it into a reasonable format for model training. The selection of dialogue history sequence length (i.e., *n*) also follows the same criteria as in the adjusted CPED

doi:10.6342/NTU202304293

Table 5.4: *Details of Simulated Dialogue Scenes*

ID	Scene	Role	Task
1	Bargaining at a holiday market	Tourist / Vendor	Negotiating for a lower price or selling at the original price
2	Bargaining on an E-commerce platform	Customer / Seller	Seeking expired discounts or attempting to sell at the original price
3	Dealing with return requests	Customer / Seller	Requesting seller or buyer to cover shipping fee
4	Complaining in the customer service platform	Customer / Support	Requesting or refusing compensation for product defects
5	Regretting overpaying after seeing new discount	Customer / Support	Requesting or refusing post-sale price difference compensation
6	Urgently requesting restaurant reservation	Guest / Staff	Attempting reservation or declining reservation request
7	Handling reservation matters	Staff / Guest	Explaining the failed reservation or seeking dining rights
8	Complaining in the restaurant	Guest / Manager	Requesting or refusing compensation for poor dining experience
9	Complaining in the hotel	Guest / Manager	Requesting or refusing compensation for hotel stay issue
10	Offering a discount	Salesperson / Passerby	Promoting or bargaining for a product discount
11	Complaining	Student A / Student B	Expressing frustration or comforting the other
12	Emotional sharing	Student A / Student B	Sharing emotional issues or supporting the other
13	Inviting for an outing	Student A / Student B	Persuading or refusing the invitation
14	Discussing a weekend trip	Student A / Student B	Persuading or refusing some proposal

doi:10.6342/NTU202304293

Table 5.5: *Details of Personality Traits and Their Occurrence Probabilities*

Personality Trait	Probability of Occurrence (%)
Highly Stubborn and Unyielding	20%
Rationally Argumentative	25%
Willing to Listen and Communicate	25%
Empathetic and Considerate	20%
Prefers Indirect and Non-confrontational Approaches	10%

dataset. The statistical details of this dataset are presented in Table 5.6. Also, the distribution of sentiment polarities in the training and testing data is depicted in Figure 5.3.

Table 5.6: *Statistics of the NTUBI-Diag Dataset*

	Training + Testing	Dev
# of dialogues	3,142	338
# of utterances	25,136	2,704
# of utt. per dial.	8	8
Avg. utt. length	11.0	10.8
Max utt. length	50	38

doi:10.6342/NTU202304293

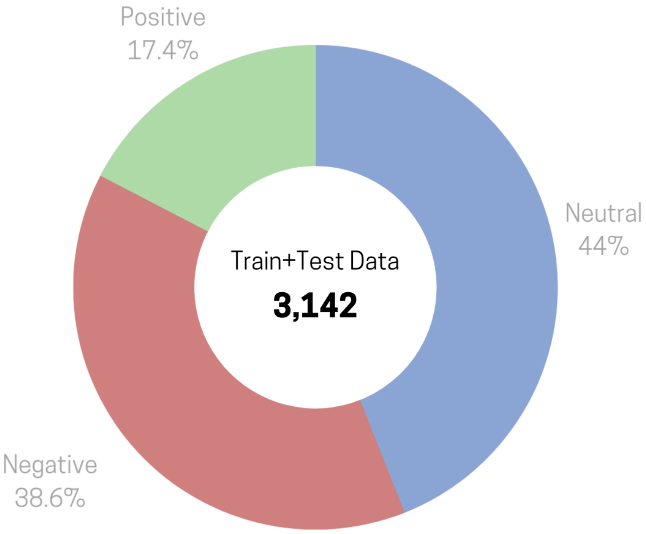


Figure 5.3: *Distribution of Sentiment Polarities (NTUBI-Diag)*

5.2 Evaluation Procedure and Metrics

In order to ensure credible and fair comparisons, we choose not to utilize the original train-dev-test splitting of the dataset. Instead, we adopt a 5-fold dataset splitting approach. First, we keep the development data unchanged, utilizing it as a reference for selecting hyperparameters. Subsequently, we divide the training and testing data into five subsets. Each experiment involves training the model on 80% of the data and testing it on the remaining 20% for five iterations, where the testing data in an iteration corresponds to each fold. Consequently, with this 5-fold cross-validation method, the performance of each experiment will be reported as the weighted average of each fold's results. We opt for using weighted averages due to the split of the dataset is based on TV series, which is designed to prevent data leakage. Given the slight variations in data quantities among different TV series, the use of weighted averages ensures the fairness and robustness of our experimental results.

doi:10.6342/NTU202304293

Regarding the evaluation of model performance, we follow the metrics adopted in previous studies on sentiment forecasting task (Wang et al., 2020; Zou et al., 2022), and these metrics are also consistent with those in Chen et al. (2022) using the CPED dataset. Specifically, we will employ precision, recall, and F1 score with macro-averaging as our evaluation metrics. Furthermore, we conducted individual calculations for these metrics within each sentiment polarity category, which enable us to gain a comprehensive understanding of how each model performs differently in forecasting various sentiment polarities. The formulations for these metrics are as follows, with polarity-*i* belonging to the set {positive, neutral, negative}:

$$Precision_{polarity-i} = \frac{\textit{\# correctly predicted samples of polarity-i}}{\textit{\# total predicted samples of polarity-i}} \tag{5.1}$$

$$Recall_{polarity-i} = \frac{\textit{\# correctly predicted samples of polarity-i}}{\textit{\# total annotated samples of polarity-i}} \tag{5.2}$$

$$F1_{polarity-i} = \frac{2 \times Precision_{polarity-i} \times Recall_{polarity-i}}{Precision_{polarity-i} + Recall_{polarity-i}} \tag{5.3}$$

5.3 Experimental Settings

5.3.1 Implementation Details

In our research, all experiments are implemented using the PyTorch (Paszke et al., 2019) deep learning framework. For model optimization, we utilize the widely adopted Adam (Kingma and Ba, 2014) optimizer for all our experiments. As for the learning rate, we employ the Linear learning rate schedule with warm-up. Specifically, we linearly increase the learning rate from zero to a predefined target learning rate *lr* during the first 3% of training steps, and gradually decrease it back to zero using a linear decay schedule

doi:10.6342/NTU202304293

until the final step. For the MTSF model and all benchmark models, we set the target learning rate lr to 5×10^{-5}. However, in the case of Transfer Learning experiments, where we deal with limited data or fine-tuning model parameters, we adjust the target learning rate lr to a lower value of 5×10^{-6}.

Furthermore, we set the maximum training epochs to 300 and implement early stopping with patience set to 50 epochs. This means that if the validation performance of the model did not improve over the last 50 consecutive epochs, the training process is terminated early to prevent overfitting. Additionally, throughout all experiments, we maintain a consistent batch size of 32. Regarding the multi-head Attention settings, we employ an 8-head multi-head attention mechanism for Dialogue Contextual Attention and a 4-head multi-head attention for Last Utterance Attention. Finally, the hyperparameter α, which is used to adjust the weights for auxiliary tasks, is set to 0.5, as it performs best in the original paper's experiments (Liu et al., 2019a).

It is noteworthy that the choice of dialogue history length (i.e., n) has been investigated in previous studies, ranging from 3 (Wang et al., 2020) and 4 (Zou et al., 2022) to 8 (Shi et al., 2020), and even beyond. However, given the distinctive nature of the dataset used in our experiments, where consecutive utterances by the same speaker occur, we believe it is necessary to select a longer dialogue history length to cover information from both speakers more comprehensively. The more detailed hyperparameter settings for our experiments are summarized in Table 5.7.

5.3.2 Benchmark Methods

We will compare our proposed MTSF model with two existing methods:

doi:10.6342/NTU202304293

Table 5.7: *Hyperparameter Settings*

Epochs	Early stopping epochs	Learning rate	Utterance embedding dim.	Speaker Turn embedding dim.
$e = 300$	$w_{es} = 50$	$lr_{multi} = 5e-5$ $lr_{trans} = 5e-6$	$d_u = 768$	$d_s = 768$

Batch size	Feed forward hidden dim.	Attention head number	Task weight balancer	Dialogue history length
$b = 32$	$d_{ff} = 128$	$h_{cont} = 8$ $h_{last} = 4$	$\alpha = 0.5$	$n = 8$

- **Neural Sentiment Forecasting (NSF)** (Wang et al., 2020): The NSF method focuses on simulating the next upcoming utterance and learning the influence of context on the upcoming utterance. Specifically, they employ attention mechanism to capture and fuse information from both "utterances spoken by the target speaker" and the "entire historical dialogue sequence" for sentiment forecasting tasks.

- **Emotion-Assisted Sentiment Forecasting (EASF)** (Zou et al., 2022): Similar to NSF, the EASF method also utilizes attention mechanism to capture and fuse information from "utterances spoken by the target speaker" and the "entire historical dialogue sequence." However, they further consider the incorporation of emotion features to assist in sentiment forecasting tasks.

It is essential to note that, to ensure result comparability and fairness, we use the same pretrained RoBERTa model, as employed in our research, to generate the underlying utterance embeddings when replicating these benchmark methods. Additionally, as our study uses a different dataset from the original research of these benchmark methods, we adopt the hyperparameter settings consistent with those listed in Table 5.7 in cases where the original papers did not provide explicit details on hyperparameter selection.

These two benchmark methods presented in this section provide valuable insights into the current landscape of sentiment forecasting in dialogue systems, and serve as benchmarks for evaluating the effectiveness of our proposed MTSF model.

5.4 Evaluation Results

In this section, we present a comprehensive performance evaluation of our proposed MTSF model and compare it with the two benchmark methods, namely NSF and EASF. The evaluation results are summarized in Table 5.8, where we report the macro F1 score, macro precision, and macro recall for each method.

Table 5.8: *Comparison of Benchmarks and Our Proposed MTSF Method (on CPED Dataset)*

Method	Macro F1	Macro Precision	Macro Recall
NSF (Wang et al., 2020)	36.47%	41.23%	37.96%
EASF (Zou et al., 2022)	37.89%	38.09%	38.00%
MTSF	**41.26%**	**41.65%**	**41.60%**

As shown in Table 5.8, our proposed MTSF method demonstrates superior performance across all three metrics compared to those attained by the benchmark methods. The bold texts denote the best performance in each evaluation criterion. Specifically, the macro F1 score of MTSF is 41.26%, indicating a significant improvement over both NSF (36.47%) and EASF (37.89%). Similarly, the macro precision of MTSF stands at 41.65%, outperforming both NSF (41.23%) and EASF (38.09%).

Moreover, MTSF achieves an impressive macro recall of 41.60%, surpassing the performance of NSF (37.96%) and EASF (38.00%). These results suggest that our proposed MTSF method strikes a better balance between precision and recall, making it well-suited for the sentiment forecasting task.

For a more detailed analysis of the performance of MTSF across different sentiment polarities, please refer to Table 5.9. In that table, we provide a comprehensive breakdown of the precision, recall, and F1 scores for each sentiment class, enabling a thorough understanding of each method's efficacy in capturing diverse sentiment patterns.

Table 5.9: *Comparison of Performance Across Different Sentiment Polarities (on CPED Dataset)*

Method	**Negative**			**Positive**			**Neutral**		
	F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
NSF	**69.60%**	59.01%	**84.83%**	14.91%	**28.33%**	10.12%	24.91%	36.35%	18.95%
EASF	62.31%	60.78%	63.91%	22.13%	20.82%	23.62%	29.24%	32.66%	26.47%
MTSF	63.67%	**62.50%**	64.89%	**28.09%**	25.08%	**31.92%**	**32.02%**	**37.36%**	**28.01%**

Starting with the **negative sentiment class**, NSF achieves the highest F1 score, indicating a good balance between precision and recall. While NSF significantly outperforms the other methods in recall, our proposed MTSF method exhibits a slightly better precision, indicating a stronger ability to correctly forecast negative sentiment instances.

Moving on to the **positive sentiment class**, the results show that our proposed MTSF method achieves the highest F1 score, significantly outperforming both NSF and EASF. Moreover, MTSF exhibits the highest recall, surpassing NSF and EASF. This indicates that MTSF is more effective in forecasting positive sentiment instances.

Regarding the **neutral sentiment class**, our proposed MTSF method also achieves the highest F1 score, significantly surpassing both NSF and EASF. Moreover, MTSF demonstrates better precision and recall compared to those of NSF and EASF. This indicates that MTSF strikes a better balance between precision and recall for forecasting neutral sentiment instances.

In summary, the detailed performance evaluation reveals that our proposed MTSF method outperforms the benchmark methods in forecasting positive and neutral sentiment, while still demonstrating a competent ability to forecast negative sentiment. This also suggests that MTSF achieves a well-balanced performance, making it suitable for the sentiment forecasting task.

5.5 Additional Evaluation Results

5.5.1 Effectiveness of Auxiliary Tasks

One of the most significant advantages of our proposed MTSF method lies in its effective utilization of a multi-task framework, which enhances the overall performance of the sentiment forecasting task. We firmly believe that learning to recognize dialogue acts and sentiments within the dialogue history sequence can provide valuable information and clues for improving the performance of sentiment forecasting. Similarly, the ability to predict upcoming dialogue acts can also offer insights that benefit the sentiment forecasting task. Motivated by this rationale, we incorporated the auxiliary tasks discussed in Chapter 3.7 into our proposed MTSF model structure, empowering the model to acquire more robust sentiment representations.

In the following experiments, we conduct a comparative analysis of the effectiveness of our proposed method by selectively excluding one or more auxiliary tasks. Specifically, we systematically remove the loss corresponding to certain auxiliary tasks from the total loss function to observe their impact on the forecast effectiveness of the main task. The ablation results are presented in Table 5.10, and for a more detailed analysis of the effectiveness of different auxiliary tasks across different sentiment polarities, please refer to Table 5.11. Note that the “w/o” prefix denotes the exclusion of specific auxiliary tasks

doi:10.6342/NTU202304293

from the model training.

Table 5.10: *Effectiveness of Different Auxiliary Tasks (on CPED Dataset)*

Method	**Macro F1**	**Macro Precision**	**Macro Recall**
w/o All Auxiliary Tasks	38.02%	38.29%	38.04%
w/o DA(Seq) & Sentiment(Seq)	38.81%	39.52%	38.83%
w/o DA(Upcoming) & Sentiment(Seq)	38.83%	39.43%	39.14%
w/o DA(Upcoming) & DA(Seq)	39.57%	39.53%	39.70%
w/o Sentiment(Seq)	39.68%	40.17%	39.92%
w/o DA(Seq)	39.88%	40.26%	39.79%
w/o DA(Upcoming)	40.02%	40.16%	39.95%
MTSF	**41.26%**	**41.65%**	**41.60%**

Table 5.11: *Effectiveness of Auxiliary Tasks Across Different Sentiment Polarities (on CPED Dataset)*

Method	**Negative**			**Positive**			**Neutral**		
	F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
w/o All Auxiliary Tasks	63.34%	60.22%	66.81%	21.24%	22.35%	20.24%	29.46%	32.31%	27.07%
w/o DA(Seq) & Sentiment(Seq)	**64.87%**	60.26%	**70.25%**	21.39%	24.55%	18.94%	30.18%	33.74%	27.29%
w/o DA(Upcoming) & Sentiment(Seq)	64.26%	60.24%	68.87%	25.34%	24.67%	26.04%	26.89%	33.39%	22.52%
w/o DA(Upcoming) & DA(Seq)	60.92%	61.57%	60.28%	25.54%	24.03%	27.25%	32.27%	32.99%	31.58%
w/o Sentiment(Seq)	64.44%	60.96%	68.33%	26.22%	25.34%	27.16%	28.40%	34.21%	24.27%
w/o DA(Seq)	64.04%	60.98%	67.43%	24.93%	**26.85%**	23.27%	30.66%	32.95%	28.67%
w/o DA(Upcoming)	61.64%	61.68%	61.61%	25.51%	26.84%	24.31%	**32.92%**	31.95%	**33.94%**
MTSF	63.67%	**62.50%**	64.89%	**28.09%**	25.08%	**31.92%**	32.02%	**37.36%**	28.01%

The experimental results in Table 5.10 clearly demonstrate that the exclusion of any auxiliary task leads to a decline in the overall effectiveness of the sentiment forecasting task, and removing all auxiliary tasks (w/o all auxiliary tasks) leads to a noticeable drop in performance. This observation highlights the importance of the auxiliary tasks in contributing to the overall effectiveness of our sentiment forecasting method.

From the experimental results in the Table 5.11, it is evident that different auxiliary tasks contribute differently to the forecasting of sentiment in various sentiment categories. Specifically, for the forecast of “negative” samples, the auxiliary task of forecasting upcoming dialogue act “DA(Upcoming)” exhibits the most significant improvement. On

doi:10.6342/NTU202304293

the other hand, for the prediction of "positive" samples, both tasks related to recognizing sentiments and dialogue acts within the dialogue history sequence demonstrate substantial benefits. Meanwhile, in the case of predicting "neutral" samples, the task of recognizing sentiments within the dialogue history sequence proves to be the most crucial. These findings suggest that each auxiliary task focuses on extracting distinct emotional features relevant to the specific sentiment categories.

Furthermore, the experimental results also reveal that incorporating all auxiliary tasks simultaneously leads to a significant performance improvement in forecasting "positive" samples. Although the performance on "negative" and "neutral" samples may not reach its optimum, it closely approaches the performance achieved by the best-performing task combination. This observation implies that incorporating all auxiliary tasks simultaneously provides a more stable overall forecasting performance across different sentiment categories.

Through the evaluation of our method's performance under various scenarios of auxiliary task exclusion, we have gained valuable insights into the contribution of each auxiliary task to the overall effectiveness of sentiment forecasting. This analysis has allowed us to understand the significance of each auxiliary task in refining sentiment representations and effectively complementing the primary sentiment forecasting task. In summary, our experimental findings demonstrate that each auxiliary task plays a meaningful role in enhancing the method's ability to perform sentiment forecasting effectively. By adopting the multi-task framework and incorporating these auxiliary tasks, our proposed MTSF model achieves superior performance, highlighting the importance of leveraging diverse sources of information to achieve more accurate sentiment forecasting results.

5.5.2 Experiments on Domain Transfer

To ensure the applicability of our proposed method in real-world scenarios, we conduct domain transfer experiments on the newly collected NTUBI-Diag dataset, as described in Chapter 4. These experiments aim to evaluate the method's ability to generalize across different domains or application contexts. Specifically, we explore two distinct domain transfer methods as follows:

- **Domain Adversarial Training with CPED-Pretrained MTSF Model (DaNN):** We first train the MTSF model on the CPED dataset, and then adapt it to the newly collected NTUBI-Diag dataset through domain adversarial training, without any labeled data on the NTUBI-Diag dataset.
- **Fine-tuning with CPED-Pretrained MTSF Model (Fine-tuning):** We first train the MTSF model on the CPED dataset, and then fine-tune it using the newly collected NTUBI-Diag dataset, assuming that we have access to labeled data for the sentiment forecasting task on the NTUBI-Diag dataset.

Furthermore, to validate the effectiveness of our chosen domain transfer methods, we compare the results with two baseline approaches:

- **Direct Testing with CPED-Pretrained MTSF Model (Direct Testing):** We directly apply the MTSF model trained on the CPED dataset to the newly collected NTUBI-Diag dataset for testing.
- **Training from Scratch with NTUBI-Diag Data (Training from Scratch):** We exclusively construct the MTSF model using the newly collected NTUBI-Diag

doi:10.6342/NTU202304293

dataset for both training and testing, and only consider the sentiment forecasting task.

The experimental results are presented in Table 5.12 and Table 5.13, showcasing the performance of each domain transfer method and the baselines. From the experimental results, it is evident that our proposed **Fine-tuning method** outperforms the baseline approaches on the newly collected NTUBI-Diag dataset. This method not only achieves better overall performance but also demonstrates more consistent performance across different sentiment categories. These results highlight the effectiveness of fine-tuning in transferring knowledge learned from the CPED dataset to the new domain.

On the other hand, the **DaNN method**, which utilizes unlabeled data from the new application domain for training, exhibits inferior performance compared to Fine-tuning and Training from Scratch, as expected. However, it still outperforms the Direct Testing baseline, indicating its ability to leverage the unlabeled data from the new domain to acquire more general knowledge for sentiment forecasting.

Table 5.12: *Domain Transfer Experiments (on NTUBI-Diag Dataset)*

Method	**Macro F1**	**Macro Precision**	**Macro Recall**
Direct Testing	38.41%	44.73%	40.48%
Training from Scratch	46.81%	48.85%	46.15%
DaNN	40.27%	41.81%	41.23%
Fine-tuning	**48.36%**	**49.65%**	**47.67%**

Table 5.13: *Domain Transfer Experiments Across Different Sentiment Polarities (on NTUBI-Diag Dataset)*

Method	**Negative**			**Positive**			**Neutral**		
	F1	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
Direct Testing	28.68%	**57.74%**	19.08%	32.79%	32.55%	33.04%	**53.75%**	43.88%	**69.33%**
Training from Scratch	**52.11%**	48.65%	56.11%	40.71%	**51.24%**	33.78%	47.59%	46.67%	48.55%
DaNN	51.38%	43.08%	**63.66%**	32.57%	37.84%	28.59%	36.84%	44.50%	31.43%
Fine-tuning	51.49%	51.38%	51.60%	**43.98%**	50.48%	**38.96%**	49.62%	**47.10%**	52.44%

Chapter 6

Conclusion

6.1 Conclusion

In this research, we have addressed the task of sentiment forecasting in dialogue and proposed a novel method called MTSF, which is based on the architecture of multi-task learning. Our proposed method incorporates multiple auxiliary tasks, including sentiment and dialogue act recognition within dialogue history sequences, as well as predicting the dialogue act of the upcoming utterance. With these auxiliary tasks, our method effectively extracts informative sentiment features by leveraging diverse sources of information, thereby optimizing the performance of the main sentiment forecasting task. The experimental results have demonstrated the superiority of our proposed MTSF method compared to two salient benchmarks. Moreover, through the ablation studies, we have extensively explored the contribution of each auxiliary task to the main sentiment forecasting task.

To ensure the real-world applicability of our model, we also collected a new dialogue dataset that simulates various common dialogue scenarios. After, considering different data annotation situations in real-world scenarios, including both labeled and unlabeled, we demonstrated different domain transfer methods. The results further validate the

doi:10.6342/NTU202304293

effectiveness of our proposed domain transfer methods, solidifying the potential for practical deployment in diverse application domains.

In conclusion, our research findings highlight the importance of multi-task learning and domain transfer techniques in the context of sentiment forecasting in dialogue. With our proposed model architecture and dataset, we have demonstrated the potential to achieve more accurate sentiment forecasting results in various practical dialogue system scenarios. Our research paves the way for developing more sophisticated sentiment analysis techniques in real-world applications and contributes to the advancement of dialogue systems for improved user experiences.

6.2 Future Works

In terms of future works, beyond forecasting the sentiment elicited in the user by system response, a more practical solution would involve generating candidate system responses first and subsequently performing sentiment forecasting. However, due to the current limitations stemming from insufficient data availability, training a response generation model that is suitably robust remains a challenge. Thus, we propose that future works could focus on this direction and try to overcome these limitations, to enhance the applicability of the task of sentiment forecasting in dialogue. Furthermore, future works may also delve into exploring the extent to which our proposed MTSF method's effectiveness is sensitive to the length of dialogue history.

In the pursuit of advancing sentiment forecasting in dialogue, there are also several promising avenues for further research and exploration. First, a fruitful direction involves investigating additional auxiliary tasks. By incorporating more diverse and relevant tasks,

doi:10.6342/NTU202304293

we can potentially uncover new insights and enhance the model's ability to capture sentiment-related information within dialogues. Another critical aspect to explore is the adoption of other advanced model architectures specifically tailored for this task. By delving into alternative architectures, we can better leverage the context and structure of dialogues to achieve more accurate sentiment forecasting results.

Furthermore, extending the scope of research from "sentiment forecasting" to "emotion forecasting" represents a significant opportunity. Fine-grained emotion forecasting task holds immense potential in enabling more nuanced understanding and analysis of sentiments within dialogues. In addition to sentiment and emotion forecasting, extending research into other sentiment-related tasks within dialogues can also be fruitful. For instance, exploring such research tasks as "emotion-cause pair extraction in dialogues" or "quit intention detection in dialogues" would offer valuable insights and contribute to a deeper understanding of sentiment dynamics in the conversational context. By addressing these research directions, we can push the boundaries of sentiment analysis in dialogues and pave a way for more sophisticated and context-aware dialogue systems with enhanced sentiment forecasting and understanding capabilities.

References

Abouzeid, A., Granmo, O.-C., and Goodwin, M. (2021). Modelling emotion dynamics in chatbots with neural hawkes processes. In Proceedings of International Conference on Innovative Techniques and Applications of Artificial Intelligence, pages 146-151. Springer.

Adam, E. (2019). Improving the user experience of chatbots with tone analysis. Unpublished Thesis, Department of Computer Science and Business Information Systems, Karlsruhe University of Applied Sciences, Karlsruhe, Baden-Württemberg, Germany.

Belharbi, S., Hérault, R., Chatelain, C., and Adam, S. (2016). Deep multi-task learning with evolving weights. In Proceedings of European Symposium on Artificial Neural Networks (ESANN).

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. Language Resources and Evaluation, 42(4):335-359.

Cao, S., Qu, L., and Tian, L. (2021). Causal relationships between emotions and dialog acts. In Proceedings of 9th International Conference on Affective Computing and Intelligent Interaction (ACII), pages 1-8. IEEE.

Chen, Y., Fan, W., Xing, X., Pang, J., Huang, M., Han, W., Tie, Q., and Xu, X. (2022). CPED: A large-scale Chinese personalized and emotional dialogue dataset for conversational ai. arXiv preprint arXiv:2205.14727.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Ganin, Y. and Lempitsky, V. (2015). Unsupervised domain adaptation by backpropagation. In Proceedings of International Conference on Machine Learning, pages 1180-1189. PMLR.

Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., Marchand, M., and Lempitsky, V. (2016). Domain-adversarial training of neural networks. Journal of Machine Learning Research, 17(1):2096-2030.

Guibon, G., Labeau, M., Flamein, H., Lefeuvre, L., and Clavel, C. (2021). Few-shot emotion recognition in conversation with sequential prototypical networks. arXiv preprint arXiv:2109.09366.

He, W., Dai, Y., Zheng, Y., Wu, Y., Cao, Z., Liu, D., Jiang, P., Yang, M., Huang, F., Si, L., et al. (2022). Galaxy: A generative pre-trained model for task-oriented dialog with semi-supervised learning and explicit policy injection. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 10749-10757.

He, Z., Tavabi, L., Lerman, K., and Soleymani, M. (2021). Speaker turn modeling for dialogue act classification. arXiv preprint arXiv:2109.05056.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8):1735-1780.

Kim, M. and Kim, H. (2018). Integrated neural network model for identifying speech acts, predicators, and sentiments of dialogue utterances. Pattern Recognition Letters, 101:1-5.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kongyoung, S., Macdonald, C., and Ounis, I. (2020). Multi-task learning using dynamic task weighting for conversational question answering. In Proceedings of the 5th International Workshop on Search-Oriented Conversational AI (SCAI), pages 17-26.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278-2324.

Liu, B. (2020). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge University Press.

Liu, S., Liang, Y., and Gitter, A. (2019a). Loss-balanced task weighting to reduce negative transfer in multi-task learning. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 33, pages 9977-9978.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019b). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

Naar, H. (2018). Sentiments. In The Ontology of Emotions. Cambridge University Press.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. In Proceedings of 33rd Conference on Neural Information Processing Systems.

Peng, B., Zhu, C., Li, C., Li, X., Li, J., Zeng, M., and Gao, J. (2020). Few-shot natural language generation for task-oriented dialog. arXiv preprint arXiv:2002.12328.

Ramponi, A. and Plank, B. (2020). Neural unsupervised domain adaptation in NLP—A survey. arXiv preprint arXiv:2006.00632.

Saha, T., Gupta, D., Saha, S., and Bhattacharyya, P. (2021). Emotion aided dialogue act classification for task-independent conversations in a multi-modal framework. Cognitive Computation, 13(2):277-289.

Salmeron, J. L. (2012). Fuzzy cognitive maps for artificial emotions forecasting. Applied Soft Computing, 12(12):3704-3710.

Shahriar, S. and Kim, Y. (2019). Audio-visual emotion forecasting: Characterizing and predicting future emotion using deep learning. In Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), pages 1-7. IEEE.

Shi, X., Li, S., and Dang, J. (2020). Dimensional emotion prediction based on

doi:10.6342/NTU202304293

interactive context in conversation. In Proceedings of INTERSPEECH Conference, pages 4193-4197.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017).

Wang, T., Hou, Y., Zhou, D., and Zhang, Q. (2021). A contextual attention network for multimodal emotion recognition in conversation. In Proceedings of 2021 International Joint Conference on Neural Networks (IJCNN), pages 1-7. IEEE.

Wang, Z., Zhu, X., Zhang, Y., Li, S., and Zhou, G. (2020). Sentiment forecasting in dialog. In Proceedings of the 28th International Conference on Computational Linguistics, pages 2448-2458.

Wen, Z., Cao, J., Yang, R., Liu, S., and Shen, J. (2021). Automatically select emotion for response via personality-affected emotion transition. In Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 1-12.

Zhang, J., Gan, R., Wang, J., Zhang, Y., Zhang, L., Yang, P., Gao, X., Wu, Z., Dong, X., He, J., Zhuo, J., Yang, Q., Huang, Y., Li, X., Wu, Y., Lu, J., Zhu, X., Chen, W., Han, T., Pan, K., Wang, R., Wang, H., Wu, X., Zeng, Z., and Chen, C. (2022). Fengshenbang 1.0: Being the foundation of Chinese cognitive intelligence. arXiv preprint arXiv:2209.02970.

Zou, C., Yin, Y., and Huang, F. (2022). Attention-based emotion-assisted sentiment forecasting in dialogue. In Proceedings of 2022 International Joint Conference on Neural Networks (IJCNN), pages 1-8. IEEE.

doi:10.6342/NTU202304293