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's Thesis

# 運用多任務方法進行對話失焦與脫身意圖偵測

# MAD-detect: A Multi-task Approach for Dialogue Disengagement Detection

莊承叡

Cheng-Ruei Chuang

指導教授：魏志平 博士

Advisor: Chih-Ping Wei, Ph.D.

中華民國 113 年 7 月

July, 2024

doi:10.6342/NTU202403676

# 誌謝

首先，我想要感謝魏志平老師，帶領我完成這個研究，當我因為研究遇到的困難而徘徊時，每次與老師討論完都能馬上找到方向，謝謝您無論有多忙碌總是能騰出時間與我們開會，還會找我們去打球運動，這幫助我們在漫長的研究時間中能維持良好的心態去繼續完成剩下的任務，我認為自己在這段時間內從老師身上學習到的遠不止於此。還要感謝口試委員楊錦生教授、胡雅涵教授遠從桃園前來，在口試時提供許多寶貴的意見，讓本篇論文可以更加完善。

接下來要感謝的是實驗室的同學們：大禾、子寬、心瑋、啟宏、苡菱、品君，謝謝你們營造了良好的實驗室氛圍，正是因為我們每天早九晚十二的在教研館唱歌做研究、打球，一起買飲料、健身運動，才好好地充實豐富這段每天不斷重複的生活日常，也感謝冠均在後來加入我們一起研究，分享他的進度和經驗，和我在最後階段好幾次待到冷氣都關了，還有梓旭來找我玩遊戲跟吃飯，以及家瑋在口試前聽我練習好幾次，讓我在最後口試時能夠安心上場，也謝謝虹鈞協助處理實驗室的大小事，最後感謝 BI Lab 的學弟妹們，尤其是幫助我收集標注資料的日明，要不是有他完整又仔細的規劃，這項研究沒辦法完成。

此外，我想要感謝我的家人，你們總是給我許多鼓勵和溫暖，讓我知道無論我最後做得如何，你們都會在我身邊支持我，並關心我的近況，為了讓我能更順利的完成體諒我帶來的許多不便。感謝這一路走來的所有好朋友、同學們，是你們幫助我慢慢拼湊出這個成果，也謝謝撐過這些的自己，每天規律地完成自己的目標，希望能好好保留這段回憶！

莊承叡 謹誌

于國立臺灣大學資訊管理研究所

中華民國一一三年七月

doi:10.6342/NTU202403676

# 摘要

對話系統已成為企業不可或缺的工具，然而維持使用者與系統之間高品質的互動仍是一個重大挑戰。雖然現有的對話系統提升了回應速度和效率，但其無法偵測和應對使用者是否參與在其中可能會對系統的效用產生負面的影響，是影響整體使用者體驗和滿意度的重要因素，倘若能在對話進行的過程中及時偵測到使用者的對話開始失焦並意圖脫離，就能使對話系統快速地作出相應的策略進行調整，改善整體對話的體驗。

因此在本研究中，我們的目標是提出一個有效的對話失焦與使用者脫身意圖偵測模型，並利用多任務學習方法，將對話行為分類和情感識別作為輔助任務試圖提升模型在主任務上的表現，這種多任務框架不僅提高了脫離偵測的準確性，還透過共享相關任務的學習結構，提供了對使用者在對話中行為更多的理解。在實驗過程中，我們提出一個特殊的兩階段訓練策略：初始階段專注於輔助任務，隨後整合主任務進行第二階段的訓練。實驗結果顯示，我們提出的多任務學習模型加上兩階段訓練策略的表現，在偵測使用者產生對話失焦和意圖脫身的行為上表現出色，這一方法證明了專注於輔助任務在提升主任務能力方面的有效性。除此之外，我們創建了一個中文對話資料集，模擬真實世界場景，確保研究的實用性，該資料集包括明確的定義與標註指南，為未來持續擴增和對話系統研究提供有用的資源。

總之，本論文通過新穎的多任務學習方法推進了對話失焦與使用者脫身意圖偵測的領域，提供了一個能夠滿足現實場景中用戶多樣化需求的強大解決方案。我們的研究為未來發展奠定了堅實基礎，有助於創建更具回應性和適應性的對話系統。

關鍵字：對話系統、對話失焦與脫身意圖偵測、情感分析、對話行為分類、多任務學習、兩階段訓練策略

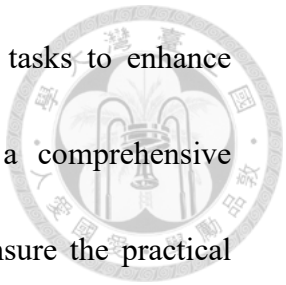
doi:10.6342/NTU202403676

# Abstract

Dialogue systems have become indispensable tools for businesses; however, maintaining high-quality interactions between users and these systems remains a significant challenge. While current dialogue systems have improved response speed and efficiency, their inability to detect and respond to user engagement can negatively impact their effectiveness, influencing overall user experience and satisfaction. Timely detection of users' disengagement intention during a conversation can enable dialogue systems to quickly adjust dialogue strategies, thereby enhancing the overall experience in systems.

This research aims to develop an effective model for detecting user disengagement in dialogues, employing a multi-task learning approach, MAD-detect. By incorporating dialogue act classification and sentiment recognition as auxiliary tasks, we seek to improve the model's performance on the main task, disengagement detection. This multi-task framework not only enhances the performance of disengagement detection but also provides a deeper understanding of user behavior by sharing learned structures across related tasks. We propose a unique two-stage training strategy: the initial phase focuses on auxiliary tasks, followed by the integration of the primary task in the second phase. Experimental results demonstrate that MAD-detect combined with the two-stage training strategy outperforms baseline models, achieving higher recall and balanced.

doi:10.6342/NTU202403676



This approach proves the effectiveness of emphasizing auxiliary tasks to enhance disengagement detection capabilities. Furthermore, we created a comprehensive Chinese dialogue dataset that simulates real-world scenarios to ensure the practical applicability of our research. This dataset includes clear definitions and annotation guidelines, making it a valuable resource for future expansions and dialogue system research.

In summary, this research advances the field of dialogue disengagement detection through an innovative multi-task learning approach, offering a robust solution that meets diverse user needs in real-world scenarios. Our research lays a solid foundation for future developments, contributing to the creation of more responsive and adaptable dialogue systems.

**Keywords:** Dialogue System, Disengagement Detection, Sentiment recognition, Dialogue Act Classification, Multi-task Learning, Two-stage Training Strategy

doi:10.6342/NTU202403676

# Table of Contents





doi:10.6342/NTU202403676

doi:10.6342/NTU202403676

# List of Figures





doi:10.6342/NTU202403676

# List of Tables





doi:10.6342/NTU202403676

# Chapter 1 Introduction

## 1.1 Background

Dialogue systems have become indispensable tools for companies, serving as key components in understanding and engaging with customers in today's dynamic business landscape. These systems are pivotal across various sectors, from customer service platforms to virtual assistants in tourism, reshaping how businesses respond to user queries and directly influencing customer satisfaction, thus positioning these systems as strategic assets (Adam et al., 2021; Zhang et al., 2024).

As companies increasingly rely on dialogue systems, the quality and efficacy of interactions emerge as primary concerns. Antonio et al. (2022) explored the impact of integrating chatbots into e-commerce customer service systems on customer satisfaction, addressing the challenge of providing round-the-clock support within human resource constraints. Nordheim et al. (2019) emphasized the importance of user trust in chatbots for effective customer service, proposing a model of user trust that includes factors such as perceived expertise, responsiveness, and environmental aspects like risk and brand perceptions. Jiang et al. (2022) highlighted the significant influence of responsiveness and conversational tone on customer satisfaction, demonstrating their effects on user experiences, purchase intentions, and willingness to pay price premiums.



doi:10.6342/NTU202403676

Despite the benefits of dialogue systems, there are potential drawbacks. Antonio et al. (2022) identified issues such as unsatisfactory responses and the perception of inhuman behavior, which can negatively impact customer satisfaction. Similarly, Deng, Lei, et al. (2023) pointed out that inappropriate responses can erode user trust, foster dissatisfaction, and diminish brand loyalty. Therefore, user satisfaction is an important metric for evaluating dialogue systems, aiming to enhance overall user experience.

However, the quality of a dialogue system is reflected not only in its response speed and accuracy but also in its ability to maintain user engagement. Within this context, the detection of engagement in open-domain dialogue systems is a critical aspect. Engagement represents the system's ability to captivate users' attention, fostering positive and enjoyable interactions and making users more willing to communicate with the dialogue system. Rather than relying on subjective human judgments, using automatic dialogue evaluation metrics to measure engagement has been suggested as a robust approach, qualifying how a dialogue system responds to users (Ghazarian et al., 2020). Detecting user disengagement has accumulated significant attention due to its potential negative effects on commercial applications (Forbes-Riley et al., 2012). In daily conversation, dialogues do not necessarily conclude when both speakers wish to end the conversation, leading to emotional fatigue and a decline in conversation quality (Mastroianni et al., 2021). Detecting disengagement has



doi:10.6342/NTU202403676

become a crucial factor in both human-agent and human-human interactions, as both speakers should actively engage in the conversation.

NTUBI-Diag Dataset (Excerpt from Dialogue #0027)

| Turn | Speaker | Utterance | Disengaged |
|---|---|---|---|
| #13 | 語者一 | 我要昨晚的住房費用全額退費 | No |
| | Speaker 1 | I want a full refund for last night's accommodation fee. | |
| #14 | 語者二 | 抱歉，這部分可能沒有辦法哦 | No |
| | Speaker 2 | Sorry, it may not be possible. | |
| | 語者二 | 因為這部分真的不屬於我們旅館的責任歸屬 | No |
| | Speaker 2 | Because this situation really does not belong to the responsibility of our hotel. | |
| #15 | 語者一 | 那這樣我該找誰求償 | No |
| | Speaker 1 | Then who should I ask for compensation? | |
| | 語者一 | 還是我就只能摸摸鼻子自認倒霉？ | No |
| | Speaker 1 | Or should I just consider myself unlucky?. | |
| #16 | 語者二 | 這部分我們會幫您跟隔壁反應 | No |
| | Speaker 2 | We will help you react to your neighbor. | |
| | 語者二 | 希望今天不會遇到一樣的問題 | No |
| | Speaker 2 | I hope you won't encounter the same problem today. | |
| | 語者二 | 很抱歉造成您的不滿 | No |
| | Speaker 2 | Sorry for causing your dissatisfaction. | |
| #17 | 語者一 | 算了我覺得我們也只是在鬼打牆 | Yes |
| | Speaker 1 | Forget it, I think we are just going around in circles. | |
| | 語者一 | 反正我以後是不會來住這裡了 | Yes |
| | Speaker 1 | Anyway, I won't stay here again. | |

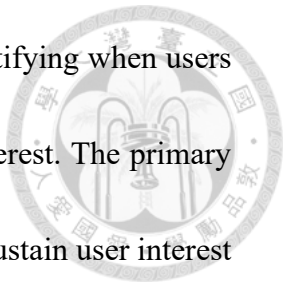
**Figure 1:** *Scenario of Disengagement in Dialogue*

Previous studies on dialogue systems have primarily concentrated on two main approaches, each addressing different aspects of dialogue interaction to enhance system performance and user experience: dialogue breakdown detection and dialogue disengagement detection.

In the domain of dialogue breakdown detection, the objective is to identify inappropriate utterances that cause the dialogue system to fail (Higashinaka et al., 2016). However, this approach faces significant challenges in comprehensively detecting all situations and reasons that lead to dialogue breakdowns. The inherent complexity of human language and the contexts in which dialogues occur make it difficult to anticipate and address every potential breakdown scenario effectively.



doi:10.6342/NTU202403676



Conversely, dialogue disengagement detection focuses on identifying when users intend to exit the ongoing dialogue, signifying a decline in user interest. The primary goal of this approach is to enhance the overall user experience and sustain user interest throughout the interaction (Ghazarian et al., 2020). Recognizing signs of disengagement enables dialogue systems to adapt their strategies to re-engage users, thereby improving user satisfaction and the effectiveness of the interaction. (Figure 1)

## 1.2 Research Motivation

Detecting engagement and disengagement in dialogue systems holds significant importance and offers numerous advantages across various applications. One of the primary benefits is the ability to adjust dialogue strategies in real-time. When a system detects signs of user disengagement, it can proactively change the topic, enhance interactivity, or provide more engaging content to recapture the user's interest. This responsive adaptation ensures that the dialogue remains engaging and relevant, maintaining the user's participation and interest.

In counseling and psychological support, the ability to detect engagement levels is particularly critical. For online psychological counseling, understanding a user's engagement can help counselors better grasp the user's emotional state, allowing them to tailor their strategies accordingly. Engagement detection enables these systems to convey empathy effectively, which is essential for providing meaningful emotional



doi:10.6342/NTU202403676

support (Deng, Zhang, et al., 2023). Proactively assisting users in exploring and addressing their problems becomes more efficient when the system can accurately assess and respond to their engagement levels.

Building on existing techniques and insights in dialogue breakdown detection, various research efforts have focused on utilizing additional features. For instance, Sugiyama (2015) considered dialogue act annotated corpora as input features, and Matsumoto et al. (2022) explored emotion analysis in dialogue by calculating the similarity between utterances. Studies on detecting disengagement have shown that dialogue acts (DA) correlate with expert judgments on engagement and provide valuable insights when used as weak labels for disengagement detection (Liang et al., 2021). The work of Liang et al. (2021) demonstrated the effectiveness of using dialogue acts as indicators of user engagement levels. Most of these studies employ English and Japanese datasets and often treat these tasks as single-task scenarios.

Our research aims to extend these existing techniques by incorporating additional tasks such as sentiment recognition and dialogue act classification. These tasks have shown promise in related works and offer a more comprehensive understanding of user engagement. By leveraging multi-task learning, our approach seeks to enhance the performance of disengagement detection by sharing the learned structure across multiple related tasks. This multi-task model not only improves detection accuracy but



doi:10.6342/NTU202403676

also underscores the benefits of integrating auxiliary tasks in dialogue systems, ultimately contributing to a more engaging and satisfying user experience.

## 1.3 Research Objective

In this paper, we aim to detect user disengagement in text-based dialogue systems, representing an indicator that users intend to exit the ongoing dialogue, which poses a potential threat to the overall effectiveness of dialogue systems. Identifying and understanding how and when disengagement occurs is crucial for system improvement, enabling proactive measures to maintain user interest and enhance the overall user experience.

Our primary objective is to develop a model for detecting user disengagement in dialogue systems, addressing a critical gap in current dialogue system evaluation and improvement efforts. To achieve this, we propose a multi-task model, MAD-detect, which incorporates dialogue acts classification and sentiment recognition as auxiliary tasks. This approach leverages the complementary strengths of these auxiliary tasks to enhance both the accuracy and interpretability of disengagement detection.

Integrating dialogue acts classification and sentiment recognition can provide a more explicit understanding of the motivations behind user disengagement. This allows the model to capture characteristics of engagement that may be overlooked by single-task approaches. By understanding the interplay between user sentiment and dialogue



doi:10.6342/NTU202403676

acts, our model can more effectively identify and respond to signs of disengagement, thereby improving the overall user experience.

In addition to developing the MAD-detect model, we have created a Chinese dialogue dataset that closely simulates real-world dialogue scenarios. This dataset includes clear guidelines for future expansions and annotations, ensuring its relevance and utility for ongoing research and practical applications. By focusing on real-world scenarios, our dataset aims to reflect the diverse and changing needs of speakers in open-domain topics, enhancing the applicability of our model across various application domains.

We believe that this approach will significantly advance the quality and effectiveness of dialogue systems. By addressing the complexities of user engagement through a multi-task framework, our research aims to provide a robust solution that meets the diverse needs of users in practical, real-world scenarios. Ultimately, this will contribute to the development of more responsive and adaptable dialogue systems, capable of delivering superior user experiences across different contexts.



doi:10.6342/NTU202403676

# Chapter 2 Related Works

## 2.1 Dialogue Breakdown Detection

In Japan, dialogue breakdown detection has gathered significant attention, particularly through the Dialogue Breakdown Detection Challenge established by Higashinaka et al. (2016). This challenge categorized dialogue breakdowns into three levels:

- Not a Breakdown (NB): It is easy to continue the conversation.
- Possible Breakdown (PB): It is difficult to continue the conversation smoothly.
- Breakdown (B): It is difficult to continue the conversation.

To address the task of identifying inappropriate utterances in dialogue systems, Higashinaka et al. (2016) organized this challenge and leveraged an evaluation workshop. However, defining whether a dialogue is "likely" to break down faced varying subjective human opinions, highlighting the difficulty in achieving definitive assessments. For this reason, a group of annotators provided collective judgments. Despite this approach bringing together various methods, dialogue breakdowns can occur in diverse ways, such as misinterpreting user intent or failing to provide contextually relevant information, making it challenging to comprehensively detect all causes of system breakdowns.



doi:10.6342/NTU202403676

Various methods have been employed to predict whether an utterance might cause a dialogue breakdown, particularly using deep neural networks (DNN). Sugiyama (2015) proposed a DNN-based method incorporating extensive external knowledge as additional features, such as dialogue-act annotated corpora and question-answer databases, to enhance prediction accuracy. Inaba and Takahashi (2015) introduced an approach utilizing Long Short-Term Memory (LSTM) networks. Their method uses word embeddings generated by word2vec and processes user and system utterances separately, allowing for a more implicit understanding of interaction dynamics and potential breakdown points. Despite improvements in detection and annotation distribution estimation, challenges persisted, particularly in detecting certain breakdown types where it is difficult to assess the likelihood of a breakdown. These works highlighted the potential of neural network architectures for dialogue breakdown detection but underscored the need for further refinement.

Furthermore, Matsumoto et al. (2022) identified the lack of awareness of emotional changes or patterns as a significant factor contributing to dialogue breakdowns. They explored emotion analysis as a complementary feature to dialogue breakdown detection, utilizing deep neural networks and distributed representation vectors, such as emotion similarity vectors and utterance similarity vectors. This approach demonstrated superior performance compared to methods relying solely on



doi:10.6342/NTU202403676

utterance embeddings. The study also outlined future research directions, including the relationship between emotional understanding and dialogue breakdowns.

In conclusion, the existing research landscape in dialogue breakdown detection showcases a progression from challenge initiation to the exploration of advanced neural network architectures and the incorporation of emotion analysis. However, opportunities for further refinement remain, particularly in detecting severe breakdowns, extending methods to different linguistic and modal contexts, and exploring the relationship between emotion and dialogue breakdowns. Recognizing the differentiation of severe breakdown labels caused by a lack of understanding as a critical trigger for disengagement, future work can elaborate on the concept of emotional changes in dialogue systems.

## 2.2 Dialogue Disengagement Detection

Glas and Pelachaud (2015) provide a comprehensive overview of engagement definitions in human-agent interaction, emphasizing the diverse features explored in prior research on user-system communication. Among these studies, Yu et al. (2004) defined engagement as "capturing a participant's interest and attentiveness in a conversation," which has served as a guiding principle for our subsequent work.

Researchers have utilized various methods to assess user engagement in systems, including behavioral metrics, neurophysiological techniques, and self-reports (O'Brien



doi:10.6342/NTU202403676

et al., 2018). Despite this, most existing literature on engagement and disengagement detection emphasizes multimodal approaches incorporating audio and video data. For instance, Fedotov et al. (2018) developed a multimodal regression model that leverages speech, facial expressions, body language, lip movements, and eye movements to determine engagement levels. Similarly, Hsiao et al. (2012) employed external sensors to capture nonverbal behaviors and recognize social engagement in face-to-face conversations.

Previous research has shown that engagement with a dialogue system reflects user satisfaction at the session level and is a favorable indicator of users' willingness to interact at the turn level (Ma, 2018; Yi et al., 2019). Traditional methods for measuring engagement in text-based dialogue systems include human ratings and proxy metrics such as the number of dialogue turns and topical diversity. However, Ghazarian et al. (2020) highlighted the limitations of these metrics, advocating for predictive engagement, which focuses on utterance-level engagement estimation. This shift from conversation-level to utterance-level engagement metrics allows for real-time feedback in dialogue model training and represents a move towards more sophisticated and precise evaluation methods. Figure 2 demonstrates the general structure of models used in dialogue disengagement detection from input encoding to final classification. Starting with the input layer that choose a window of dialogue history before utterance



doi:10.6342/NTU202403676

n, followed by encoding layers and pooling layers that distill the representation of each utterance, fusion layers that integrate the information in this window, and finally, classification layers that predict the engagement state.

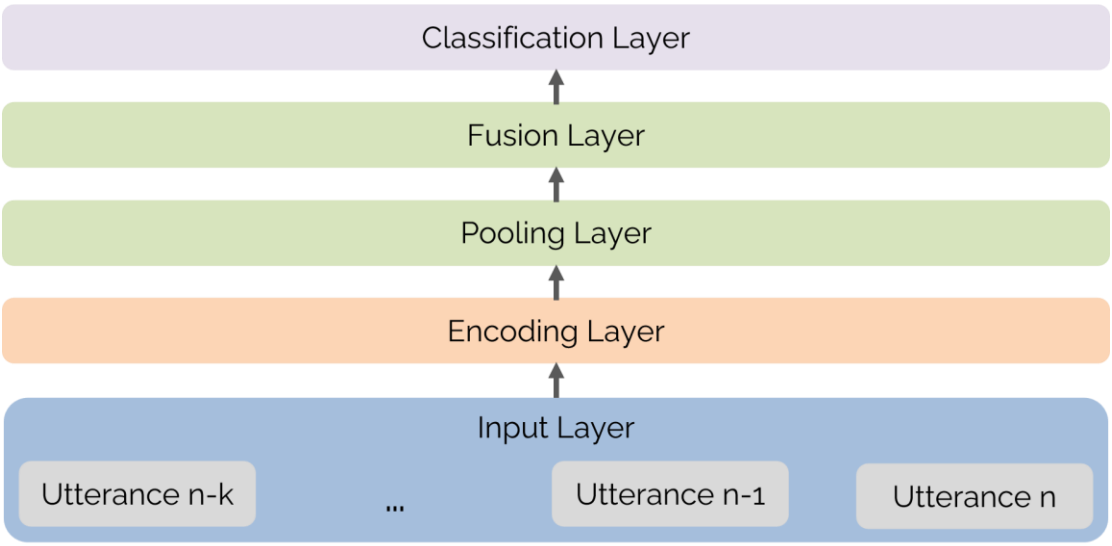



**Figure 2:** *General Architecture of Disengagement Detection in Dialogue*

In parallel, Liang et al. (2021) introduced HERALD, an annotation-efficient method for detecting user disengagement in social conversations. HERALD redefines the annotation process as a denoising problem within a two-stage pipeline. In the first stage, heuristic regular expressions and dialogue act classification provide weak labeling. The second stage employs the Shapley algorithm for denoising. This pipeline significantly reduces the manual labeling burden by leveraging machine learning, inspiring the potential integration of dialogue acts in future models. Notably, the model, which learns from weak labels automatically assigned by heuristic Regexes and DA classification, outperforms models trained on a limited number of clean labels. This has motivated us



doi:10.6342/NTU202403676

to consider dialogue acts as a helpful component related to disengagement and use it as an auxiliary task in our proposed model.

Jiang et al. (2023) focused on turn-level engagingness evaluation, emphasizing the importance of continuous user engagement monitoring throughout a dialogue session. They introduced WeSEE, a weakly supervised engagingness evaluator that uses the remaining depth for each turn as a heuristic weak label for engagingness. By framing engagingness prediction as a regression task with automatically generated labels, WeSEE eliminates the need for human annotations. The results showcase the effectiveness of leveraging implicit signals in multi-turn dialogue data. However, the weak label's characteristic tends to direct the model's attention to utterances that mark the beginning or end of dialogues, which does not always translate well to handling general responses.

Table 1 shows that previous studies on detecting dialogue breakdown and disengagement have established a solid foundation for future research. These studies primarily utilize single-task learning approaches with additional features, and recent works concentrate on reducing annotation costs. The focus has been on improving efficiency and leveraging various forms of additional features to enhance the accuracy of engagement detection. Despite these advancements, existing methods exhibit certain limitations. They often treat disengagement detection as isolated tasks, relying heavily



doi:10.6342/NTU202403676

on additional features without fully integrating auxiliary tasks that could provide a more implicit understanding of user behavior. While effective in certain contexts, these single-task approaches may overlook the complex interaction of factors contributing to user disengagement.

**Table 1:** *Summary of Previous Studies*

| Reference | Task | Method | Additional features |
|---|---|---|---|
| Sugiyama (2015) | Dialogue Breakdown Detection Single-task learning | Word2Vec + DNN | • Dialogue-act annotated corpus<br>• Q&A database |
| Inaba and Takahashi (2015) | Dialogue Breakdown Detection Single-task learning | Word2vec+ BiLSTM | |
| Matsumoto et al. (2022) | Dialogue Breakdown Detection Single-task learning | Sentence2Vec + Logistic Regression | • Emotion Similarity<br>• Utterance Similarity |
| Ghazarian et al. (2020) | Dialogue Disengagement Detection Single-task learning | BERT+SVM | |
| Liang et al. (2021) | Dialogue Disengagement Detection Single-task learning | BERT | • Heuristic Regexes<br>• Dialogue act classification<br>• Sharpley denoising |
| Jiang et al. (2023) | Dialogue Disengagement Detection Single-task learning | BERT | • Remaining dialogue depth |
| This study | Dialogue Disengagement Detection **Multi-task learning** | BERT+BiLSTM | • Sentiment<br>• Dialogue Act |

In contrast to these existing methods, our study aims to address these gaps by designing a multi-task learning model. This model seeks to provide a deeper comprehension of why users disengage from conversations. By incorporating additional features such as sentiment recognition and dialogue acts classification as auxiliary tasks, our approach offers a more comprehensive analysis of engagement dynamics. The integration of sentiment and DA as auxiliary tasks allows for a richer representation of the dialogue context, capturing subtle cues and patterns that single-



doi:10.6342/NTU202403676

task models might miss. This multi-task framework not only enhances the accuracy of disengagement detection but also improves the interpretability of the results, providing clearer insights into the underlying reasons for user disengagement.

Through this innovative approach, our research aims to advance the field of dialogue engagement and disengagement detection, contributing to the development of more effective and user-centric dialogue systems.



doi:10.6342/NTU202403676

# Chapter 3 Methodology

## 3.1 Problem Formulation

We define the problem of detecting disengagement in dialogues as follows: Given a window of dialogue history as a sequence of utterances $U = [u_1, u_2, \dots, u_n]$, where each $u_i$ represents the $i$-th utterance in this dialogue history, with a maximum size $n$, restricting it to contain at most $n$ utterances. Each utterance $u_i$ has a corresponding speaker $s_i$ from the list of speakers $S = [s_1, s_2, \dots, s_n]$. Utterances are grouped into the window based on turns $T = [t_1, t_2, \dots t_m]$, where each $t_i$ represents consecutive utterances spoken by the same speaker $s_i$. These turns are included in the window until adding the next turn would exceed the window size. Each window needs to contain at least three turns, as fewer turns do not provide sufficient information for accurate analysis and will not be used.

| | Turn 1 | | Turn 2 | | | Turn 3 | | Turn 4 | | Turn 5 | | | Turn 6 | Turn 7 | Turn 8 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Window 1 | $s_1$ | $s_1$ | $s_2$ | $s_2$ | $s_2$ | $s_1$ | $s_1$ | $s_2$ | $s_2$ | $s_1$ | $s_1$ | $s_1$ | $s_2$ | $s_1$ | $s_2$ | $s_2$ |
| Window 2 | $s_1$ | $s_1$ | $s_2$ | $s_2$ | $s_2$ | $s_1$ | $s_1$ | $s_2$ | $s_2$ | $s_1$ | $s_1$ | $s_1$ | $s_2$ | $s_1$ | $s_2$ | $s_2$ |
| Window 3 | $s_1$ | $s_1$ | $s_2$ | $s_2$ | $s_2$ | $s_1$ | $s_1$ | $s_2$ | $s_2$ | $s_1$ | $s_1$ | $s_1$ | $s_2$ | $s_1$ | $s_2$ | $s_2$ |
| Window 4 | $s_1$ | $s_1$ | $s_2$ | $s_2$ | $s_2$ | $s_1$ | $s_1$ | $s_2$ | $s_2$ | $s_1$ | $s_1$ | $s_1$ | $s_2$ | $s_1$ | $s_2$ | $s_2$ |
| Window 5 | $s_1$ | $s_1$ | $s_2$ | $s_2$ | $s_2$ | $s_1$ | $s_1$ | $s_2$ | $s_2$ | $s_1$ | $s_1$ | $s_1$ | $s_2$ | $s_1$ | $s_2$ | $s_2$ |

window = 10

**Figure 3:** *Illustration of Generating Dialogue History Window*

The objective of this task is to detect disengagement in the dialogue, classified as either 0 (engaged) or 1 (disengaged), based on the turns $t_i \in T$. This classification aims to identify whether the user is engaged or disengaged at any given point within the



doi:10.6342/NTU202403676

dialogue window, providing valuable insights for improving dialogue system performance and user satisfaction.

## 3.2 Overview of MAD-detect Architecture

We propose a **M**ulti-task **A**pproach for dialogue **Di**sengagement **detect**ion (MAD-detect), treating it as a multi-task problem that leverages the interrelated nature of disengagement, sentiment, and dialogue acts. Our model is designed to handle one main task and two auxiliary tasks.

The primary objective of the model is disengagement detection, which aims to identify whether a speaker is engaged or disengaged. The classification labels for this task are binary, where 0 represents engaged turns and 1 represents disengaged turns. The main task operates at the turn level, represented by the sequence of turns $T = [t_1, t_2, \dots, t_m]$, with $m \leq n$, where n stands for the number of utterances in the history window.

The first auxiliary task involves sentiment recognition, focusing on determining the sentiment polarity of each utterance, classifying them as positive, negative, or neutral. Sentiment recognition is performed at the utterance level, represented by the sequence $U = [u_1, u_2, \dots, u_n]$. The second auxiliary task is dialogue act classification, which involves classifying each utterance into specific dialogue acts such as questions, greetings, command-and-requests, etc. Similar to sentiment recognition, dialogue act



doi:10.6342/NTU202403676

classification is performed at the utterance level, using the same sequence $U = [u_1, u_2, \dots, u_n]$.

The objective of incorporating auxiliary tasks into MAD-detect is to enhance the performance of the main task, disengagement detection, by utilizing a shared learning structure across related tasks. The rationale for combining these tasks lies in the interdependence between sentiment and dialogue acts, as both are influential factors in determining engagement levels in dialogue. By leveraging these relationships, MAD-detect aims to provide a more comprehensive and accurate model for detecting speaker disengagement in dialogues. The architecture of our proposed model is illustrated in Figure 4. It begins with the utterance encoder, which extracts the representation of each utterance. Next, speaker embedding concatenation integrates the speaker's features, resulting in speaker-aware representations. A BiLSTM layer follows, understanding the flow of the dialogue over utterances and generating utterance-level representations. The turn pooling layer then leverages these utterance-level representations to create turn-level representations. Finally, the multi-task classification layers predict the auxiliary tasks based on the utterance-level representations, while the main task is predicted based on the turn-level representations.



doi:10.6342/NTU202403676

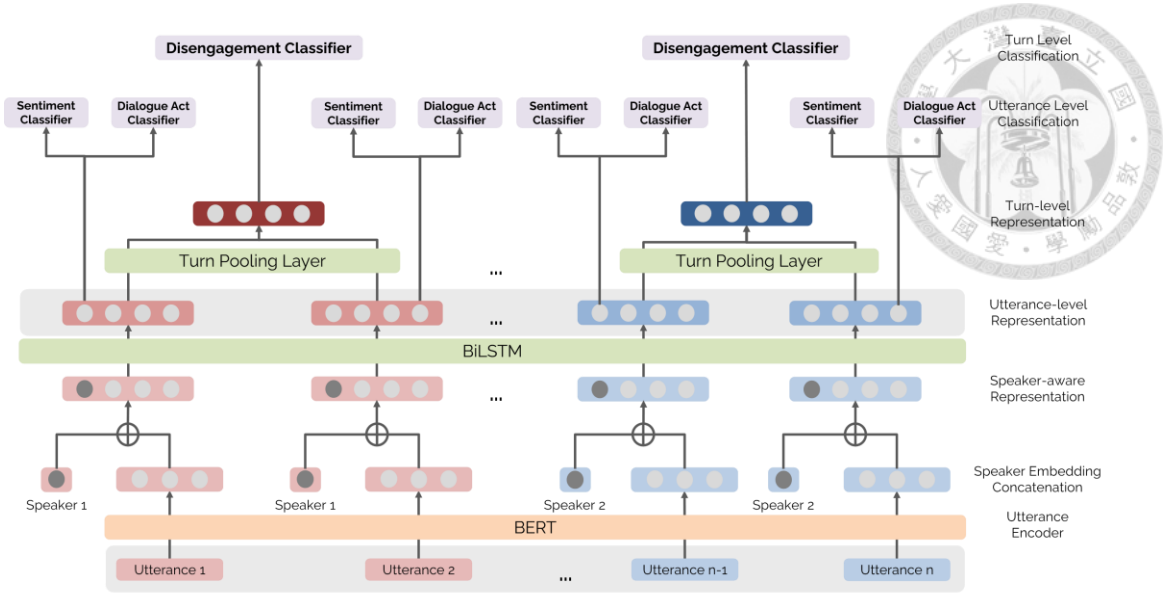



**Figure 4:** *Architecture of Proposed Multi-task Model* ***MAD-detect***

## 3.3 Utterance Encoder

We employ the *bert-base-chinese* model, a pre-trained BERT model specifically optimized for the Chinese language, to obtain utterance embeddings. This model is utilized due to its robust performance in capturing the semantics of Chinese text. To extract meaningful and overall embeddings for each utterance, we aggregate the word embeddings by applying mean pooling. This process generates a single vector representation for each utterance, encapsulating its semantic content in a concise form.

Given a window of dialogue utterances $[u_1, u_2, \ldots, u_n]$, each utterance $u_i$, is initially transformed into a sequence of word embeddings $[w_{i1}, w_{i2}, \ldots, w_{ik_i}]$, where $k_i$ represents the number of words in utterance $u_i$. The transformation captures the complex word-level details necessary for understanding the utterance. The next step



doi:10.6342/NTU202403676

involves applying mean pooling to these word embeddings to derive a single utterance embedding $e_i$ for each utterance $u_i$. The formula for this process is given by:

$$e(u_i) = e\left(\left[w_{i1}, w_{i2}, \dots, w_{ik_i}\right]\right) = \frac{1}{k_i}\sum_{j=1}^{k_i} w_{ij} \tag{3.1}$$

This mean pooling operation ensures that the resulting utterance embedding $e_i$ encapsulates the overall semantic information of the utterance, considering all words equally. The final set of embeddings $[e(u_1), e(u_2), \dots e(u_n)]$, captures the semantic information of each utterance within the dialogue window. These embeddings serve as the foundational input for subsequent components in the MAD-detect architecture, enabling the model to effectively analyze and understand the dialogue's context and flow.

## 3.4 Speaker Embedding Concatenation

Speaker information plays a critical role in enhancing the understanding of the interactive context of dialogues. Given the set of speaker labels $S = [s_1, s_2, \dots, s_n]$, where $s_i$ represents the speaker label for the $i$-th utterance, we incorporate speaker embeddings to enrich the representation of each utterance.

To obtain speaker-aware utterance representations, we define the transformed embedding for each utterance $u_i$ as follows:

$$e'(u_i) = concatenate\left(e(u_i), f(s_i)\right) \tag{3.2}$$



doi:10.6342/NTU202403676

In this equation, $e(u_i)$ represents the utterance embedding obtained from the Utterance Encoder, while $f(s_i)$ denotes the embedding vector associated with the speaker label $s_i$. Each speaker label is assigned a learnable embedding layer, allowing the model to capture the unique characteristics or speaking styles of the speakers involved in the conversation. This embedding layer is updated during training, enabling the model to adaptively learn the characteristics associated with each speaker. By concatenating the speaker embedding with the corresponding utterance embedding, we generate speaker-aware utterance representations, denoted as $e'(u_i)$. This integration of speaker information is anticipated to assist the model in understanding that different speakers may exhibit distinct patterns in their expressions of engagement or disengagement throughout the dialogue. Consequently, the inclusion of speaker embeddings enhances the model's ability to analyze interactions more completely, facilitating improved detection of user disengagement within the dialogue system.

## 3.5 BiLSTM

The BiLSTM (Bidirectional Long Short-Term Memory) component (Figure 5) of the MAD-detect architecture demonstrates the effectiveness in capturing contextual information from both past and future utterances. This dual-context capability is important for understanding the dynamic flow of dialogue over the ongoing



doi:10.6342/NTU202403676

conversation, making BiLSTM particularly well-suited for tasks that require a comprehensive knowledge of the sequence of interactions.

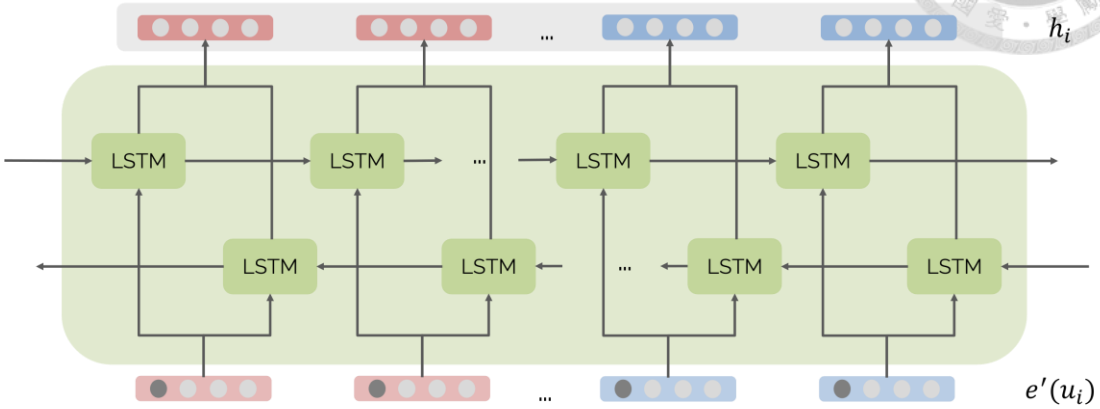



**Figure 5:** *Structure of BiLSTM*

Given the input embeddings $[e'(u_1), e'(u_2), \dots, e'(u_n)]$ corresponding to $n$ utterances, the BiLSTM processes these embeddings through two separate layers: a forward layer and a backward layer. The forward layer generates the hidden state for each utterance $u_i$ as follows:

$$\vec{h}_i = LSTM_{forward}\left(e'(u_i), \vec{h}_{i-1}\right) \tag{3.3}$$

Simultaneously, the backward layer computes the hidden state for each utterance $u_i$ using the subsequent utterances:

$$\overleftarrow{h}_i = LSTM_{backward}\left(e'(u_i), \overleftarrow{h}_{i+1}\right) \tag{3.4}$$

By combining the outputs of both the forward and backward layers, we obtain the final representation for each utterance:

$$h_i = \left[\vec{h}_i;\ \overleftarrow{h}_i\right] \tag{3.5}$$



doi:10.6342/NTU202403676

This concatenation results in a comprehensive representation $h_i$ for each utterance $u_i$ which is utilized for downstream utterance-level tasks. The incorporation of BiLSTM in the MAD-detect architecture thus enables a richer understanding of the dialogue context, ultimately enhancing the model's capability to detect user disengagement effectively.

## 3.6 Turn pooling layer

In natural, open-domain conversations, speakers often express multiple sentences in succession. To effectively capture the intent behind a speaker's contributions, consecutive utterances by the same speaker are aggregated into a unit known as a speaker turn. This aggregation process enables the model to construct a more consistent representation of the speaker's overall message.

To achieve this, the turn pooling layer employs mean pooling on the utterance-level representations. This method generates a turn-level representation that integrates the speaker's intended meaning while minimizing the influence of irrelevant utterances. By transitioning from utterance-level to speaker turn-level representation, the model effectively extracts the contextual information relevant to each speaker's turn Figure 6.

Given a window of dialogue representations at the utterance level, denoted as $H = [h_1, h_2, \dots, h_n]$, and the corresponding speakers $S = [s_1, s_2, \dots s_n]$, the first step involves identifying the turn boundaries, represented as $B = [b_1, b_2, \dots, b_k]$, where



doi:10.6342/NTU202403676

$s_x \neq s_{x+1}$. This identification allows the model to depict the segments of dialogue attributed to each speaker. For each identified turn $j$, the turn-level representation $t_j$ is calculated using the mean of the hidden states within the corresponding boundaries:

$$h'_j = mean\,(h_x | b_j \leq x < b_{j+1}) \tag{3.6}$$

The resulting turn-level representations are consolidated into $H' = [h'_1, h'_2, \ldots, h'_m]$, where $m$ stands for the total number of turns identified. This pooling mechanism enhances the model's ability to capture the collective context of a speaker's contributions, thereby facilitating more accurate detection of user disengagement in dialogue systems.

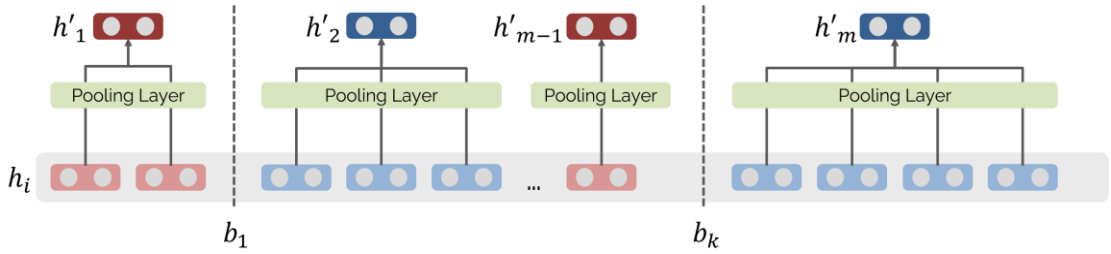


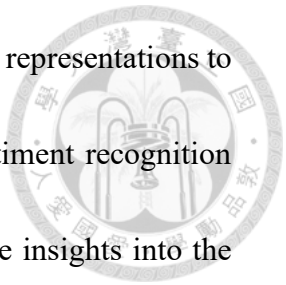

**Figure 6:** *Illustration of Turn Pooling Layer*

## 3.7 Multi-task Classification

Since the proposed MAD-detect architecture successfully generates both utterance-level and turn-level representations within the defined dialogue window through its preceding modules. These representations serve distinct purposes for the auxiliary and main tasks of the model.



doi:10.6342/NTU202403676



For the auxiliary tasks, the architecture employs utterance-level representations to analyze shifts in speaker expressions, specifically focusing on sentiment recognition and dialogue act classification. These tasks are vital as they provide insights into the details of communication that may indicate varying actions or information in one expression. In contrast, the main task classifier utilizes turn-level representations to detect speaker disengagement. This differentiation in representation levels is crucial, as disengagement may not be immediately apparent in individual utterances but instead requires a comprehensive assessment across multiple turns of dialogue. By examining the aggregated context of a speaker's contributions, the model can more effectively identify disengagement.

The architecture implements a hard parameter-sharing strategy, wherein a shared BiLSTM layer is trained simultaneously across all tasks. This shared layer forms the foundation for learning robust representations that are beneficial for each task. Subsequently, task-specific heads branch off from this shared layer, enabling the model to tailor its outputs to the unique requirements of each classification task. This design aims to enhance the performance of the main task of disengagement detection by leveraging the interconnectedness of the tasks and the shared knowledge derived from the training process.



doi:10.6342/NTU202403676

## 3.8 Loss Functions and Weighting Strategies

In the proposed MAD-detect model, distinct loss functions are utilized for the main task and auxiliary tasks to optimize overall performance effectively. For the auxiliary tasks, which include sentiment recognition and dialogue act classification, CrossEntropyLoss serves as the chosen loss function. This function is particularly well-suited for multi-class classification problems, as it provides a robust probability distribution over the various classes, enabling accurate differentiation among the outputs. For the main task of detecting speaker disengagement, we employ BCEWithLogitsLoss. This loss function is advantageous for binary classification tasks, as it combines a sigmoid layer with binary cross-entropy loss into a single, efficient computation. This integration not only enhances numerical stability but also streamlines the training process. To summarize, CrossEntropyLoss is utilized for auxiliary tasks due to its effectiveness in managing multi-class classification and its ability to yield a probability distribution over classes. Conversely, BCEWithLogitsLoss is employed for the main task, capitalizing on its suitability for binary classification by merging the sigmoid activation and binary cross-entropy into one step for efficiency and stability.

To address class imbalance within the main task, class weighting is implemented by assigning higher weights to less frequent classes. By doing so, the model is encouraged to pay more attention to the positive class, which is critical for accurately



doi:10.6342/NTU202403676

detecting disengagement, especially given its relative infrequency in typical datasets. The total loss for the main task can be represented as:

$$L_{main} = w_c \cdot l_1 + l_0 \tag{3.7}$$

Here, $L_{main}$ signifies the total loss for the main task, $w_c$ represents the class weight, $l_1$ denotes the loss for disengagement labels, and $l_0$ refers to the loss for non-disengagement labels.

Additionally, to balance the importance of various tasks within our multi-task learning framework, we implement task weighting. Each task is assigned a specific weight that reflects its significance, allowing us to prioritize the main task while benefiting from the auxiliary tasks. The task weighting approach is designed to ensure balanced contributions across tasks, with the overall loss calculated as:

$$L_{total} = \frac{(w_{aux1} \cdot L_{aux1} + w_{aux2} \cdot L_{aux2} + w_{main} \cdot L_{main})}{(w_{aux1} + w_{aux2} + w_{main})} \tag{3.8}$$

Static task weight methods calculate the loss by weighing each task's loss with predetermined static weights, ensuring a balanced contribution. However, to enhance the adaptability of our model, we also employ dynamic adjustment of task weights. As the losses evolve during training, the weights are dynamically updated to reflect the relative importance of each task. This method ensures that tasks with higher current losses contribute more significantly to the overall loss calculation, effectively balancing the training process in response to varying task difficulties.



doi:10.6342/NTU202403676

The dynamic task weight calculation is represented as:

$$L_{total} = w_t[0] \cdot L_{aux1} + w_t[1] \cdot L_{aux2} + w_t[2] \cdot L_{main} \tag{3.9}$$

where the weights are initialized and updated as:

$$initial\ w_t = [w_{aux1}, w_{aux2}, w_{main}] \tag{3.10}$$

$$updated\ w_t = \begin{bmatrix} \frac{L_{aux1}}{L_{aux1} + L_{aux2} + L_{main}}, \\ \frac{L_{aux2}}{L_{aux1} + L_{aux2} + L_{main}}, \\ \frac{L_{main}}{L_{aux1} + L_{aux2} + L_{main}} \end{bmatrix} \tag{3.11}$$

This dynamic adjustment mechanism provides several advantages: it enhances flexibility by adapting to changing task dynamics, improves training efficiency by focusing on tasks with higher losses, and optimizes the learning process by allocating resources more effectively to tasks that require greater attention. By integrating these loss functions and weighting strategies, the MAD-detect architecture aims to achieve a comprehensive and robust model for dialogue disengagement detection.



doi:10.6342/NTU202403676

# Chapter 4 Experiment

## 4.1 Data Collection

In our research, we undertook the task of collecting a Chinese dialogue dataset by simulating real-world conversations. This dataset extends and refines the NTUBI-Diag dataset, aiming to enhance the variety and depth of conversational scenarios, particularly those involving speaker disengagement.

To achieve this, we created dialogues using specific prompts that provided detailed settings and situations. Each dialogue was generated by two individuals guided by these prompts. The simulated dialogues were categorized into 46 scenarios, divided into the daily conversation section and the customer service section, each corresponding to distinct fields and situations. Importantly, for the daily conversation scenarios, we incorporated speaker settings in the prompts to ensure that certain dialogues had a clear intention of disengagement.

**Table 2:** *Comparison between two NTUBI-Diag collection*

| **NTUBI-Diag collection** | | |
|---|---|---|
| | First time | Second time |
| # of participants | 10 | 19 |
| # of dialogues | 600 | 950 |
| Sentiment annotation | ✓ | ✓ |
| Dialogue act annotation | ✗ | ✓ |
| Disengagement annotation | ✗ | ✓ |

The results of the two rounds of data collection for NTU-BI-Diag are summarized in Table 2. We engaged 19 participants in this data collection process, each responsible



doi:10.6342/NTU202403676

for generating 50 dialogue samples, resulting in a total of 950 new dialogues. These new samples were combined with the previous NTUBI-Diag dataset to create the updated NTUBI-Diag v2. This integration expanded and enriched the dataset with more varied conversational instances. The statistical details are presented in Table 3.

**Table 3:** *Statistics of the NTU-BI-Diag v2 Dataset*

| **NTUBI-Diag v2** | |
|---|---|
| # of dialogues | 1,170 |
| # of utterances | 29,313 |
| # of utterances per dialogue | 25.05 |
| # of speaker turns | 13,781 |
| # of utterances per turn | 2.13 |

For scenarios where the content to be expressed was extensive, we allowed the same speaker to continue speaking across multiple utterances. This approach provided a more realistic flow of conversation, capturing the natural ebb and flow of dialogue. Specifically, in the customer service scenario involving a service agent and a customer, the prompts provided fields and situations, resulting in 22 unique combinations. Participants had the flexibility to choose their character, with reference characters provided if they encountered difficulty in making a decision (Table 4). The daily conversation scenario was designed to include a topic starter and a responder. To ensure the involvement of disengaged speakers, we crafted prompts that divided the responder into those willing and unwilling to engage in the conversation. This scenario covered 24 combinations, each designed to reflect various degrees of engagement and disengagement (Table 5).



doi:10.6342/NTU202403676

**Table 4:** *Detail of Simulated Dialogue Scene: Customer Service Scenario Section*

| Field | Category | Situation |
|---|---|---|
| Online shopping | Product | Product Defects |
| | | Wrong Product |
| | | Return/Exchange/Compensation |
| | | Product Inquiry |
| | Price | Wrong Price |
| | | Questions about Coupons/Discounts |
| | | Bargaining |
| | Other | Special Situation |
| Restaurant | Product | Product Defects |
| | | Wrong Product |
| | Price | Wrong Price |
| | | Questions about Coupons/Discounts |
| | Other | Questions about Reservations |
| | | Helping Find Items |
| | | Special Situation |
| Hotel | Product | Product Defects |
| | | Wrong Product |
| | | Product Inquiry |
| | Other | Helping Find Items |
| | | Special Situation |
| Online platform | Other | Function Usage Problems |
| | | Special Situation |

By precisely integrating these comprehensive data collection methods and refining the existing dataset, NTUBI-Diag v2 stands as a robust foundation for analyzing dialogue engagement and disengagement in various conversational contexts. This



doi:10.6342/NTU202403676

enriched dataset is instrumental for training models to detect and respond to speaker disengagement effectively, thus advancing the field of dialogue system research.

**Table 5:** *Detail of Simulated Dialogue Scene: Daily Conversation Scenario Section*

| Field | Situation | Speaker |
|---|---|---|
| School/Office/Public | Discussion | Sharer |
| | | Responder with interest |
| | | Sharer |
| | | **Responder without interest** |
| | Promotion/ Recommendation | Recommender |
| | | Responder with interest |
| | | Recommender |
| | | **Responder without interest** |
| | Invitation | Inviter |
| | | Invitee with will |
| | | Inviter |
| | | **Invitee without will** |
| | Inquiry/Request | Questioner |
| | | Aggressively responding responder |
| | | Questioner |
| | | **Unwillingly helping responder** |

## 4.2 Data Annotation

In our research, we aimed to ensure precise and meaningful annotations for disengagement intention, sentiment, and dialogue acts. To achieve this, we employed a methodical approach that maintained high consistency and reliability throughout the process.



doi:10.6342/NTU202403676

Disengagement intention was annotated directly by the speaker who generated the utterances during the creation of the simulated dialogue. This method was chosen to accurately capture the speaker's intention behind their words, as only the speaker can genuinely mark their own disengagement intention, providing a closer approximation to the real situation. We defined disengagement as a state in which the speaker has lost focus, attempted to quit the conversation, or shifted focus to disrupt the conversation. Utterances with disengagement intention were labeled as 1, providing a clear binary indicator of disengagement within the dataset. Table 6 shows the distribution of the annotated turns, revealing a significant imbalance between turns with disengagement intention and those without. This imbalance presents challenges for training models but also reflects real-world scenarios where disengagement is less frequent than engagement.

**Table 6:** *Statistics of Disengagement Intention Labels*

| Disengagement | Turns | Percentage (%) |
|---|---|---|
| **0** | 13,781 | 88% |
| **1** | 1,654 | 12% |

For sentiment annotation, we utilized the existing sentiment labels from the NTUBI-Diag dataset, which include the classic sentiment polarity categories of negative, positive, and neutral. Sentiment recognition is a well-established task in natural language processing that involves determining the polarity of a given text. By leveraging these existing labels, we ensured that our sentiment annotations were



doi:10.6342/NTU202403676

grounded in well-recognized standards and methodologies. Table 7 illustrates the distribution of sentiment polarity labels within our dataset. The majority of the labels are neutral, followed by negative, and then positive. This distribution highlights the predominance of neutral sentiment in our data, which aligns with typical conversational dynamics where neutral statements are more common than overtly positive or negative ones.

**Table 7:** *Statistics of Sentiment Polarity Labels*

| Sentiment | Utterances | Percentage (%) |
|---|---|---|
| **Positive** | 4,091 | 13.96% |
| **Neutral** | 18,884 | 64.42% |
| **Negative** | 6,338 | 21.62% |

Our dialogue act annotations were based on the schema used in the Chinese dataset CPED (Chen et al., 2022). Given the complexity of this annotation task, which involves nineteen different labels, and the subjective nature of interpreting dialogue behaviors, we designed comprehensive guidelines for each label to ensure consistent and high-quality annotations. These guidelines were created to maintain consistent annotation standards for future dataset expansions, helping annotators understand and apply the labels uniformly. The nineteen dialogue acts and their corresponding definitions are shown in Table 8. To ensure a general and clear distinction among these dialogue acts, we adopted a two-phase annotation process. This process helped maintain consistent



doi:10.6342/NTU202403676

interpretation of each label among different annotators, ensuring the reliability and accuracy of the annotations.

**Table 8:** *Definition of Dialogue Acts*

| Dialogue Act | Definition |
|---|---|
| **Statement-opinion (sv)** | （帶有意見想法/價值判斷的）陳述；聲明；表態 |
| **Question (q)** | 提出仍未有答案/需要被回應的疑問 |
| **Answer (ans)** | 針對某個疑問的回應 |
| **Statement-non-opinion (sd)** | （不帶有意見想法的）陳述；聲明；表態 |
| **Command/request (c)** | 要求；指使他人（以自身期待的方式）行動 |
| **Acknowledge (a)** | 承認；認可某些事物確實存在/屬實為真；<br>自然的回應、附和 |
| **Reject (rj)** | 拒絕；駁回；排斥去接受/相信/使用其他人事物 |
| **Agreement/acceptance (aa)** | 表明意見一致，認同；相信；<br>接受對方的想法/提議 |
| **Conventional-closing (fc)** | 常用於結束；收尾；作結的語句 |
| **Apology (fa)** | 承認錯誤並傳達表明後悔/歉意 |
| **Greeting (g)** | 問候；招呼；迎接；表示歡迎 |
| **Interjection (ij)** | （表達突然、短暫情緒的）感歎詞 |
| **Disagreement (dag)** | 表明意見分歧，否定；爭論；<br>批判對方的想法/提議 |
| **Thanking (ft)** | 對他人的行動表達感激；感恩；感謝 |
| **Appreciation (ba)** | 理解；欣賞；賞識某些事物的價值與重要性 |
| **Comfort (cf)** | 安慰；撫慰，使人感覺到慰藉；鼓勵 |
| **Irony (ir)** | 嘲諷、諷刺；反語，<br>表達與字面上相反意義的語境 |
| **Quotation (^q)** | （從其他著作來源）引用；<br>引述的短句；特別表達是別人的意見 |
| **Other (oth)** | 其他 |

The annotation process for dialogue acts was conducted in two phases to ensure accuracy and consistency. During the initial phase, a team member and I annotated the dialogues according to the provided guidelines. To measure inter-annotator agreement, we calculated Cohen's kappa, which resulted in a value of 0.624. This moderate level



doi:10.6342/NTU202403676

of agreement highlighted areas where the guidelines needed clarification. Consequently, we updated the guidelines to address ambiguities and improve consistency.

In the second phase, a new annotator was introduced to the updated guidelines. This annotator, along with me, re-annotated the dialogues. The agreement level in this phase significantly improved, with Cohen's kappa reaching 0.924. This high consistency demonstrated the effectiveness of the updated guidelines. Following this, the new annotator took over the remaining annotations for the dialogue act labels, ensuring the uniform application of the guidelines. The details of the two-phase annotation process are presented in Table 9. The statistics of the dialogue act labels are shown in Table 10, which includes the distribution and percentage of each label

**Table 9:** *Details of Two-phase Annotation Process*

| Inter-annotator agreement | | |
|---|---|---|
| | Initial Phase | Subsequent Phase |
| Annotators | 2 team members | 1 new annotator<br>1 team member |
| # of dialogues | 30 | 100 |
| # of utterances | 647 | 2,191 |
| Cohen's kappa | 0.624 | 0.924 |

By implementing this systematic and iterative approach to data annotation, we were able to achieve a high level of consistency and reliability in our dataset. This meticulous process not only enhanced the quality of our annotations but also provided a robust foundation for future research and applications on this dataset.



doi:10.6342/NTU202403676

Table 10: *Statistics of Dialogue Act Labels*

| Dialogue Act | Utterances | Percentage (%) |
|---|---|---|
| **Statement-opinion (sv)** | 7,629 | 26.03% |
| **Question (q)** | 4,659 | 15.89% |
| **Answer (ans)** | 3,956 | 13.50% |
| **Statement-non-opinion (sd)** | 2,646 | 9.03% |
| **Command/request (c)** | 2,205 | 7.52% |
| **Acknowledge (a)** | 1,306 | 4.45% |
| **Reject (rj)** | 1,198 | 4.09% |
| **Agreement/acceptance (aa)** | 1,151 | 3.93% |
| **Conventional-closing (fc)** | 721 | 2.46% |
| **Apology (fa)** | 595 | 2.03% |
| **Greeting (g)** | 576 | 1.96% |
| **Interjection (ij)** | 572 | 1.95% |
| **Disagreement (dag)** | 550 | 1.88% |
| **Thanking (ft)** | 551 | 1.88% |
| **Appreciation (ba)** | 385 | 1.31% |
| **Comfort (cf)** | 304 | 1.04% |
| **Irony (ir)** | 180 | 0.61% |
| **Quotation (^q)** | 90 | 0.31% |
| **Other (oth)** | 39 | 0.13% |

## 4.3 Evaluation and Metrics

To ensure a robust and reliable assessment of our model's performance, we employ a 10-fold cross-validation method. This approach involves partitioning our dataset into 10 equal sets, where in each iteration, 9 sets are used for training, and 1 set is reserved for testing. This process is repeated 10 times, with each set serving as the test set once. By averaging the results across these 10 iterations, we obtain a reliable estimate of our model's overall performance. To prevent data leakage during training, we generate dialogue history windows within each partitioned set. This step is crucial to ensure that



doi:10.6342/NTU202403676

the model does not inadvertently learn from information it would not have access to during actual deployment, thereby maintaining the integrity of our evaluation.

Given the class imbalance inherent in our auxiliary tasks, we use the weighted-F1 score to evaluate their performance. The weighted-F1 score aggregates contributions from all classes based on their frequencies, providing a balanced evaluation that accounts for the uneven distribution of classes. This metric ensures that the performance is not overly influenced by the majority class, offering a better understanding of the model's effectiveness across all classes. For the main task, which involves detecting disengagement, we calculate precision, recall, and F1 score for both the disengagement (label 1) and non-disengagement (label 0) classes. This comprehensive analysis allows us to thoroughly assess the model's effectiveness in identifying both engaged and disengaged states. By evaluating these metrics for both classes, we can better understand the model's strengths and weaknesses, ensuring a balanced and thorough assessment.

The combined use of these metrics provides a detailed and comprehensive evaluation of our model. The 10-fold cross-validation method offers a robust estimate of performance, while the weighted-F1 score addresses class imbalance in the auxiliary tasks, and precision, recall, and F1 score for both labels ensure a thorough assessment of the main task. This systematic approach helps us to accurately evaluate the



doi:10.6342/NTU202403676

effectiveness and reliability of our model, paving the way for further refinements and improvements.

## 4.4 Experimental Settings

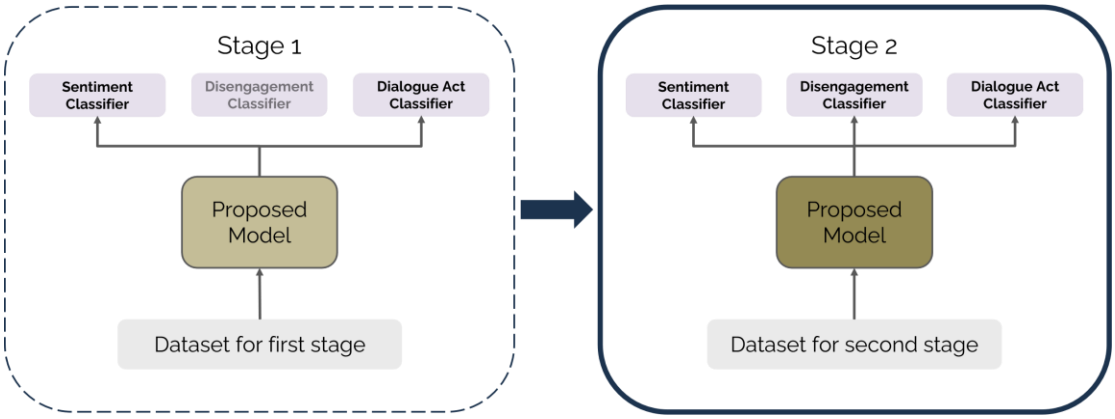



**Figure 7:** *Demonstration of the Proposed Two-stage Training Strategy*

### 4.4.1 Two-stage Training Strategy

In our experiments, we employed a two-stage training strategy to optimize our multi-task learning approach, as demonstrated in Figure 7. This method aims to enhance the performance of the main task by leveraging the learning from auxiliary tasks in a structured manner.

In the first stage, we trained the model exclusively on the auxiliary tasks. This phase allowed the model to learn useful representations from the auxiliary tasks without the added complexity of the main task. By focusing solely on the auxiliary tasks, the model effectively captured the underlying patterns and features relevant to these tasks. Following the auxiliary tasks training, we proceeded to the second stage, where we used the parameters learned from the first stage to initialize the shared base of the multi-task



doi:10.6342/NTU202403676

model. We then continued training the model on both the auxiliary tasks and the main task simultaneously. This approach ensured that the shared base started with a solid foundation of useful representations, significantly enhancing the performance on the main task. The primary advantage of this two-stage approach is that it allows the shared base to develop robust representations from the auxiliary tasks before optimizing for the main task. This often leads to better performance on the main task compared to training the full multi-task model from scratch.

To implement our two-stage training strategy, we first split the dialogues based on the presence of disengagement intention. For dialogues without disengagement intention, totaling 531, we applied an 80-20 train-test split. The performance of the first stage was evaluated based on this validation split, which helped us determine the optimal model parameters to carry forward to the next stage. Following this, we performed a 10-fold cross-validation on the remaining 639 dialogues that included disengagement intention. This precise evaluation method ensured that our model's performance was robust and generalizable. The evaluation of the two-stage approach was based on the performance observed during this 10-fold cross-validation phase.

The two-stage training strategy in our multi-task learning approach provided a structured method to enhance the main task's performance. By initially training on the auxiliary tasks and then on both auxiliary and main tasks, the model developed a strong



doi:10.6342/NTU202403676

foundation of useful representations. Although this method involves additional computational resources, the improved performance on the main task often justifies the extra effort.

**Table 11:** *Statistics of Dataset Used in Two-stage Training Strategy*

| | # of dialogues | % of dialogues |
|---|---|---|
| Total data | 1,170 | 100% |
| Data for first stage | 531 | 45.38% |
| Data for second stage | 639 | 54.62% |

### 4.4.2 Experimental Setup

In our experimental setup, we consistently utilized the Adam optimizer across all our models. To further enhance our training process in first stage, we implemented an early stopping mechanism. This mechanism monitors the loss on the validation data during training and stops training if the loss does not improve for a predefined number of epochs. This approach prevents overfitting and ensures that the model generalizes well to unseen data.

In our multi-task learning framework, task weights were carefully assigned to balance the importance and difficulty of the main task and auxiliary tasks. The task weights were set initially as follows: $w_t = [w_{aux1}, w_{aux2}, w_{main}]$, where $w_{aux1} = 0.2$, $w_{aux2} = 0.3$, $w_{main} = 0.5$ This weighting scheme prioritizes the main task while still allowing significant contributions from the auxiliary tasks.



doi:10.6342/NTU202403676

For comparative purposes, we benchmarked our proposed method against several established models. Prior research indicates that most models employ BERT followed by pooling for predictions. Additionally, some studies use sentiment or dialogue acts as features for the task. We adopted these approaches as our baselines due to the differences in the dataset utilized in our study compared to the original research on these benchmark methods. The hyperparameter settings of the experiments were aligned with those listed in Table 12, ensuring a fair and consistent comparison across different models.

In summary, our experimental settings were meticulously designed to optimize the performance of our multi-task learning model. By utilizing a two-stage training strategy, carefully assigned task weights, and benchmarking against established methods, we ensured a robust and comprehensive evaluation of our proposed approach.

**Table 12:** *Hyperparameter Settings*

| Hyperparameter settings | | | |
|---|---|---|---|
| Dialogue history window length | Epochs | Early stopping epochs | Learning rate |
| $n = 10$ | $e_1 = 300$<br>$e_2 = 5$ | $w_{es1} = 50$ | $lr_1 = 5e-5$<br>$lr_2 = 1e-3$ |
| Batch size | Utterance embedding dim. | Speaker embedding dim. | LSTM hidden dim. |
| $b = 32$ | $d_u = 768$ | $d_s = 2$ | $d_{lstm} = 128$ |



doi:10.6342/NTU202403676

## 4.5 Experimental Results

The experimental results presented in this research underscore the effectiveness of our multi-task learning approach, MAD-detect, for dialogue disengagement detection. We evaluated performance across three tasks: sentiment recognition, dialogue act classification, and disengagement detection. Our focus was on comparing the performance of our multi-task model against various baselines in different experiments. The results consistently show that MAD-detect outperforms the baselines, highlighting the benefits of incorporating auxiliary tasks to enhance the main task of disengagement detection.

### 4.5.1 MAD-detect

**Table 13:** *Comparison of Baselines and Our Proposed MAD-detect Method*

| Task | | | Main Task | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | Disengagement | | | | | | |
| Method | | | Non-disengagement | | | Disengagement | | | Accuracy |
| First stage | Second stage | Weighting strategy | Precision | Recall | F1 | Precision | Recall | F1 | |
| | Baseline (BERT only) | class weight=1.5 | 94.53% | **97.87%** | 96.17% | 46.28% | 24.49% | 32.03% | 92.75% |
| | Baseline (feature) | class weight=1.5 | 95.12% | 97.58% | 96.33% | 50.69% | 33.16% | 40.09% | 93.09% |
| | MAD-detect w/o auxiliary tasks | class weight=1.5 | 95.75% | 96.53% | 96.14% | 48.04% | 42.82% | 45.28% | 92.78% |
| | MAD-detect | class weight=1.5 | **95.96%** | 96.86% | **96.41%** | 52.06% | 45.51% | 48.56% | **93.28%** |
| **Sentiment DA** | **MAD-detect** | **dynamic task weight class weight=1.5** | 93.58% | 92.83% | 93.20% | **55.63%** | **58.52%** | **57.04%** | 88.26% |

Table 13 presents the performance comparison between our proposed MAD-detect model and several baselines. The evaluated methods include a simple BERT-only baseline, a baseline that concatenate the sentiment and dialogue act as additional features, and our proposed multi-task model with and without auxiliary tasks. The



doi:10.6342/NTU202403676

objective was to assess how well our multi-task model leverages the additional information from auxiliary tasks to improve disengagement detection.

For non-disengagement detection, the proposed MAD-detect model without two-stage training strategy achieved the highest precision (95.96%) and F1 score (96.41%), while the highest recall (97.87%) was achieved by the baseline BERT-only model. For disengagement detection, the highest precision (55.63%), the highest recall (58.51%), and the best F1 score for disengagement (57.04%) was obtained with the complete MAD-detect model.

The MAD-detect model's strength lies in its best performance across all metrics, particularly excelling in recall for disengagement detection, which is the most critical part in a dialogue system. This suggests that our multi-task approach effectively captures the detailed dependencies between tasks, thereby enhancing the model's ability to detect disengagement. The comprehensive evaluation underscores the effectiveness of our multi-task learning strategy, demonstrating its capability to leverage auxiliary tasks to improve the primary task's performance.

### 4.5.2 Weighting strategy

The experimental results presented in Table 14 compare the performance of different weighting strategies in our MAD-detect model. By examining various



doi:10.6342/NTU202403676

weighting strategies, we aim to determine how the assignment of weights affects the overall performance of the model across both the primary and auxiliary tasks.

**Table 14:** *Comparison of Performance Across Different Weighting Strategy*

| Task | | | Auxiliary Tasks | | Main Task | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Sentiment | Dialogue Act | Disengagement | | | | | | |
| Method | | | Weighted-F1 | Weighted -F1 | Non-disengagement | | | Disengagement | | | Accuracy |
| First stage | Second stage | Weighting strategy | | | Precision | Recall | F1 | Precision | Recall | F1 | |
| | MAD-detect | equal weight | 77.01% | 62.78% | 95.53% | 97.49% | 96.50% | 53.81% | 39.10% | 45.29% | 93.42% |
| | MAD-detect | class weight=6 | 76.55% | 62.84% | **96.32%** | 94.97% | 95.64% | 43.43% | 51.55% | 47.14% | 91.94% |
| | MAD-detect | task weight | 77.06% | 62.23% | 95.47% | 97.31% | 96.38% | 51.64% | 38.37% | 44.03% | 93.20% |
| | MAD-detect | dynamic task weight | 76.65% | **62.95%** | 94.63% | **98.58%** | **96.57%** | **57.31%** | 25.37% | 35.17% | **93.48%** |
| | MAD-detect | dynamic task weight class weight=1.5 | 76.63% | 61.26% | 93.18% | 92.83% | 93.00% | 54.41% | 55.74% | 55.07% | 87.89% |
| **Sentiment DA** | **MAD-detect** | **dynamic task weight class weight=1.5** | **77.25%** | 62.06% | 93.58% | 92.83% | 93.20% | 55.63% | **58.52%** | **57.04%** | 88.26% |

In the auxiliary tasks, the weighted-F1 scores for sentiment recognition varied across the weighting strategies. The highest score, 77.25%, was achieved with the dynamic task weight and class weight with two-stage training strategy, indicating the robustness of this combined approach. For dialogue act classification, the highest weighted-F1 score was 62.95%, obtained with the dynamic task weight strategy, closely followed by the class weight strategy at 62.84%.

In the primary task of disengagement detection, results varied based on the weighting strategies. For non-disengagement detection, the equal weight strategy achieved a high precision of 95.53% and recall of 97.49%, resulting in an F1 score of 96.50%. However, the highest F1 score for non-disengagement, 96.57%, was obtained using the dynamic task weight strategy, indicating that dynamically adjusting weights based on task performance can enhance the model's ability to correctly identify non-disengagement cases. In contrast, for disengagement detection, the dynamic task weight



doi:10.6342/NTU202403676

combined with class weight strategy achieved the second highest F1 score of 55.07%, with a precision of 54.41% and recall of 55.74%, only lower than the dynamic task weight and class weight with two-stage training strategy. These results suggest that dynamically adjusts based on task performance while also considering class imbalances, can significantly improve the detection of disengagement.

Incorporating class weights into our model significantly improved recall, although it caused a slight reduction in precision. This trade-off is advantageous in our context as it increases the model's sensitivity to disengagement instances, which are often underrepresented. The enhanced recall ensures that more instances of disengagement are correctly identified, contributing to a more comprehensive detection system. The dynamic weighting mechanism prioritizes harder tasks during training, such as dialogue act classification and disengagement detection. This approach ensures that the model allocates more resources and attention to challenging elements, fostering a more balanced and effective learning process. As a result, the model shows marked improvement in tasks that are inherently more complex, without neglecting the easier tasks.

The dynamic task weight approach, particularly when combined with class weight adjustments, proves to be the most effective in enhancing the performance of our MAD-detect model. This approach allows the model to adaptively balance the contributions



doi:10.6342/NTU202403676

of each task, leveraging the contextual information provided by auxiliary tasks to improve the detection of disengagement. The comparative analysis of different weighting strategies demonstrates the critical role of weight assignment in multi-task learning. The dynamic task weight strategy, especially when combined with class weight adjustments, offers a balanced and robust solution, making our MAD-detect model the best choice for dialogue disengagement detection.

### 4.5.3 Ablation study

In this section, we explore the impact of different first-stage training and overall configurations on the performance of our multi-task learning model, MAD-detect. By conducting an ablation study, we aim to understand the contributions of each auxiliary task, sentiment recognition and dialogue act classification, to the primary task of disengagement detection. We compare the performance of our model when trained with only one auxiliary task at a time versus training with both auxiliary tasks.

**Table 15:** *Comparison of Performance Across Different First Stage Tasks*

| Task | | Auxiliary Tasks | | Main Task | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Sentiment | Dialogue Act | Disengagement | | | | | | |
| Method | | Weighted-F1 | Weighted-F1 | Non-disengagement | | | Disengagement | | | Accuracy |
| First stage | Second stage | | | Precision | Recall | F1 | Precision | Recall | F1 | |
| | MAD-detect | 76.63% | 61.26% | 93.18% | 92.83% | 93.00% | 54.41% | 55.74% | 55.07% | 87.89% |
| Sentiment | MAD-detect | 77.32% | 60.54% | 93.35% | 91.67% | 92.50% | 51.46% | 57.46% | 54.29% | 87.12% |
| Sentiment | MAD-detect w/o DA | **77.38%** | | **93.63%** | 92.20% | 92.61% | 53.83% | **59.18%** | 56.38% | 87.81% |
| DA | MAD-detect | 77.00% | 61.83% | 92.98% | 92.34% | 92.66% | 52.28% | 54.61% | 53.42% | 87.32% |
| DA | MAD-detect w/o Sentiment | | **62.42%** | 93.42% | **93.06%** | **93.24%** | **55.94%** | 57.34% | 56.63% | **88.31%** |
| **Sentiment + DA** | **MAD-detect** | 77.25% | 62.06% | 93.58% | 92.83% | 93.20% | 55.63% | 58.52% | **57.04%** | 88.26% |

For sentiment recognition, the highest weighted-F1 score (77.38%) was achieved when the MAD-detect model was trained with only the sentiment task in the first stage



doi:10.6342/NTU202403676

and remove dialogue act classification in the second stage. The inclusion of both auxiliary tasks yielded a slightly lower weighted-F1 score of 77.25%. Similarly, for dialogue act classification, the best performance (weighted-F1 score of 62.42%) was observed when only dialogue act classification task was included. When the model was trained on both tasks, the weighted-F1 score was slightly lower at 62.06%.

When evaluating the main task of disengagement detection, the benefits of the full training strategy become even more apparent. For non-disengagement detection, the model that undergoes pre-training on auxiliary tasks demonstrates higher precision, recall, and F1 scores compared to the model trained exclusively on the main task. This suggests that pre-training on auxiliary tasks helps the model to accurately identify non-disengagement instances by providing richer contextual cues.

Disengagement detection, which is inherently more challenging, also shows significant improvements with the full training strategy. The model trained on all stages exhibits superior performance in terms of precision, recall, and F1 scores, clearly indicating that auxiliary task learning aids the model in better understanding and detecting disengagement. The substantial increase in these metrics highlights the added value of incorporating auxiliary task pre-training. Overall accuracy further proves these findings, with the model trained on all stages outperforming the model trained only on one stage. This demonstrates that the inclusion of auxiliary task pre-training not only



doi:10.6342/NTU202403676

enhances specific metrics but also contributes to a more robust and accurate model overall. The comprehensive improvement across different evaluation metrics underscores the efficacy of our multi-task learning approach.

In conclusion, the ablation study clearly illustrates the advantages of including the first stage of pre-training on auxiliary tasks. Models that undergo full training on both auxiliary and main tasks consistently outperform those trained only on the main task. This highlights the effectiveness of our multi-task learning strategy, where pre-training on auxiliary tasks such as sentiment recognition and dialogue act classification enriches the model's contextual understanding, leading to improved performance in the primary task of disengagement detection. These findings establish our approach as a robust and superior solution for dialogue disengagement detection, showcasing the significant benefits of leveraging multi-task learning with well-structured training stages.



doi:10.6342/NTU202403676

# 5 Conclusion

## 5.1 Conclusion

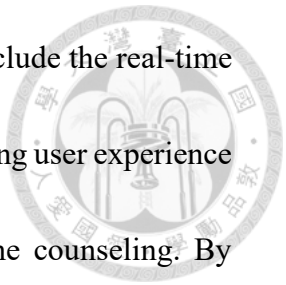
In this paper, we introduce a comprehensive approach to dialogue disengagement detection, addressing a critical aspect of dialogue systems by employing multi-task learning to incorporate auxiliary tasks such as sentiment recognition and dialogue act classification. Our methodology is based in a two-stage training strategy: the initial phase focuses on auxiliary tasks, while the subsequent phase integrates the primary task. This two-staged approach refines the model for robust and accurate detection of disengagement.

We have created a diverse and extensive dataset that includes a wide range of real-world scenarios. This dataset is instrumental in training and evaluating the proposed models and serves as a valuable resource for future research in dialogue systems. By leveraging real-world simulated data, we ensure that our models are not only theoretically sound but also practically applicable.

Our experiments demonstrate that the proposed model outperforms baseline models, particularly in detecting disengagement, achieving higher performance on recall, precision, and F1-score. The two-stage training strategy proves to be effective in enhancing the model's ability to identify disengagement, confirming the advantage of focusing on auxiliary tasks before integrating the main task.



doi:10.6342/NTU202403676



Practical applications of this research are far-reaching. They include the real-time adjustment of dialogue strategies to re-engage users, thereby improving user experience in various interactive systems such as customer service and online counseling. By accurately detecting disengagement, systems can prompt appropriate interventions, enhancing user satisfaction and overall interaction quality.

In summary, our work not only advances the field of dialogue disengagement detection but also provides a solid foundation for future developments in this area. The integration of multitask learning, the creation of a robust dataset, and the successful application of a two-stage training strategy collectively contribute to the effectiveness and practical relevance of our proposed approach.

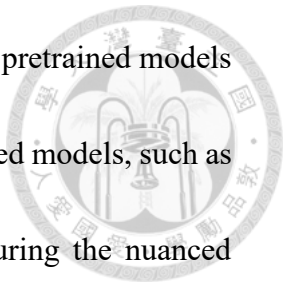
## 5.2 Future Works

There are several avenues for future research that can further enhance the robustness and applicability of our models. In expanding the dataset, annotating the remaining data will provide better coverage and a more comprehensive training set, which will improve the model's generalizability across different scenarios. This step is crucial for capturing a wider range of disengagement signals, leading to more accurate detection.

Future research should explore the use of existing models trained for auxiliary tasks to generate weak labels for unannotated data. This approach can significantly



doi:10.6342/NTU202403676



reduce the manual labeling burden while leveraging the strengths of pretrained models to enhance data quality. Additionally, incorporating more sophisticated models, such as attention-based models, may improve performance by better capturing the nuanced relationships between dialogue turns and speaker intentions. Simplifying the complexity of dialogue act annotations by merging some labels can streamline the annotation process and reduce ambiguity. This refinement may lead to more consistent and reliable annotations, contributing to the overall quality of the dataset.

Enhancing dialogue representation learning is another critical area for future work. One promising approach is to incorporate unsupervised tasks, such as next sentence prediction, in the first stage of the two-stage training strategy. This can help the model learn richer contextual representations and improve its understanding of dialogue flow and speaker intentions.

By addressing these areas, future research can build on the foundation laid by this paper, advancing the field of dialogue disengagement detection and contributing to the development of more sophisticated and effective dialogue systems. These improvements will not only enhance the robustness of the models but also broaden their practical applications, making them more adaptable and efficient in real-world scenarios.



doi:10.6342/NTU202403676

# References


Adam, M., Wessel, M., & Benlian, A. (2021). AI-based chatbots in customer service and their effects on user compliance. *Electronic Markets*, *31*(2), 427-445.

Antonio, R., Tyandra, N., Nusantara, L. T., & Gunawan, A. A. S. (2022). Study literature review: Discovering the effect of chatbot implementation in e-commerce customer service system towards customer satisfaction. In *Proceedings of 2022 International Seminar on Application for Technology of Information and Communication (iSemantic)*, pp. 296-301.

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

Deng, Y., Lei, W., Lam, W., & Chua, T.-S. (2023). A survey on proactive dialogue systems: Problems, methods, and prospects. *arXiv preprint arXiv:2305.02750*.

Deng, Y., Zhang, W., Yuan, Y., & Lam, W. (2023). Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. *arXiv preprint arXiv:2305.10172*.

Fedotov, D., Perepelkina, O., Kazimirova, E., Konstantinova, M., & Minker, W. (2018). Multimodal approach to engagement and disengagement detection with highly imbalanced in-the-wild data. In *Proceedings of the Workshop on Modeling Cognitive Processes from Multimodal Data*, pp 1-9.

Forbes-Riley, K., Litman, D., Friedberg, H., & Drummond, J. (2012). Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 91-102.

Ghazarian, S., Weischedel, R., Galstyan, A., & Peng, N. (2020). Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. In *Proceedings of the AAAI Conference on Artificial Intelligence,* vol. 34, pp. 7789-7796.

Glas, N., & Pelachaud, C. (2015). Definitions of engagement in human-agent interaction. In *Proceedings of 2015 International Conference on Affective Computing and Intelligent Interaction (ACII),* pp. 944-949. IEEE.

Higashinaka, R., Funakoshi, K., Kobayashi, Y., & Inaba, M. (2016). The dialogue breakdown detection challenge: Task description, datasets, and evaluation metrics. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16),* pp. 3146-3150.




doi:10.6342/NTU202403676

Hsiao, J. C.-y., Jih, W.-r., & Hsu, J. Y.-j. (2012). Recognizing continuous social engagement level in dyadic conversation by using turn-taking and speech emotion patterns. Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence.

Inaba, M., & Takahashi, K. (2015). Long Short-Term Memory Recurrent Neural Network を用いた対話破綻検出. 人工知能学会研究会資料 言語・音声理解と対話処理研究会 75 回 (2015/10) (p. 13). 一般社団法人 人工知能学会.

Jiang, H., Cheng, Y., Yang, J., & Gao, S. (2022). AI-powered chatbot communication with customers: Dialogic interactions, satisfaction, engagement, and customer behavior. *Computers in Human Behavior*, *134*, 107329.

Jiang, S., Vakulenko, S., & de Rijke, M. (2023). Weakly supervised turn-level engagingness evaluator for dialogues. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, Austin, Texas, USA,* pp. 258-268.

Liang, W., Liang, K.-H., & Yu, Z. (2021). HERALD: An annotation efficient method to detect user disengagement in social conversations. *arXiv preprint arXiv:2106.00162*.

Ma, X. (2018). Towards human-engaged AI. In *Proceedings of International Joint Conferences on Artificial Intelligence*, pp. 5682-5686.

Mastroianni, A. M., Gilbert, D. T., Cooney, G., & Wilson, T. D. (2021). Do conversations end when people want them to? In *Proceedings of the National Academy of Sciences*, *118*(10).

Matsumoto, K., Sasayama, M., Yoshida, M., Kita, K., & Ren, F. (2022). Emotion analysis and dialogue breakdown detection in dialogue of chat systems based on deep neural networks. *Electronics*, *11*(5), 695.

Nordheim, C. B., Følstad, A., & Bjørkli, C. A. (2019). An initial model of trust in chatbots for customer service—findings from a questionnaire study. *Interacting with Computers*, *31*(3), 317-335.

O'Brien, H. L., Cairns, P., & Hall, M. (2018). A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form. *International Journal of Human-Computer Studies*, *112*, 28-39.

Sugiyama, H. (2015). Chat-oriented dialogue breakdown detection based on combination of various data. In *Proceedings of Sixth Dialogue System symposium (SIG-SLUD), The Japanese Society for Artificial Intelligence, 6th Dialogue System Symposium*, pp. 51-56.

Yi, S., Goel, R., Khatri, C., Cervone, A., Chung, T., Hedayatnia, B., Venkatesh, A., Gabriel, R., & Hakkani-Tur, D. (2019). Towards coherent and engaging spoken



doi:10.6342/NTU202403676

dialog response generation using automatic conversation evaluators. *arXiv preprint arXiv:1904.13015*.

Yu, C., Aoki, P. M., & Woodruff, A. (2004). Detecting user engagement in everyday conversations. *arXiv preprint cs/0410027*.

Zhang, J., Chen, Q., Lu, J., Wang, X., Liu, L., & Feng, Y. (2024). Emotional expression by artificial intelligence chatbots to improve customer satisfaction: Underlying mechanism and boundary conditions. *Tourism Management*, *100*, 104835.



doi:10.6342/NTU202403676