

基於情緒支持的對話：融合情緒感知和關鍵字識別的研究
**Emotionally Supportive Dialogue: A Study of Integrating Emotion
Perception and Keyword Recognition**

巫宇哲

國立雲林科技大學

gehnew0916418068@gmail.com

陳重臣

國立雲林科技大學

jcchen@yuntech.edu.tw

摘要

隨著科技進步，生活變得更便利，但同時社會和個人健康壓力也不斷增加。面對醫療資源短缺的挑戰，本研究開發了一個創新的對話系統。該系統改良了傳統語言模型，能識別使用者的情緒並融合術語識別技術，從而提升資料整合效果。在生成模型的損失函數設計中，結合了預測下一詞彙的原始訓練目標與新增的關鍵字辨識目標。此外，系統整合了 RoBERTa 模型進行情緒分類和 BART 模型生成對話，構建了一個端到端的系統。實驗顯示，模型自動評估提升 0.3% 至 2%，BLEU 指標甚至提升 7% 至 10%，人工評估亦取得優異成績，證實兩種方法融合有效增強模型理解能力及情緒支持。

關鍵詞：對話系統、RoBERTa、BART、術語識別、情緒標籤

基於情緒支持的對話：融合情緒感知和關鍵字識別的研究

Emotionally Supportive Dialogue: A Study of Integrating Emotion Perception and Keyword Recognition

Yu-Zhe Wu

National Yunlin University of Science and Technology
gehnew0916418068@gmail.com

Jong-Chen Chen

National Yunlin University of Science and Technology
jcchen@yuntech.edu.tw

Abstract

With technological advancements, life has become more convenient, yet the pressures on societal and personal health continue to escalate. Facing the challenge of scarce medical resources, this study has developed an innovative dialogue system. This system has enhanced traditional language models by recognizing users' emotions and incorporating Terminology-Aware technology, thereby improving data integration efficiency. In designing the loss function for the generative model, we merged the original training goal of predicting the next word with the added objective of keyword recognition. Additionally, the system integrates the RoBERTa model for emotion classification and the BART model for dialogue generation, creating an end-to-end system. Experiments show that automatic evaluations of the model have improved by 0.3% to 2%, and the BLEU score has even increased by 7% to 10%. Manual evaluations have also achieved excellent results, confirming that the integration of these two methods effectively enhances the model's comprehension abilities and emotional support.

Keywords: Dialogue System, RoBERTa, BART, Terminology-Aware, Emotion Label

壹、導論

一、研究背景與動機

隨著時代快速的變遷，科技日新月異，不僅僅讓生活變得更便利，工作效率也有了顯著的提升，然而處在這種現代化的生活方式下，我們也開始面臨各種文明病的挑戰，其中憂鬱症的患者數近年來迅速增加，已成為一個備受全球關注的問題。根據聯合國世界衛生組織 (WHO, 2021) 統計，憂鬱症在 21 世紀被列為人類健康方面的首要威脅之一。在台灣方面據健保署統計，因憂鬱症而需用藥的患者數也是逐年增加，依據健保署統計資料 2017 年到 2021 年，這 5 年用藥人數成長 17.4% (衛生福利部中央健康保險署, 2021)。另外，用藥的原因雖然眾多，並不能直接等同於臨床身心疾患診斷，但卻意味著有更多人尋求醫療手段來處理情緒困擾的議題。因此，相比於憂鬱症的問題，更多的人還受到情緒問題的困擾。美國 2019 年的研究報告 (Center for Collegiate Mental Health, 2020) 指出，就讀於美國的大學生在心理諮詢、藥物治療和住院治療的比例在過去九年中有顯著增加。接受心理諮詢的學生增加了 10%，而其中一些學生甚至表現出自殺的傾向。類似的心理健康問題在台灣的年輕人中也逐漸顯現。

由於醫療資源有限和亞洲較為保守的文化背景，並非每個人都能從中受益。基於這些考慮，本研究提出了一個想法：如果能開發一款能夠隨時聆聽並對情緒困擾者提供正面回應的聊天機器人，可能會對他們擺脫抑鬱情緒有所幫助。正如心理治療專家 Brandon Santan 博士所指出，向他們展示關心和提供支持，讓他們知道有人理解和共鳴他們的感受，是對待受情緒困擾者的最佳方法 (DALAL, 2021)。這種支持和理解對於他們以正面情緒面對生活可能至關重要。

自然語言處理 (Natural Language Processing, 簡稱 NLP) 是人工智慧領域中的一個熱門研究方向。由於人類語言的高度複雜性和非結構化特點，許多研究者開始開發模型來讓電腦理解人類語言。Vaswani et al. (2017) 提出的 Transformer 模型在語言翻譯等 NLP 任務中取得了顯著進展。基於此架構，後續研究人員開發了多種變體，如 BERT 和 GPT 等，在不同 NLP 任務上也展現出卓越表現。

在自然語言生成和理解領域取得顯著進展的同時，對於情緒困擾者而言，找到能理解並提供情感支持的「傾聽者」至關重要。然而，現有模型通常只能讀取使用者輸入的文字。但正如 Lin et al. (2019) 和 Zandie & Mahoor (2020) 所指出，將情感元素融入對話生成過程，不僅能讓患者感受到同理心和情感支持，還能激發他們積極的情感態度，從而提高對話系統的使用動機。考慮情緒因素是為情緒困擾者設計對話生成系統的關鍵部分。

二、研究目的

基於前面的討論，不論是憂鬱症患者還是經歷情緒困擾的人們，許多人可能傾向於認為藥物治療能帶來康復。然而，心理治療的重要性不容忽視。如 (Engel, 1980) 提出的「生物—心理—社會」模式 (Biopsychosocial Model) 所強調的，除了生物因素外，心理和社會因素也是影響個人健康和恢復過程的關鍵因素。因此，主要研究目的為：

(三) 建立一個端到端的、具有即時陪伴和同理心的日常陪伴聊天機器人

(四) 將情緒資訊和術語識別技術結合到模型中，以增強模型回覆能力

貳、文獻探討

五、社會支持

社會支持 (陳德倫, 2020) 作為心理學領域的一個概念, 旨在降低壓力對個人產生負面影響方面起著保護作用。這種支持包括正面情緒支持、適度的肯定對當事人的感受、提供資訊、建議或實質幫助。本研究的主要關注點在於情緒性支持, 這意味著提供一種「鼓勵、個人溫暖、愛或情感的支持」。這種支持包括表達關心、建立信任和展現同理心 (Leavy, 1983)。希望透過情緒性的支持來幫助憂鬱症患者或情緒困擾者, 使他們在面對任何事情時都能保持正向的態度。

六、Seq2Seq 模型

Seq2Seq 模型是由 (Sutskever et al., 2014) 提出的一種時間循環神經網路的變型。這個模型通過引入編碼器-解碼器架構來改進傳統的循環神經網路。編碼器負責將輸入轉換成一個嵌入空間中的向量, 即上下文向量, 以捕捉輸入的語義信息。解碼器則使用這個向量來生成對應的輸出序列。這種模型特別適合於處理自然語言中的序列對序列任務, 如機器翻譯。Seq2Seq 模型的首要應用是在機器翻譯領域, 它讓機器可以不依賴人工規則而直接從大量訓練數據學習語言之間的特徵轉換。

(Cho et al., 2014) 提出了一個融合傳統統計機器翻譯模型(SMT)與深度學習循環神經網路(RNN)的混合架構。這種架構結合了傳統方法的穩定性與深度學習的學習能力, 從而顯著提升了機器翻譯的效能。此外, Seq2Seq 模型也已應用於其他領域, 例如問答系統。在這方面, (Ngai et al., 2021) 採用了基於 Seq2Seq 的 Transformer 模型來處理 COVID-19 相關的問答任務。

七、預訓練模型

隨著硬體技術的進步和數據量的增加, 預訓練模型已成為深度學習的主流策略之一。這種方法基於遷移學習的概念, 即先在一個特定的源任務上訓練模型, 然後將其遷移到其他目標任務。這種策略的核心是利用源任務中獲得的知識和特徵來加速目標任務的學習效果。預訓練模型通常在大量未標籤數據上進行自監督學習, 成為一個高效的特徵提取器, 在特定任務上進行微調以快速達到最優效果。

預訓練模型的策略已經被證明在多個領域, 如自然語言處理和電腦視覺, 具有極高的效能。特別在自然語言處理領域, BERT 和 GPT 這樣的模型已經顯示了其在各種語言任務中的優越性能。例如, (Liu et al., 2019) 提出的 RoBERTa 模型, 可說是 BERT 的進階版本, 它不僅利用了更大的數據集和更長的訓練時間, 更特別的是它採用了動態遮罩來預測詞彙, 並且選擇不使用 NSP 任務, 這些調整使 RoBERTa 在自然語言處理任務上比 BERT 展現出更為優異的性能。(Zhang et al., 2019) 提出了 DIALOGPT, 這是一個在 Reddit 對話資料上訓練和微調的大規模、可調整的對話模型。該模型基於 Transformer 的解碼器 GPT 模型, 它在單輪對話設定中的自動和人類評估方面實現接近人類的性能, 也比一般的基礎模型更能產生相關、內容更豐富且情境一致的回應。下表 1 為 BERT 和

GPT 兩者比較。

表 1：BERT 和 GPT 差異

	BERT	GPT
訓練策略	Masked Language Model (MLM)	Autoregressive Language Model (ALM)
考慮上下文	雙向	單向
預訓練目標	預測被遮蓋的詞語	預測句子中的下一個詞語
使用的網路層	Transformers 的編碼器	Transformers 的解碼器
用途	自然語言理解	自然語言生成

基於上述 2 種不同模型所導致的優劣勢，(Lewis et al., 2019) 提出 Bidirectional and Auto-Regressive Transformers (BART) 模型，它是一種從 Transformers 架構衍生出來的生成式預訓練模型，其結合了雙向和自回歸(即 Seq2Seq 架構)，也就是除了有 BERT 可以查看整段句子的資訊外，也有 GPT 的特性，使生成的文本只考慮當前時間點之前的詞彙，從而既可以用於理解文本，也可以用於生成文本。訓練策略方面，則比原始的 BERT 克漏字填空 (MLM) 還要更多樣性，包括遮罩一段連續的詞彙或是將句子順序打亂等操作，透過不同的給噪音方式，可以讓 BART 更加學習到語義和語法的結構。該模型結果也表明與 RoBERTa 和 XLNet 等大型語言模型相比，BART 在自然語言理解任務上的效果與它們相當，但在生成文本的任務上，它的表現更為突出。

八、標籤資訊引入模型

儘管 Seq2Seq 模型已被廣泛應用在各個領域，由於該模型都是藉由最大概似機率 (Maximum likelihood probability) 來產生回應，這使得它的輸出容易受到訓練資料中高頻句子的影響。這導致了兩個主要問題：一是生成的句子傾向於重複和模仿資料集中的常見句型；二是其回應往往缺乏多樣性和創意，使得結果看起來較為單一。為了克服生成回應的單調性，(Holtzman et al., 2019) 在解碼器生成部分擺脫了最大概似機率 Beam Search 等方式，他們引入了 Nucleus Sampling 方法，這種方法不僅僅選擇最高概率的詞彙，而是在一個概率核心集中隨機選擇詞彙。這允許模型生成更具多樣性和新穎性的句子，減少了生成結果的單調性。

而 (Li et al., 2016) 研究中嘗試將個人資訊也加入到模型訓練中，藉由個人資訊讓模型產生一致性和多樣性。(Lin et al., 2019) 研究中將 32 種情緒標籤進行模型訓練，他們設計了一種模型，該模型能夠在編碼階段識別使用者的當前情緒，然後使用多個解碼器專門處理不同的情緒。(Zandie & Mahoor, 2020) 則是將主題、意圖和情緒 3 種標籤轉為

詞嵌入後加入到文本中，在自動評估指標上有超越其他模型的表現。上述研究都表明通過融入額外的標籤，不僅可以提高模型的理解和生成能力，還可以使其生成的句子更具一致性和多樣性。

九、術語識別

預訓練模型基於大規模的數據集而建立，所以當要將其適用於某特定任務或領域時，往往需要額外的調整，先前的工作主要有 2 種方式，第一種就是透過為微調的方式 (Zeng et al., 2020; Li & Liang, 2021)，也就是拿預訓練好的模型應用在自己特定的任務上面在做訓練，第二種是將知識注入到具有額外訓練目標的語言模型中以提高語言理解。

第二種方式近年來也受到越來越多的關注，尤其應用在醫療領域，(Tang et al., 2023) 提出一種術語識別的醫療對話生成，由於醫療諮詢涉及到大量的特定領域知識，因此如果沒有相關醫學背景知識，模型很難產生有品質回復，另外醫生和病人的交流中，可能會涵蓋醫學和非醫學相關的話題，某個程度上模型或許只需要理解關鍵醫學詞彙就可以，其他詞彙有沒有理解可能不重要，因此該研究利用注意力機制來專注於醫學術語，並使模型強化對這些術語的學習，從而提高模型的整體理解和回應品質。

在本研究的情緒支持對話中，確保語言模型能夠準確識別並回應患者的當前情感非常重要。由於這些對話經常包含大量非特定或情緒中立的詞彙，這可能使模型難以捕捉到患者的真實情感。因此，導入識別關鍵字的策略(即前面提及的術語識別技術)作為額外的訓練目標，有助於模型更專注於患者的核心情感，從而提高回應的相關性和精準度。

參、研究方法

七、研究架構

本研究架構分為兩大階段。首先，在初階階段，系統對使用者輸入的對話進行情緒識別，將識別的情緒標籤轉化為特定 token，並與原對話內容結合。次階階段中，對合併的句子實施關鍵字註釋，確保每個關鍵詞彙都被明確標記。最後，這些整合後的內容輸入到深度學習模型 BART，生成回應給使用者。完整的聊天系統流程詳見圖 1。

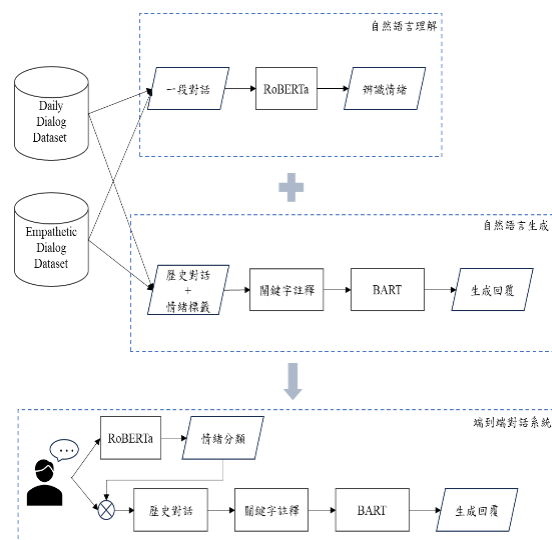


圖 1：端到端對話系統流程圖

八、資料集

本研究主要使用了 Rashkin et al. (2018) 提出的同理心對話數據集 (Empathetic Dialogues Dataset) 作為訓練資料來源。考慮到該數據集的規模較小，可能不足以支持模型達到理想的泛化能力，因此我們額外引入了 DailyDialog 數據集 (Li et al., 2017)，希望透過結合這兩大數據源來提升模型的泛化表現。由於這兩個數據集的附加標籤不一致，本研究僅針對情緒標籤進行考慮。

九、情緒標籤前處理

本研究使用的兩組資料集各有獨特的情緒標籤分類。為提高模型訓練效益並減少特殊 token 的使用，我們整合了情緒標籤，將原始的 39 個情緒標籤綜合為 9 個主要類別。這一整合參考了 Goel et al. (2021) 的方法，旨在平衡不同情緒標籤的資料量。針對數據量分布不均的問題，本研究採用資料擴增和降採樣策略。具體措施包括使用同義詞替換和故意錯拼單字，增加稀少標籤的數量，並透過降採樣解決樣本過多的情況。這些策略不僅豐富了資料集，也提高了模型對不同情緒的識別能力。資料擴增後各情緒標籤的新數量分布詳見下表 2。

表 2：經資料擴增後的情緒標籤數量分布

情緒類別	資料擴增前	資料擴增後
No_Emotion	61001	23000
Joyful	12842	20842
Afraid	2174	23914
Guilty	2279	20511
Angry	2959	20713
Hopeful	2270	20430
Sad	4034	20170
Proud	1624	19488
Confident	1600	19200

十、模型架構

(一) 情緒分類模型

採用了由 (Loureiro et al., 2022) 提出的基於 RoBERTa 的情感分析模型。然而，考慮到該模型是做 3 種主要情感分類，與我們資料集的情緒標籤不完全吻合，為了更好地適應我們的需求，我們對模型進行了必要的修改。具體來說，我們調整了模型的最後一層分類器，將其輸出類別從原本的三類調整為九類，以符合本研究資料集的情緒標籤需求。

在訓練策略方面，目標是這樣描述的：假設一個標記資料集 C ，給定一段使用者輸入序列 $X = \{x_1, x_2, \dots, x_n\}$ ，該目標是產生對應的情緒標籤 Y 。我們使用交叉熵損失函數，其實質上是在進行最大似然估計。我們的目標是最大化對應情緒標籤 Y 的條件概率，即 $P(Y|X)$ 。損失函數如公式 1 所示。

$$L_m(C) = - \sum_{(X,Y)} \log P(Y|x_1, x_2, \dots, x_n)$$

(1)

(二) 對話生成模型

1. 輸入格式

模型輸入格式中，我們將前一節情緒分類模型得到的輸出轉換為特定 token，並將其置於句首以引導情緒相關回應的生成。此外，透過自動註釋機制，特定關鍵詞前加入標記[TERM]，以強調對話中的關鍵信息。例如，句子 X 輸入模型時，會按以下公式 2 進行處理：

$$X_{term} = Identify(X), x_i = \begin{cases} x_T, x_i, x_i \text{ is term} \\ x_i, \text{ otherwise} \end{cases} \quad (2)$$

其中 x_T 表示 [TERM]， x_i 表示一個序列中的第 i 個 token， X_{term} 代表對輸入序列 X 經過函數後形成的。 $Identify$ 可以視為一個遠距離監督函數，這個方法的基本假設是，可以利用現有的知識庫（例如 Wikipedia、Freebase 等）來推斷出文本數據中實體之間的關係。在本研究中，使用了 Tang et al. (2023) 提供的字典，並添加了自定義的情緒標籤做為本研究的字典，以利後續註釋。最終模型輸入格式如圖 2 所示。

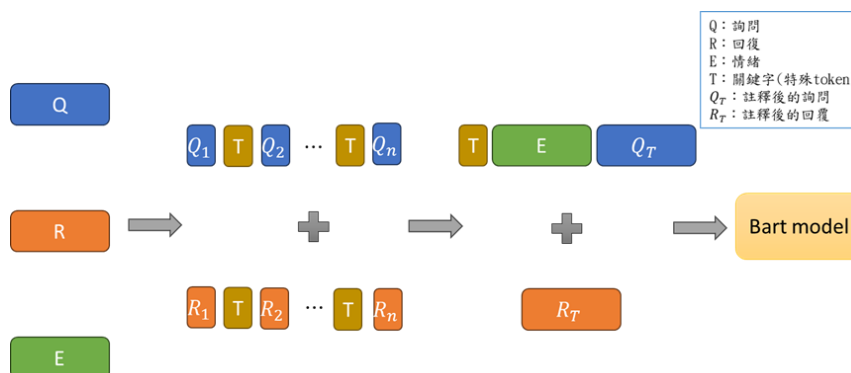


圖 2：模型輸入格式

2. 模型設計

本研究以 BART 模型為基礎上改良，除了保留了原始 BART 模型的語言模型頭部 (Language Modeling Head)，其負責基於上下文預測下一個詞。另外，我們在編碼器的輸出階段增加了一層新的線性轉換層。這一層的關鍵作用是對編碼器輸出的特徵向量 (Representation) 進行進一步處理，賦予模型識別關鍵詞的能力。這一新增層的架構與實現細節在圖 3 中展示。

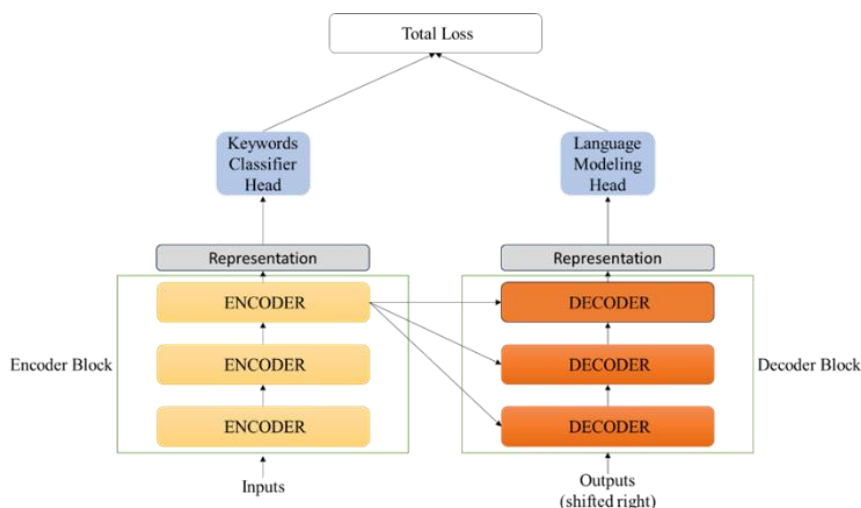


圖 3：模型設計圖

3. 損失函數設計

本研究採用 BART (編碼器-解碼器結構) 作為基礎模型。每個輸入序列的特徵都使用注意力機制進行編碼，並且這些特徵被自回歸解碼以預測下一個詞彙，如公式 3~6 中對此計算進行了具體的說明。

$$F = \text{Encoder}(X_{term}) \quad (3)$$

$$O_i = \text{Decoder}(y_{<i}, F) \quad (4)$$

$$P(y_i | y_{<i}, X) = \text{softmax}(W O_i + b) \quad (5)$$

$$L_{lm} = -\frac{1}{N} \sum_{n=1}^N \log P(Y | x_1, x_2, \dots, x_n) \quad (6)$$

其中， X_{term} 代表為加入情緒資訊和關鍵字註釋後的輸入， F 代表經過編碼器處理後的特徵，而 O_i 表示在第 i 個位置的解碼特徵。在計算下一個詞的概率 $P(y_i | y_{<i}, X)$ 時，我們使用 softmax 函數和訓練參數 W 和 b ，這裡指的是神經網路，也就是 Language Modeling Head。最終， L_{lm} 表示的是對於模型預測的 token 與標準答案之間的交叉熵損失，用於衡量和優化模型的性能。

在模型設計中，本研究額外新增了一個 Keywords Classifier 層，目的在於訓練模型編碼器識別出各個 token 是否為關鍵字。於此過程中，我們採用交叉熵作為損失函數。其計算方法在公式 7 中所示。

$$\tilde{k}_i = \text{classifier}(F_i), \quad L_{kc} = -k \log P(\tilde{k}_i) - (1 - k) \log P(1 - \tilde{k}_i) \quad (7)$$

其中第 i 個位置的 token F_i 編碼特徵被映射到分類標籤 \tilde{k}_i ，表示它是否為關鍵字， classifier 由一個全連接網路組成，用作將得到的編碼器特徵做關鍵字分類，且我們將 \tilde{k}_i 定義如公式 8， L_{kc} 表示的是對於二元分類結果的交叉熵損失。

$$\tilde{k}_i = \begin{cases} 1, & x_i \text{ is term} \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

最終，將這兩個訓練目標在微調階段透過最小化 $L_{overall}$ 結合，以整合學習效果：

$$L_{overall} = L_{lm} + L_{kc} \quad (9)$$

肆、實驗結果

五、實驗設定

本研究使用 RoBERTa 進行情緒分類，為 BART 生成模型提供情緒脈絡。此外，研究也探討了其他知名模型的效能，包括：

- (一) GPT2：一個全基於 Transformer decoder 結構，採用 cross-entropy 作為損失函數。
- (二) DialoGPT：類似 GPT2，但專門用 Reddit 對話進行訓練。
- (三) T5：採用 Transformer 的 encoder 和 decoder 結構，將所有任務統一為文本到文本格式處理。
- (四) UniLM：單一 Transformer encoder 支援多種 NLP 任務，通過不同預訓練策略達成統一性。
- (五) BART：基於 Transformer 的 encoder 和 decoder 所組成。

六、參數設定

實驗使用的預訓練模型基於 Hugging Face 的公開 checkpoints，在 NVIDIA GTX1080Ti 上運行，以 Adam 優化器進行訓練。情緒分類採用 $1e-6$ 的學習率和 8 的 batch size，並設定梯度累積為 8。訓練過程中，我們使用 early stopping，在 5 個 epoch 無進展時停止並儲存權重。對話生成的學習率設為 $2e-5$ ，batch size 為 2，梯度累積為 12，經過 6 個 epoch 的訓練後，選擇驗證集中 Loss 最低的權重作為最終模型。

七、自動評估結果

本研究以 Perplexity (PPL) 衡量模型對生成句子的信心，PPL 越低表示信心越高。同時，運用 BLEU 和 ROUGE 評估指標來測量生成句子。其各個模型跑出來的結果如下表 3。表 4 為本研究額外針對提出的兩種方法進行分別試驗。

表 3：模型自動評估結果

	ppl↓	B-1↑	B-2↑	B-3↑	B-4↑	R-1↑	R-2↑	R-L↑
GPT2	3.351	0.089	0.044	0.028	0.018	13.803	2.104	10.110
DialoGPT2	3.331	0.089	0.045	0.028	0.018	14.039	2.216	10.287
T5	2.771	0.162	0.083	0.052	0.036	13.399	2.301	11.872
UniLM	45.617	0.167	0.038	0.006	0.002	13.044	0.701	11.272
BART	3.027	0.166	0.095	0.065	0.047	18.100	4.612	16.476
Our(BART) (terms+label)	2.765	0.182	0.104	0.072	0.051	18.448	4.650	16.528

表 4：關鍵字註釋和情緒標籤之 BART 模型性能分析

	ppl↓	B-1↑	B-2↑	B-3↑	B-4↑	R-1↑	R-2↑	R-L↑
BART	3.027	0.166	0.095	0.065	0.047	18.100	4.612	16.476
BART+terms	2.773	0.169	0.095	0.065	0.046	18.269	4.663	16.514
BART+label	3.019	0.164	0.092	0.063	0.045	18.057	4.770	16.464

BART+terms +label	2.765	0.182	0.104	0.072	0.051	18.448	4.650	16.528
----------------------	-------	-------	-------	-------	-------	--------	-------	--------

其中表格裡的 terms 表示加入了關鍵字註釋，label 則是加入了情緒標籤給模型，此外，在計算 BLEU 得分時，B-2 之後的數字表示我們使用連續的字數來評估文本的相似度。舉例來說，B-2 意味著我們比較的是連續的兩個字與標準答案中對應的部分。如果標準答案中也出現了這樣的連續兩個字，這將使得 BLEU 得分增加，表示與真實答案的相似度更高。ROUGE 評分標準同樣是基於這種比較連續詞序列的方法。

從表 3 結果中，我們的模型在多項評估指標上都超越了其他模型，其中 UniLM 的性能最弱。進一步的比較顯示，UniLM 在單詞相似度上表現尚可，甚至在 BLEU-1 指標上領先。但對於多詞連貫性，其表現銳減，這或許與其作為編碼器的設計有關，並不適合生成流暢對話。而結合了編碼器和解碼器的 T5 與 BART 在生成任務上明顯優於僅用解碼器的 GPT 和 DialoGPT2，顯示出整合編碼器和解碼器架構在對話生成方面的潛在優勢。

從表 4 的結果來看，引入關鍵字註記確實提升了模型的效能，無論是從困惑度還是相似度指標來看，這種方法都使模型的回答更加貼近真實答案。此外，加入情緒標籤對提升模型性能也有幫助，儘管在某些指標上仍優於原始的 BART 模型，但由於標籤數據的多樣性和數量限制，其效果有限。觀察其他研究發現，它們通常會向模型輸入加入大量標籤，這不僅限於當前用戶的情緒，還可能包括討論的主題或句子展現的行為等。因此，未來可以提供更多樣標籤供模型有更多資訊可學習。

八、情緒標籤分析

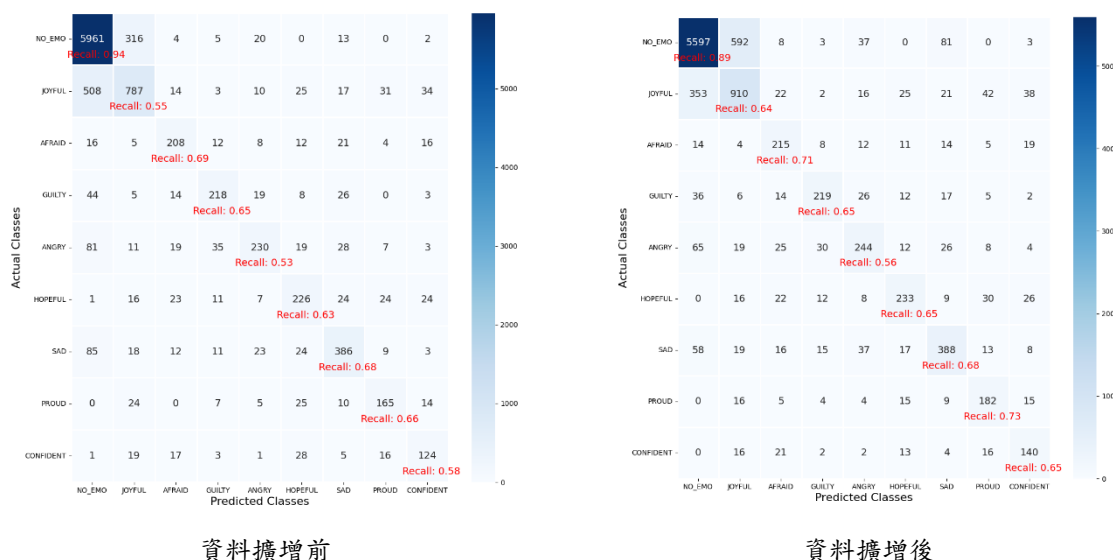


圖 4：資料增強技術實施前後可視化混淆矩陣的比對

就圖 4 結果來看，儘管總體準確率僅相差 1%，但考慮到本研究聚焦於提供情緒支持的對話，關鍵在於模型能否廣泛識別可能面臨情緒困擾或憂鬱傾向的個體。如果召回率低，則可能忽略需要幫助的個體。總體而言，各個情緒指標的召回率至少持平或有所提升。因此，本研究確認資料擴增對我們任務是有幫助的。然而，進一步分析，可以看出該召回率提升主要都是情緒正面類別，而較負面情緒類別，也被視為該任務中較需

要識別出來的情況下，但在此次資料擴增後並未顯示出相應程度的改善。這一發現也指出了未來研究中需要改進的方向。

九、情緒標籤對生成模型影響

為了驗證模型是否真正掌握了那些標籤所表達的含義，我們對一個初始句子指定的特定情緒標籤進行了隨機更換，觀察是否會影響模型生成品質，結果如下表 5。

表 5：隨機情緒標籤對生成影響

	B-1↑	B-2↑	B-3↑	B-4↑	R-1↑	R-2↑	R-L↑
BART + 情緒標籤	0.182	0.104	0.072	0.051	18.448	4.650	16.529
BART + 情緒標籤 (隨機)	0.180	0.102	0.070	0.050	18.413	4.648	16.590

表 5 展示的數據通過比較在句子中精確放置情緒標籤與隨機指定情緒標籤的效果，顯示出即便相似度等核心指標上只有輕微的降低，也突顯了準確識別使用者情緒在使模型生成適切回應方面的重要性。這一結果激發我們尋求方法提高模型的情緒識別精準度（例如前面用到的資料擴增技術），以期進一步提升對話生成的相關性與品質。

本研究認為，目前使用的單一標籤對於結果的影響似乎有限。若能在標籤系統中引入更多元的資訊，如當前討論主題、使用者的具體行動等，並且在模型的輸出中也考慮到回覆的情緒識別，則隨機分配標籤的影響可能變得更加顯著。這樣的擴充不但能提高模型對於情緒的敏感度，也可能進一步提升對話產生的品質和相關性。

十、關鍵字註釋對生成模型影響

為了增進模型對對話內容的理解深度，本研究運用關鍵字註釋技巧以增強模型識別對話中詞彙相互關係的能力。下圖將展示詞彙關聯度的可視化，以揭示模型如何處理這些資訊。

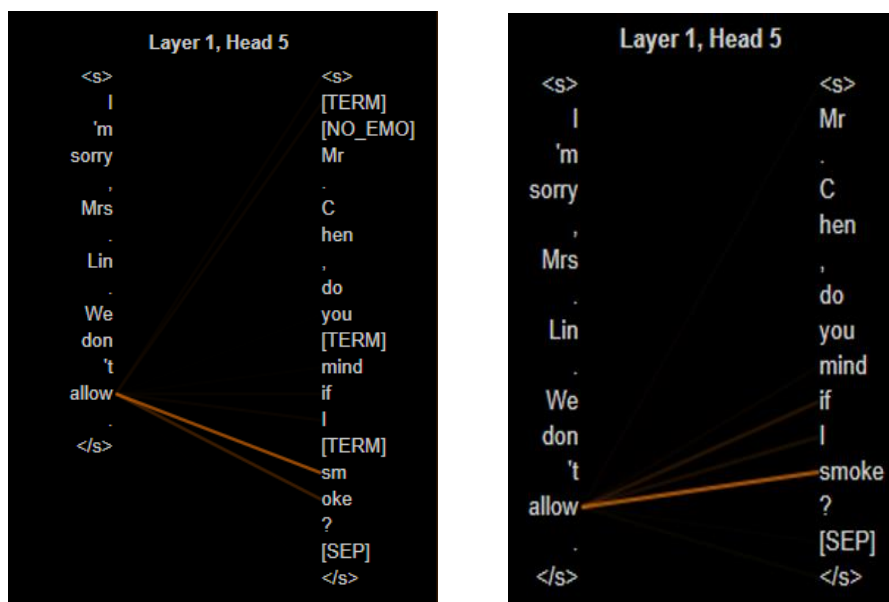


圖 5：詞彙關聯度可視化

圖 5 左側顯示的是我們基於 BART 所修改的模型，而右側則展示了原始 BART 模型的結果。例句中的“we don’t allow”指涉的是禁止吸煙這一行為，這一點從兩個模型的視覺化都可以看出。然而，在確認這一詞群關聯上，修改後的 BART 模型確定性似乎更高，這可以從其注意力分布的集中度看出，儘管兩模型顏色相近，修改後的 BART 模型的注意力分支較少。本研究推斷，所提出的模型之所以表現得更佳，是因為引入了關鍵字註釋，使得模型在處理關鍵信息時能夠更集中，不易受到其他非關鍵詞彙的分散。

值得注意的是，使用於本研究中的關鍵字註釋字典最初是為醫療領域設計的，若此字典能夠針對情緒支持對話進行專門的定制，其效果可能會更加顯著。

十一、 人工評估結果

儘管自動化評估指標提供了一定的量化分析，它們可能無法全面反映模型生成回答的實際品質。考慮到人類情感因素，程序化的回答可能不總是高品質的。為了獲得更深入的評價，我們邀請了十位具有大專院校以上學歷的受試者對選自測試集的 30 條單輪對話和 20 條透過人機互動生成的多輪對話資料進行人工評估，共計 50 條對話。評估方式包括對模型回覆的幫助性、品質和相關性三個指標，分數範圍從 1（最低）到 5（最高）。最終結果將展示在後續的表格中，以評價交互式對話系統的整體表現。

表 6：人工評估指標定義

	指標定義
幫助性	評估此對話是否有幫助(得到情緒上的支持)
品質	評估此回覆的內容品質、互動性和完整性。
相關性	評估此回覆與問題是否有相同主題。

表 7：對話系統串連範例

對話







 ：我馬上就要考試了，我已經準備好了，但我還是很緊張。	情緒：緊張
 ：我相信你會做得很好，只是為真正的事情做好準備。這將是值得的。	
 ：你能祝我考試取得好成績嗎？謝謝你，先生。	情緒：開心
 ：不客氣，我相信你會做得很好。我希望你考試愉快。	
 ：謝謝，再見~	情緒：開心
 ：再見，下次考試時再見。希望你今天過得好！	

表 8：對話評估結果

	幫助性	品質	相關性
總體	3.670	3.456	3.698
單輪對話	3.710	3.500	3.720
多輪對話	3.615	3.395	3.660

根據表 8 的評估結果顯示，使用者普遍對模型在回應問題的相關性給予正面評價，顯示模型能根據主題進行相應的回覆。然而，在幫助性指標上得分高於品質，這暗示了雖然模型的回答在一定程度上滿足了使用者的情緒需求，但可能因回答表面化，沒有深入探討或真正貼近使用者內心的感受。觀察表 7 中的對話範例，我們可以看到，儘管模型的回應在某程度上是支持性的，內容卻普遍籠統且缺少深入分析，加上後續回覆的缺乏連貫性及出現重複性，都是導致品質分數不佳的因素。

評估結果從單輪對比多輪對話的得分來看，顯示出單輪對話在所有評分指標上都優於多輪對話。在這些指標中，相關性得分最高，其次是幫助性，品質則排在後面。這一趨勢進一步揭示了，多輪對話中由於回答的重複性和缺乏流暢性，可能對整體評分產生了不利影響，從而使得在多個評估標準上，多輪對話的得分落後於單輪對話。

綜合來說，本研究開發的模型在回答用戶問題上展現了一定的能力。然而，在對話的流暢連接與討論深度方面，表現並不盡如人意。此外，本研究發現，模型面對訓練集中較少見或新領域的問題時，常出現答非所問。而且，模型似乎也有一種固定模式，會在回應負面評語時經常使用“聽到這個消息我很遺憾”作為開場白，這顯示出資料集在負面情境下的回答過於單一。此外，情緒辨別的模型也並不是都很準確，往往有時候

簡單的句子也會辨識錯誤，這些因素共同導致了模型在保持對話連貫性和深入討論方面的不足。未來的研究方向可以考慮擴充和豐富資料集，提升對話的品質和多樣性，進而提高模型的回答品質。

伍、結論

在本文中，我們設計並實現了一個創新的聊天機器人框架，此機器人具備情感支持功能，旨在幫助現今社會中日益增多的情緒困擾患者。面對情緒壓力及憂鬱症趨勢的上升，這種對話系統提供了一個及時且有效的溝通渠道，對於緩解個人的情緒問題具有潛在的重大意義。

本研究主要貢獻如下：

1. 融合情緒標籤與關鍵字註釋的對話生成技術：本研究首次結合使用者的情緒狀態與術語識別技術，將這些資訊融入對話模型訓練中。結果顯示，模型在自動評估中，相較於基線模型，BLEU 指標提升 7% 至 10%，ROUGE 指標提升 0.3% 至 2%，人工評估表現也取得中上分數。
2. 分析關鍵字註釋與情緒標籤對生成模型的影響：實驗探討了情緒標籤和關鍵字註釋對對話生成的影響。結果發現，精確的情緒標籤顯著提高模型對語境的理解及對話質量。關鍵字註釋利用注意力機制增強模型捕捉關鍵語義的能力。
3. 端到端整合的對話系統設計：研究成功整合情緒分類和術語識別技術到對話生成模型中，實現端到端的對話系統。系統能即時識別使用者情緒，並將這些情緒資訊有效地融入到對話中並做註釋，以提供更加準確及人性化的回應。

參考文獻

- 陳德倫. (2020, October 19). 如何為情緒受苦者「撐傘」？陪他們走一段聆聽和理解的路. 報導者. <https://www.twreporter.org/a/high-academic-achievement-students-psychological-distress-medical-counseling-resources>
- 衛生福利部. (2023). 「年輕的心，有我傾聽」 「年輕族群心理健康支持方案」. 衛生福利部. <https://www.mohw.gov.tw/cp-16-75401-1.html>
- Dalal, C. (2021, April 19). 陪伴「憂鬱症」患者的 5 種方法，最重要的是傾聽、表達同理心與支持. Womenshealth. <https://www.womenshealthmag.com/tw/mental/relationship/g36108730/5-tips-for-helping-loved-one-with-depression/>
- Center for Collegiate Mental Health. (2020). 2019 Annual Report [PDF file]. Pennsylvania State University. <https://bpb-us->

e1.wpmucdn.com/sites.psu.edu/dist/3/3058/files/2020/03/2019-CCMH-Annual-Report_3.17.20.pdf

- Cho, K., Merriënboer, B.V., Gülçehre, Ç., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Conference on Empirical Methods in Natural Language Processing*.
- Engel, G.L. (1980). The clinical application of the biopsychosocial model. *The American journal of psychiatry*, 137 5, 535-44 .
- Goel, Raman et al. “Emotion-Aware Transformer Encoder for Empathetic Dialogue Generation.” *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (2021) : 1-6.
- Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The Curious Case of Neural Text Degeneration. *ArXiv, abs/1904.09751*.
- Leavy, R.L. (1983). Social support and psychological disorder: a review. *Journal of community psychology*, 11 1, 3-21 .
- Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Annual Meeting of the Association for Computational Linguistics*.
- Li, J., Galley, M., Brockett, C., Spithourakis, G.P., Gao, J., & Dolan, W.B. (2016). A Persona-Based Neural Conversation Model. *ArXiv, abs/1603.06155*.
- Li, Y., Su, H., Shen, X., Li, W., Cao, Z., & Niu, S. (2017). DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. *ArXiv, abs/1710.03957*.
- Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.
- Lin, Z., Madotto, A., Shin, J., Xu, P., & Fung, P. (2019). MoEL: Mixture of Empathetic Listeners. *ArXiv, abs/1908.07687*.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv, abs/1907.11692*.
- Loureiro, D., Barbieri, F., Neves, L., Anke, L.E., & Camacho-Collados, J. (2022). TimeLMs: Diachronic Language Models from Twitter. *Annual Meeting of the Association for Computational Linguistics*.
- Ngai, H., Park, Y., Chen, J., & Parsapoor, M. (2021). Transformer-based models for question answering on covid19. *arXiv preprint arXiv:2101.11432*.
- Rashkin, H., Smith, E.M., Li, M., & Boureau, Y. (2018). Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. *Annual Meeting of the Association for Computational Linguistics*.

- Sutskever, I., Vinyals, O., & Le, Q.V. (2014). Sequence to Sequence Learning with Neural Networks. *ArXiv, abs/1409.3215*.
- Tang, C., Zhang, H., Loakman, T., Lin, C., & Guerin, F. (2022). Terminology-aware Medical Dialogue Generation. *ArXiv, abs/2210.15551*.
- Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *NIPS*.
- WHO. (2021). WHO depression statistics.
- Zandie, R., & Mahoor, M.H. (2020). EmpTransfo: A Multi-head Transformer Architecture for Creating Empathetic Dialog Systems. *The Florida AI Research Society*.
- Zeng, G., Yang, W., Ju, Z., Yang, Y., Wang, S., Zhang, R., ... & Xie, P. (2020, November). MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 9241-9250).
- Zhang, Y., Sun, S., Galley, M., Chen, Y., Brockett, C., Gao, X., Gao, J., Liu, J., & Dolan, W.B. (2019). DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. *ArXiv, abs/1911.00536*.