

國立臺灣大學管理學院資訊管理研究所



碩士論文

Graduate Institute of Department of Information Management

College of Management

National Taiwan University

Master Thesis

以憂鬱症患者為對象之聊天機器人對話分析

Dialogue Analysis of

A Chatbot for Patients with Depression

黃郁云

Yu-Yun Huang

指導教授：莊裕澤 博士

Advisor: Yuh-Jzer Joung, Ph.D.

中華民國 111 年 9 月

September, 2022

誌謝



時光流逝，隨著論文的完成，在研究所的學習階段即將宣告結束，也必須告別陪伴我兩年時光的老師、同學夥伴與校園，非常不捨但也感謝這段旅程。

首先，能完成這篇論文，我要特別感謝我的指導教授莊裕澤老師，從碩一開始每周的 meeting 老師都會提供我們很多協助，從研究問題的發想開始，總是會提醒我們應該要如何了解論文並找到研究的重點，再到可以完整說明整篇論文的能力，有了這些幫助我才得以有研究方向並完成我的論文。除此之外，在寫這份論文的過程中，也謝謝老師給的許多意見及指導，提供不一樣的思維，讓我們可以有更多嶄新的想法並能在實驗過程中及時修正實驗方向，獲得更好的實驗結果。

另外，也要特別感謝實驗室的所有成員，尤其是我的論文合作夥伴成恩，我們從剛開始的發想和資料集的尋找上都遇到了一些困難，幸好最後都會討論出不錯的解決方案，謝謝我們都能默契合作才得以完成目標；還有人傑、安浩與翔岳，在各種課堂以及 meeting 準備中，我們都能彼此鼓勵，有你們一起學習、陪伴，讓我的碩班生活非常豐富精采，祝福大家未來都能順利、健康、平安、快樂。還有要感謝在國泰數據部實習遇到的同事，讓我學到很多事情，有了不一樣的經歷，並留下很多有趣的回憶。

最後要感謝我的家人們，有了你們的支持，我才能順利完成學業，取得碩士學位。尤其是我的阿嬤總是會打電話來關心我的論文進度，她其實不太知道論文到底是什麼，但我把論文形容成一個很大的考試，她總是會希望我能趕快考過，我終於要完成了，謝謝家人們的關心與協助。

黃郁云謹誌

國立臺灣大學資訊管理學系

民國一百一十一年九月



國立臺灣大學（碩）博士學位論文
口試委員會審定書

中文題目：以憂鬱症患者為對象之聊天機器人對話分析

英文題目：Dialogue Analysis of A Chatbot for Patients with
Depression

本論文係 黃郁云 君（學號 R09725053）在國立臺灣大學資訊管理學系、所完成之碩（博）士學位論文，於民國 111 年 8 月 22 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

（簽名）

黃郁云
盧信敏

陳建銘

系主任、所長

陳建銘

（簽名）

摘要



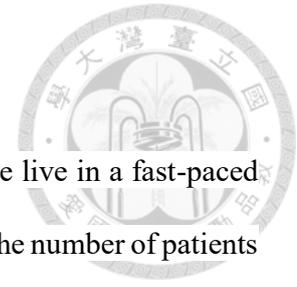
在這個科技發達的時代，現代人生活在充滿壓力的快步調社會，這也導致憂鬱症等精神疾病的患病者持續增加，而精神疾病的盛行所導致的自殺率提高，也會使社會負擔加重。雖然心理疾病的患病者越來越多，本研究發現許多患者是不願意求助的，甚至沒有病識感，不認為自己已經患病，他們缺乏一個管道可以抒發情緒，並感受到被理解與陪伴。

為了解決上述問題，本研究與研究室夥伴(Su, 2022)共同提出了一個針對憂鬱症患者的聊天機器人架構，我們認為一個完善的對話系統可以解決他們即時的情緒需求，給予他們所需要的理解與陪伴。為了設計一個符合憂鬱症患者需求的對話系統，本研究認為對話中的情緒、意圖、主題、態度是有助於對話理解的，並能達成同理患者心情的對話生成。本研究主要透過這些標籤進行對話的分析，期望能透過對話準確的標籤分類，生成正面積極的回覆，滿足患者的傾訴需求。除了要回應使用者的需求，本研究也期望能喚起民眾的病識感，許多人在對話中會透露出自殺想法，我們期望能在對話中完成憂鬱程度的判斷，先確立患者的憂鬱程度，再進行對話的分析與生成，一旦判斷出對話中帶有自殺想法，更能達到即時通報，減少自殺事件發生的可能。

本研究主要的目標是達到準確的判斷與生成，因此我們使用 BERT 預訓練模型來進行對話的分析與判斷。本研究總共使用三個自行整理的資料集來進行憂鬱程度的判斷，分別為 RSP、GSN 和 MTD 資料集；並使用與(Su, 2022)共同進行標記的 ChinesePsyQA 來進行對話中的主題、情緒、意圖、態度四個標籤的分類。本研究針對標籤進行相關性的人工評估，結果顯示資料集的標籤是與大眾的理解相符的，而在對話分類的評估中，結果顯示我們選用的模型在兩個分類任務上，與之前相似的任務相比都取得了相當好的成果，最後從與(Su, 2022)合作完成的研究成果來看，本論文中準確的分類，是影響生成結果好壞的主要因素之一，也說明了主題、意圖、態度這三個標籤對於此架構來說是不可或缺的一部分。

關鍵字：憂鬱症、對話系統、聊天機器人、自然語言處理、深度學習、BERT 預訓練模型、文字探勘

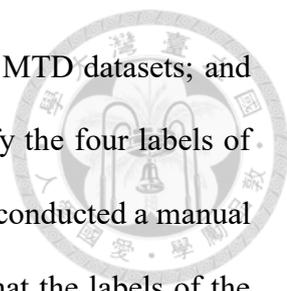
ABSTRACT



Today, in the age of advances in information technology, people live in a fast-paced society which full of stress, which also leads to a continuous increase in the number of patients with mental illnesses such as depression, and the increase in suicide rates caused by the prevalence of mental illnesses will also increase the social burden. Although there are more and more patients with mental illness, this study found that many patients are reluctant to seek help, and do not even have a sense of illness. They don't think they are sick. They lack a channel to express their emotions and feel understood and accompanied.

To solve the above problems, the study and my lab partners (Su, 2022) co-proposed a chatbot architecture for depression patients. We believe that a complete dialogue system can solve their immediate emotional needs and give them the understanding and companionship they need. To design a dialogue system that meets the needs of patients with depression, this study believes that sentiment, dialogue act, topic, and attitude in dialogue are helpful for dialogue comprehension, and can achieve dialogue generation that empathizes with patients' feelings. This study mainly analyzes the dialogue through these tags, hoping to generate positive responses through accurate label classification of the dialogue to meet the needs of patients. In addition to responding to the needs of users, this study also expects to arouse the public's sense of illness. Many people reveal suicidal thoughts in conversations. We expect to be able to analyze the depression level in the dialogue. First, we determine the depression level of the patients, and then analyze and generate the dialogue. If it is analyzed that there are suicidal thoughts in the dialogue, it will be able to achieve immediate notification and avoid something regrettable happening again.

The main goal of this study is to achieve accurate classification and generation, so we use the BERT pre-trained model to analyze the dialogue. This study uses three self-



labeled datasets to classify the depression level, the RSP, GSN, and MTD datasets; and uses ChinesePsyQA, which is co-labeled with (Su, 2022), to classify the four labels of sentiment, intention, topic, and attitude in the dialogue. This study conducted a manual evaluation of the confidence of the labels, and the results showed that the labels of the dataset were consistent with the public's understanding. In the evaluation of dialogue classification, the results show that our chosen model achieves fairly good results on both classification tasks compared to previous similar tasks. Finally, from the research results completed in cooperation with (Su, 2022), the accurate classification results in this study is one of the main factors affecting the quality of the generated results, and it also shows that these four labels are an integral part of this architecture.

Keywords: depression, dialogue system, chatbot, Natural Language Processing, Deep Learning, BERT pre-trained model, Text Mining

目 錄

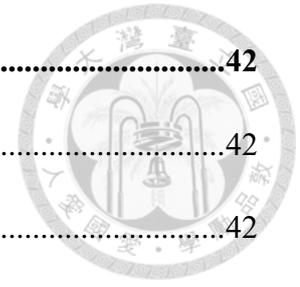


誌謝.....	i
口試委員會審定書.....	ii
摘要.....	iii
ABSTRACT.....	iv
目 錄.....	vi
圖目錄.....	ix
表目錄.....	x
第一章、緒論.....	1
1.1 研究背景與動機.....	1
1.2 研究目的.....	4
1.3 論文架構.....	5
第二章、文獻探討.....	6
2.1 憂鬱症治療.....	7
2.2 聊天機器人.....	7
2.2.1 閒聊機器人(Chit-Chat).....	8
2.2.2 醫療保健型聊天機器人.....	8
2.2.3 心理健康治療聊天機器人.....	9
2.3 基於 BERT 語言模型的對話判斷.....	10
2.3.1 自殺預防.....	11
2.3.2 情感分類.....	12
2.3.3 意圖分類.....	14



2.3.4	主題分類	14
2.4	BERT 預訓練模型	15
2.4.1	RoBERTa.....	15
2.4.2	Chinese BERT with Whole Word Masking.....	16
2.5	基於 GPT-2 的對話生成.....	17
2.6	翻譯評估	18
2.7	評估方法	18
2.8	文獻探討總結	20
第三章、研究方法		22
3.1	方法概述	22
3.2	資料集	25
3.2.1	ChinesePsyQA.....	25
3.2.2	RSP (Reddit Suicide Prevention Dataset)	27
3.2.3	GSN (Genuine Suicide Note)/ MTD (Multi-Turn Dialogue).....	29
3.3	翻譯質量評估	30
3.4	模型挑選	31
第四章、實驗細節與實驗結果		32
4.1	實驗設定	32
4.2	對話標籤相關性人工評估	33
4.3	憂鬱程度分類結果	36
4.4	對話分類結果	38
4.5	對話系統串接結果	40

第五章、結論.....	42
5.1 研究成果.....	42
5.2 研究貢獻.....	42
5.3 研究限制.....	43
5.4 未來研究方向.....	44
附錄.....	45
Reference.....	48



圖目錄



圖 1、2000 -2019 年在 Scopus 中按年份搜索 chatbot 相關關鍵字的結果.....	3
圖 2、混淆矩陣(Confusion Matrix)	18
圖 3、實作方法與架構	24
圖 4、ChinesePsyQA 主題分布情形.....	27
圖 5、ChinesePsyQA 情緒分布情形.....	27
圖 6、ChinesePsyQA 意圖分布情形.....	27
圖 7、ChinesePsyQA 態度分布情形.....	27
圖 8、RSP Dataset 類別分布	28
圖 9、MTD Dataset 各組句數分布.....	30
圖 10、資料集翻譯流程	31
圖 11、Task1 Confusion Matrix.....	38
圖 12、Task2 Confusion Matrix.....	38

表目錄



表 1 、 Whole word masking in BERT 範例.....	17
表 2 、 實驗階段內容.....	22
表 3 、 詳細分類標籤介紹.....	23
表 4 、 資料集說明.....	25
表 5 、 ChinesePsyQA 主題標記準則.....	26
表 6 、 ChinesePsyQA 回覆資料標記範例.....	26
表 7 、 ChinesePsyQA 主題標記資料範例.....	27
表 8 、 Daily Dialogue Dataset 原始資料集對話翻譯.....	28
表 9 、 Daily Dialogue Dataset 中英翻譯範例.....	29
表 10 、 Reddit 資料中英翻譯範例.....	29
表 11 、 RSP Dataset 範例.....	29
表 12 、 MTD Dataset 範例.....	30
表 13 、 憂鬱程度標籤相關性評估結果.....	33
表 14 、 憂鬱程度標籤問卷資料範例.....	34
表 15 、 對話標籤相關性評估結果.....	35
表 16 、 對話標籤相關性不佳範例.....	35
表 17 、 對話標籤相關性問卷負面情緒資料範例.....	36
表 18 、 憂鬱程度各項任務說明.....	37
表 19 、 憂鬱程度分類結果.....	37
表 20 、 對話分類結果.....	39
表 21 、 對話系統串接範例.....	40
表 22 、 對話問題人工評估結果.....	40

第一章、緒論



1.1 研究背景與動機

隨著時代變遷，科技越加發達，隨之而來的不僅僅是進步的社會，還有各式各樣的「文明病」，而憂鬱症就是一個例子。憂鬱症是一種情緒障礙，其會影響患者的感受、思考和行為方式，並可能導致各種情緒和身體問題，使患者無法正常生活。

現今，憂鬱症已經成為全球社會經濟負擔(Social Economic Burden) 最重大的疾病，患者容易受到疾病的影響，時而提不起勁、容易疲累、無法專注與負面思考，影響了社會功能與職場表現，導致生產力下降，根據哈佛大學的研究，造成人類社會整體疾病負擔 (Global burden of Disease) 前十名的疾病，第二名是憂鬱症。且根據世界衛生組織 (WHO) 研究發現，憂鬱症患者平均失能的天數為八天，顯著較沒有憂鬱者失能平均兩天為嚴重 (張家銘, 2006)；據估計，憂鬱症每年為美國雇主帶來 1000 億美元的憂鬱症成本，其中僅生產力成本一年損失就達 440 億美元 (Stewart et al., 2003)。據世界衛生組織 (WHO, 2021)對患病人數的統計，現在全世界大約有 5%的人正飽受憂鬱症所苦，且患病嚴重的人數正在繼續增加中。

另外，由於憂鬱症等精神疾病的盛行，自殺率不斷提高，心理健康的惡化常常導致人們產生無法繼續活下去的念頭。過去關於自殺死亡的心理解剖研究發現，至少有 90%的人自殺身亡前是患有精神障礙疾病的 (Louise et al., 2018)，包含憂鬱症、躁鬱症、焦慮症等疾病，但是以全國自殺防治中心之前分析自殺死亡者生前曾經被診斷憂鬱症的比例，只有 10.7%，代表了自殺者的憂鬱症診斷仍不足，也顯示醫師對於憂鬱症與自殺風險的診斷評估能力需要加強。另外在芬蘭的一研究中 (Isometsä et al., 1994)，有 75%的自殺者曾接受過精神科治療，有 2/3 的人在自殺前一年內接受過精神科治療，但是在他們自殺當下卻只有 45%正接受精神科治療，讓我們知道憂鬱症診斷的重要性，若是有充足的醫療是可以協助患者們擺脫自殺的想法的。



雖然在這個時代，由於新聞媒體與社會案件時常探討精神疾病相關議題，憂鬱症已經被大家所熟知。但仍有部分一般民眾對憂鬱症等精神疾病有嚴重的偏見，大眾對憂鬱症的誤解與標籤仍然存在，而也有很多患病的民眾不自覺、不承認或抗拒治療，導致台灣在憂鬱症醫療方面有很大的不足。另外，現今發展數位科技已成為全球趨勢，數位醫療也在疫情發展期間越受重視，全球醫療健康產業正在不斷融合人工智慧、物聯網、大數據等數位科技，使醫療服務朝智慧化的方向發展。因此我們認為可以利用數位醫療來解決憂鬱症醫療的問題。

結合前面所述的憂鬱症疾病的情況，大部分的患者缺乏溝通與理解，才會對自己的病症不夠了解，進而無法獲得醫療協助。而且目前的心理治療並無法做到無時無刻的陪伴，但是憂鬱症患者卻是非常需要溝通與陪伴的一個族群，因此本研究想要以達成線上醫療協助為目的，針對憂鬱症患者的需求，進行對話的判斷與生成，為此我們特別設計了幾項指標，用以判斷患者所想表達的完整內容。最後我們的目標是希望以完整的聊天機器人做為最終成果，補足醫療上的缺失。

聊天機器人是現在的熱門應用之一，有些機器人能夠分析高自由度的語句，但有些僅能透過一問一答逐步解決客戶問題，通常只為達成特定目標，如訂飛機票、一般常見問答機器人。2016 年是聊天機器人開始備受矚目的一年，Facebook 將 Messenger 機器人平台作為其 F8 開發者大會的核心。微軟的 Satya Nadella 將對話稱為 "third run-time" -- 操作平台不可或缺的一部分，僅次於操作系統和網絡瀏覽器。而根據 Scopus 內容分析，在 2016 年之後，人們對聊天機器人的興趣、需求迅速增長，增長情況如下圖 1 所示 (Adamopoulou & Moussiades, 2020)。越來越多企業開發、使用聊天機器人，不僅可以增加顧客滿意度，還可以降低人力成本。

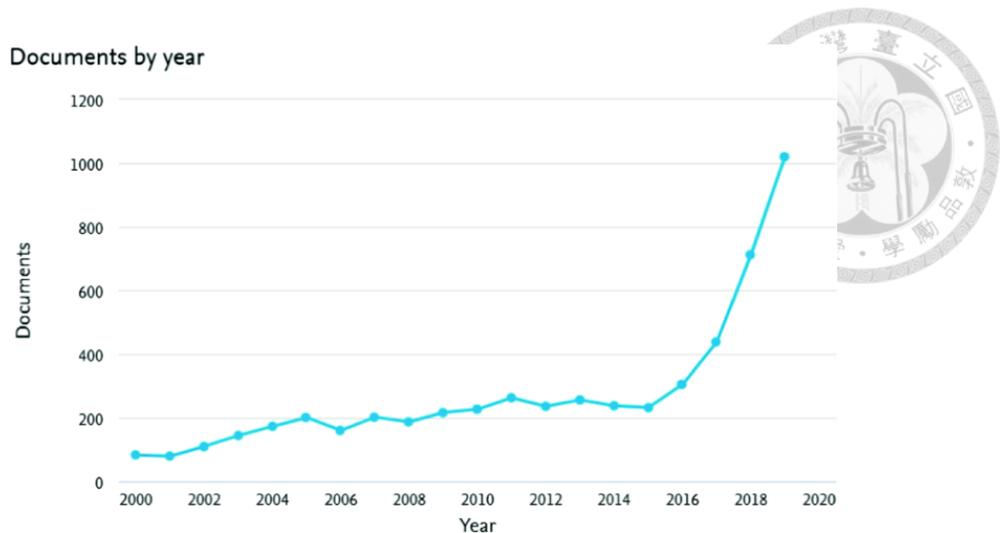
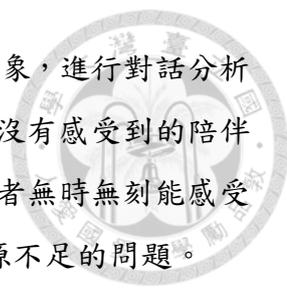


圖 1、2000 -2019 年在 Scopus 中按年份搜索 chatbot 相關關鍵字的結果
 關鍵字：“chatbot”，“conversation agent”，“conversational interface”
 (Adamopoulou & Moussiades, 2020)

在上述說明的聊天機器人中，本研究想要完成的是以自然語言生成方式來進行對話生成的聊天機器人，會運用自然語言處理 (NLP, Natural Language Processing) 技術來實現。現今也有許多應用都屬於自然語言的範疇，包含智能助理、文本生成、機器翻譯等，目的都是為一系列任務或應用程序實現與人類相似的語言 (Liddy, 2001)。

NLP 這項技術非常實用，能夠提供人類許多協助，因此也是熱門的研究題材之一，研究者們也相繼提出各式各樣的語言模型，目的是用來解決不同的任務需求。最廣為人知的是由 Google 以無監督方式訓練的 BERT (Bidirectional Encoder Representations from Transformers) Model，其架構主要由 Transformer 中的 Encoder 疊加構成，主要應用於文本的上下文填空及預測，BERT 經由上下文對語意的理解使 BERT 與前一代模型- Open AI GPT 相比成效好了許多；還有另一重要的模型，GPT-2 (Generative Pre-Training - 2)，其架構為 Transformer 中的 Decoder，主要應用於文本生成任務，值得一提的是，其訓練模型時所使用的數據以及參數量都是前所未有地龐大，是有史以來最多參數的語言模型，因為數據量足夠龐大，GPT-2 生成的文本非常流暢，甚至難以分辨是否由機器生成。

綜合上述意見，憂鬱症是一個嚴重也急需被解決的社會問題，但目前的醫療現況卻不足以應付憂鬱症患者的需求，且自殺率節節升高，也是嚴重的社會



問題。為了解決此問題，本研究期望能以憂鬱症患者為主要對象，進行對話分析與生成，從對話中獲取相應的資訊，得以提供患者日常生活中沒有感受到的陪伴與理解，並且能在與患者溝通的過程中，緩和患者情緒，讓患者無時無刻能感受到醫療的協助，最終可以彌補現今在心理治療方面，醫療資源不足的問題。

1.2 研究目的

本研究之終極目標為提出一個醫療陪伴型中文對話系統，改善現今憂鬱症患者就醫狀況與醫療資源不足的問題，藉由隨時的監測與陪伴，讓患者如同擁有一個虛擬醫生，患者可以透過與機器人對話，隨時理解自己的病症狀況，我們在解決患者情緒問題的同時也給予他們足夠的陪伴。而本研究會將問題鎖定在針對憂鬱症患者之中文的對話分類判斷，了解患者真實的溝通意圖，並提供適當的回應。另外，我們也會進行使用者對話的憂鬱程度判斷，希望能達到自殺行為的預防。

本研究主要分為三個階段。第一階段會進行憂鬱程度的判斷，第二階段則會判斷使用者對話中所提供的情緒 (Sentiment)、主題 (Topic)、意圖 (Dialogue Act)、態度 (Attitude) 方面，前兩階段為本論文主要的研究內容，我們針對使用者的對話進行分析，並會以 BERT、RoBERTa 等預訓練模型加入資料來進一步訓練。第三階段主要是由另一位研究生夥伴(Su, 2022)負責，其利用本論文第二階段所得之成果，透過 GPT-2 預訓練模型來進行回覆的生成，會以心理諮商、治療的方式與使用者進行對答。如何針對現有的對話語料，進行對話的完整分類並產生與人類對話高度相似的對話內容將會是本研究的主要目標。

總之，本研究期望能提出一個醫療陪伴型聊天機器人之完整的架構與流程，並以數量足夠的中文資料來進行語言模型的訓練，我們會藉由 BERT、GPT-2 兩種預訓練模型去學習語料內容，利用模型判斷使用者對話內容後，進而生成流暢的回覆。不僅能促進自然語言處理技術在中文方面的發展，也能補足現今醫療無法滿足憂鬱症患者的需求，並有效解決其所導致的經濟社會隱憂。

1.3 論文架構

在接下來的論文架構，將於第二章進行文獻相關的探討與解釋，包含對話憂鬱程度分析、使用者情緒分析、主題分析、聊天機器人以及相關技術文獻，於第三章詳述我們的研究方法，第四章提出實驗細節與實驗研究成果及驗證，最後於第五章說明結論與未來研究方向。

第二章、文獻探討



本研究旨在針對憂鬱症患者之中文的對話分類判斷，並利用自然語言處理技術了解使用者對話，以此達成心理治療的目的。在本章節我們將會探討過往的研究並透過分析梳理出可行的研究方法及值得關注的議題。

於本章第一節，我們會先對憂鬱症的療法做簡單的介紹，說明本研究預計採用療法。第二節，本研究將探討聊天機器人的發展現況，在這一小節主要會對過往的各種不同聊天機器人的應用進行探索。在上一章中有提到，聊天機器人主要分為兩大類，Open Domain 與 Close Domain，而本研究會以 Close Domain 中，任務導向的聊天機器人做為主要研究目標，過去已有許多研究提出相關應用的研究成果，將會在此節中深入探索。

第三、四節將以語言模型 -BERT 為討論重點，在自然語言處理的範疇裡，不管是在分類或生成的任務上，表現良好的語言模型都會對結果有很大的影響，BERT 在多數自然語言處理的任務上都有非常優異的表現。希望能從過往與其相關的研究中有所發現，找到最適合本研究的預訓練模型，進一步探討對話判斷所需要的憂鬱程度分類、情緒分類、主題分類等相關研究，如此我們就能更加了解如何在這些方面達到更卓越的表現，以利後續的對話語句生成。

於第五節將進行對話生成方面的討論，對話生成部分主要為另一研究生來完成(Su, 2022)，他們以語言模型 GPT-2 為主要討論對象。在對話語句的生成中，很容易出現語句不流暢、沒有正確回應等問題，使得無法與使用者進行完整的溝通與對話。因此，對相關研究的討論，將可以幫助思考如何以得宜的方式解決此類問題。

最後，將在第六節簡單描述翻譯內容的評估方法，並在第七節對文本分類的幾種評估方法進行說明，了解何種方法對於本研究在對話分類方面的表現評估是最為合適的。



2.1 憂鬱症治療

憂鬱症治療需要由三種方式所組成，自我協助計畫、談話療法以及藥物。首先，醫生會建議患者平時透過自助計畫來治療，醫療保健人員會定時檢查患者是否有參照計畫方法進行治療。而談話療法的配合，可以讓患者更容易發現自己的情緒問題，並用更理性的方法思考。若是有中度至重度的憂鬱症患者則須配合藥物進行治療(NHS, 英格蘭國民保健署)。本研究會採用人際取向治療做為主要治療方式，而認知行為治療則是多數現有應用使用的方法，以下會說明此兩種治療方式，並了解我們將如何在對話中呈現與實體諮商同等的效果。首先為認知行為治療 (Cognitive Behavioural Therapy, CBT)，認知行為治療主要幫助患者了解自己的想法與行為，以及這些是如何影響到患者情緒的。CBT 觀點認為過去的成長經歷會影響到現在，但主要著重在患者如何改變現在的思考、感受和行為方式，並使患者能克服消極的想法。療程通常持續 5 至 20 次，在特定情況下，也會使用團體 CBT 這個治療模式。第二種為人際取向心理治療 (InterPersonal Therapy, IPT)，人際關係治療 (IPT) 側重於患者與他人間的關係以及在人際關係中可能遇到的問題，例如溝通困難或應對喪親之痛，目的在減輕患者的痛苦並改善他們的人際功能，還有幫助患者更好地利用社會支持 (social support)。IPT 療法對於大學生與青少年的憂鬱症的成果是令人肯定的，可以提供學生一個短暫的支持網路，較關注當下的人際關係，並可以在療程中提供患者人際互動的機會 (Ravitz, 2004) (Robertson, 2006)。

2.2 聊天機器人

最初，聊天機器人只能透過幾個固定腳本理解簡單的自然語言，因而產生類似人類的互動模式來回答問題。例如，由 MIT 在 1966 年所推出的 ELIZA，可以利用腳本與人類對話 (Weizenbaum, 1976)。目前，聊天機器人已經可以透過對輸入的文本資料進行語意分析，並能自行產生相應的回覆，這都要歸功於機器學習與人工智慧的長足進步。聊天機器人現在已經可以應用在多個領域，並提供有用的資訊。在一項研究中，研究設計了一個專門針對英語學習的聊天



機器人，聊天機器人提供許多學習建議並且具有幫助使用者學習新詞彙的能力 (Pham et al., 2018)。醫院則可以利用聊天機器人來協助監測患者狀況，自動化的方式可以更有效率的掌握疾病狀況 (Piau et al., 2019)。在健康管理方面，聊天機器人也可以使人們改變每天的習慣，讓大家可以多攝取健康飲食及多運動，結果也證實是有效率的 (Fadhil & Gabrielli, 2017)。而或是應用聊天機器人提供客戶服務，進而增加公司收入(Radziwill & Benton, 2017)。此小節會將聊天機器人應用分為兩大類來進行探索，分別為閒聊聊天機器人以及具有醫療目的之聊天機器人，而本研究將會著重在目的為醫療保健與心理健康的機器人來進行討論。

2.2.1 閒聊機器人(Chit-Chat)

每天我們都會與不同的人交談，並與他人建立聯繫，可以滿足我們的歸屬感與社交需求。因此就有了閒聊機器人的出現，他們在提供使用的過程中也不停學習。著名的代表之一就是微軟在 2014 年推出的機器人小冰，為針對中文使用者的聊天機器人，她可以完成帶有情感的對話，也可以提供許多生活輔助功能。而小冰在全球已擁有約 6.6 億使用者，其中約 1.2 億月活躍使用者，占據了全球對話式人工智慧總流量中的大部分。然而，也因為機器人學習對話的能力強大，時常遭到使用者的惡意使用，小冰就曾經因為「說錯話」，而被平台禁用。另一嚴重的例子則是微軟在 Twitter 上的聊天機器人 Tay，也是針對英語使用者的美國版「小冰」。於 2016 年發布，後因被惡意操作出現種族主義言論而下線。

本研究所建立之聊天機器人以陪伴使用者為目的，希望透過閒聊方式減輕使用者的壓力，並拉近關係。因此會以閒聊為基礎，另外會在回覆中以較溫暖的文字提供心理指導以回應使用者需求，並適當過濾不當言論。

2.2.2 醫療保健型聊天機器人

為了提供更完善的醫療服務，聊天機器人被廣泛的應用在醫療健康領域。Lokman 與 Zain 等人 (Lokman et al., 2009) 提出一個為糖尿病患者設計的機

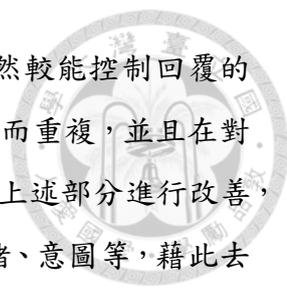


器人，可以用於控制糖尿病病提供醫師診斷建議。Belfin 等人 (Belfin et al., 2019) 設計一個癌症聊天機器人，用戶可以詢問關於癌症的任何事情，像是症狀、治療、存活率等。該機器人透過從各種癌症論壇收集的信息進行訓練，這些論壇具有大量且豐富的癌症資訊。Athota 等人 (Athota et al., 2020) 使用 AI 創建一個醫療聊天機器人，其能診斷疾病並在諮詢醫生之前提供有關疾病的基本細節。這將有助於降低醫療保健成本，並透過醫療聊天機器人提高對醫療知識的普及性。而 Ilić & Markovi (Dejan & MARKOVIÄ, 2016) 則對運用在醫療保健中的 AI 進行評估，他們發現在醫療保健中使用聊天機器人的優勢，包括降低成本、提高效率和減少做出正確診斷所消耗的時間。

2.2.3 心理健康治療聊天機器人

人工智慧與機器學習時常被用於開發心理健康照護的預測和治療方案。SimCoach 是一個由 Rizzo 等人 (Rizzo et al., 2016) 提出的虛擬智慧對話系統，目的在提高不願與傳統顧問交談的患者對症狀和治療的了解，並在最後鼓勵他們在需要時尋求其他更專業的資源協助。另外，Cameron 等 (Cameron et al., 2017) 也曾提出一個用於心理健康諮詢中的聊天機器人的設計，可以提供用戶初步資訊，並引導用戶進入諮詢服務並獲得資訊。在憂鬱症治療方面，Standford 大學的 Alison Darcy 推出了一個針對憂鬱症的治療型聊天機器人-Woebot，使用認知行為治療(Cognitive Behavioral Therapy, CBT)進行對話心理治療。當他發現你情緒低落、焦慮時，他會邀請你「寫出自己的想法」，然後透過 CBT 當中的一些技巧，來協助你調整這些想法。而根據 Darcy 等人 (Fitzpatrick et al., 2017) 的研究，他們的實驗證實使用 Woebot 確實會使患者精神症狀減輕，而此研究也間接證明了聊天機器人可以與心理治療方面結合，作為良好的治療輔助。

與上述提到的研究不同，本研究在對話的判斷分類以及生成上皆運用了語言模型進行實驗，將會使對話內容呈現較為流暢，也較能夠回應使用者的需求，而 SimCoach (Rizzo et al., 2016) 以及 Woebot (Fitzpatrick et al., 2017) 進行回覆時以規則導向為主，首先會判斷使用者是否有用到特定字詞，進而在生成



回覆時給予已在對話語料中定義好的回答，這樣的做法雖然較能控制回覆的質量，以免出現不合適的用語，但也使回覆的內容顯得單一而重複，並且在對話上，時常無法回應使用者的當下情緒。因此本研究將針對上述部分進行改善，我們所使用的方式會透過判斷使用者在對話中傳達出的情緒、意圖等，藉此去自動生成回覆，所以在回覆上較自由，也能以更貼近使用者的話語與其進行對話，期望能在回覆中讓使用者感受到同理心與陪伴，並提供針對各種情境的解決方式，而本論文也會針對不適當的用語生成進行控制，以達到通順且適當的對話。

2.3 基於 BERT 語言模型的對話判斷

近年來，深度學習的研究蓬勃發展，其在自然語言處理的應用上非常具有潛力，能夠學會語法結構、解讀語義。本論文認為將其應用在對話判斷上，可以將對話準確分類，達到最佳的成效。而在深度學習領域中，BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018) 是眾所矚目的焦點。Google 的團隊一釋出 BERT 的預訓練模型，其就在 11 項自然語言處理任務上都取得了最先進的結果。

BERT 是一個基於深度學習的語言模型，在訓練時會從一個單詞的上、下文學習，這種雙向性有助於模型更好地理解、使用句子。而這裡也體現了 BERT 的強大，與其它上下文無關模型不同，如 word2vec 或 GloVe 會為詞彙表中的每個單詞生成一個詞向量進行表示，因此容易出現單詞的歧義問題；BERT 會考慮到單詞出現時的上下文，因此會根據單詞在不同的句子中給予不同的詞向量。例如英文單詞 date 可以是日期，也可以是約會，word2vec 遇到兩種情況會生成一樣的詞向量，而 BERT 則會根據語意生成不同的詞向量，生成的向量可以更完整的表示每一單詞的情境。

以往為了解決不同的 NLP 任務，我們會為該任務設計一個最適合的神經網路架構並做訓練，而通常這會是非常耗費成本的，包含時間、計算資源等等，若是有處理各式 NLP 任務的通用架構，則可以節省重複訓練所造成的資



源浪費，而 BERT 就是以此為目標去建構的。為了使 BERT 有強大的泛化能力，其論文的作者們使用了大量文本來進行訓練，文本的訓練資料總共多達 34 億個字詞，因此由 Google 釋出的開源模型，最基本的 BERT 預訓練模型 BERT-BASE 也有 12 層的架構，並且有高達 1.1 億參數。由於 BERT 強大的泛化能力，我們只需要以這些預訓練模型為基礎，就能微調(fine-tuning)多個下游任務，包含填空、QA、文章摘要、文本分類、翻譯等等。

文本分類一直是 NLP 領域中的熱門研究議題，模型會根據文章、短句的語義、句子組成等，在訓練中學習資料類別的特性，並將文本分配到對應的類別。BERT 擁有卓越的上下文理解能力，根據 Sun, Qiu 等人 (Sun et al., 2019) 研究，他們使用了 8 個被廣泛使用的文本分類資料集對 BERT 模型的成效進行評估，包含 IMDb、Yelp 等，結果表示以 BERT 去 fine-tune 的模型在所有資料集的錯誤率都低於其他分類模型，可以得知其在文本分類任務上有出類拔萃的表現。

2.3.1 自殺預防

憂鬱程度會影響到憂鬱症患者是否會有自殺的想法，一般患者約 15% 有自殺的想法，且都是屬於偏重度憂鬱的患者，根據台灣憂鬱症防治協會統計，台灣 15 至 24 歲自殺人數，從 2018 年的 193 人，成長到 2020 年的 257 人，自殺通報人數也是逐年增加，顯示患者中出現自殺傾向的比例升高，已經形成嚴重的社會問題。許多研究中皆有提及這樣的問題，並試圖尋求解決方法，其中也包含了許多利用自然語言處理來進行分析的研究。C. M. Dharmapur 等人 (Dharmapuri et al., 2022) 於今年提出一聊天機器人作為醫療輔助，並以心理學家與患者的對話來進行機器人的訓練，使機器人可以盡可能地模仿心理學家的說詞進行對話，他們希望能達到自殺行為檢測與自殺預防的目的，而結果也指出 AI 對話系統在自殺預防是有效的。

除了上述研究，在相關的研究中，也會有許多研究是以遺書的內容來進行分析，已自殺者留下的遺書文字代表死者生前的絕望與痛苦，更能協助我們判斷與自殺想法相關的文字。Pestian, J. 等人 (Pestian et al., 2010) 即使用 33



份遺書與另外 33 份偽照的對照組文字進行研究，其研究請專家來進行判斷，並以多種 Decision Trees 與 Classification Rules 等機器學習演算法進行比較，結果表示，機器學習演算法是可以幫助區分遺書內容的。而另外於今年，Jain, P. 等人 (Jain et al., 2022) 使用了 Reddit 上的兩個版 r/depression 和 r/SuicideWatch 上面的文字進行分類任務，r/SuicideWatch 時常被有自殺想法的人使用，因此可以藉由兩個版上的資料差異來進行分類。其研究使用了如 Naïve Bayes、SVM 等 ML 演算法，分類結果也達到了約 77% 的準確率，證明了使用 NLP 方法是有效地進行兩種類別的區分，並可以達到自殺預防的目的。

本論文同樣也是使用 Reddit 上的兩個 subreddits r/depression 和 r/SuicideWatch 的文章來進行分類，但與(Jain et al., 2022) 的研究不同，我們主要是以 BERT 預訓練模型以及翻譯後的中文資料來進行實驗，本論文認為 BERT 強大的語意理解能力可以幫助我們在分類任務上獲得較佳的成果。另外，由於 Jain, P.等人未釋出資料集，因此本論文使用的資料為另外在版上進行蒐集的，無法與他們的結果進行公平的比較。

2.3.2 情感分類

情緒是我們生活中非常基本的一個面向，無論我們做任何事，都會反映我們的情緒。因此我們會從生活周遭獲取情緒數據來分析這些情緒，不管是文本、語音或是面部表情，都是很重要的分析來源。隨著數位科技的發達，關於情感分析的議題也不斷的在各個領域受到關注及討論。

Rajalakshmi 等人 (Rajalakshmi et al., 2017) 的論文將情感分類方法主要分為兩大類，一為基於情緒辭典將文本進行分類，另一種則是使用機器學習的方法來進行分類，如 Decision Tree、Neural Network 等等。而基於情緒辭典的分類有一些限制，會受到辭典中的詞彙量影響，而造成實驗結果出現偏差；機器學習的方法則可以建立多種模型，藉由訓練模型適應不同的分析目的，但需要耗費時間成本去對新資料進行情緒標記。



過去有許多研究皆將機器學習方法應用於情感分析領域。Pang, B.和 Lee, L.等人 (Pang et al., 2002), 他們透過單獨識別文本關鍵字, 更準確的判別情緒。於這項實驗中, 他們考慮了兩種機器學習方法, 分別為 Naive Bayes 以及 SVM, 結果顯示由機器學習產生的結果優於人工生成的模型。而在 Reyes, A. 和 P. Rosso (Reyes & Rosso, 2012) 的研究中, 他們在 Amazon 評論中搜尋具有諷刺意思的顧客評論, 並使用了 Naive Bayes、SVM 以及 Decision Tree 三種分類方法去做分類, 這也是情感分類議題中最常被進行比較的三種機器學習方法, 最終 Decision Tree 的效果優於其他兩者。

近年來, 除了上述傳統方法, 自然語言技術在文本分類領域也發展得很好, 尤其可以從文本分類競賽中發現, 語言模型的效果已遙遙領先其他分類演算法。除此之外, 在多數情感分析相關研究中, 語言模型對比其他方法都有較優異的表現, 其中也以 BERT 最為突出。英文情感資料集 SST-5 (Stanford Sentiment Treebank – 5 classes) 的最高準確率從 45.7% 增加到 59.1%, 分別是使用 RNTN model 以及 RoBERTaLarge+Self-Explaining 做為主要分類模型。Sousa 等人 (Sousa et al., 2019) 則利用 BERT 對金融新聞情緒分析的問題上進行評估, 以改進股票市場的預測結果, 研究結果表明 BERT 的性能優於卷積神經網絡 (CNN) 和詞向量 (Word Embedding) 方法, 大約上升了約 8.6% 準確率。由於 BERT 在情感分類上的優異表現, 我們認為使用 BERT 模型來進行對話情感判斷是最合適的, 而在 BERT 模型出現後, 也出現了許多以 BERT 為基礎的預訓練模型, 因此在本研究中, 我們會比較多個預訓練模型, 並找出最適合的模型, 來完成對話分析的任務。

在憂鬱症對話情緒方面, 我們採用 Ekman, P. (Ekman, 1971) 提出的六種基本臉部情緒, 而通過我們對資料的研究, 發現對話中有許多不帶有情緒的字詞, 且對話中不常出現驚訝這個情緒, 因此我們最終以 6 種情緒類別來進行情緒分析, 分別為中立情緒、開心正向、害怕擔憂、難過失望、噁心厭惡、生氣憤怒。

2.3.3 意圖分類

過去幾年，意圖分類領域的相關研究大部分皆以網頁搜尋引擎的改進與優化為主，分析使用者的意圖以呈現更好的搜尋結果給使用者。本研究的目的也與其相似，本研究主要將使用者的意圖進行分類以給予使用者更好且適合的回覆，針對對話中的意圖分類，Schuurmans, J.與 F. Frasincar (Schuurmans & Frasincar, 2020) 研究了幾種機器學習的方法來解決對話意圖問題，他們採用了多層分類器，根節點所建構的分類器的結果會考慮應該要選擇下一個要使用的分類器，直到葉節點，其結果會成為最終預測。研究中使用了 SVM 與 BiLSTM 兩種模型加上三種詞向量生成方式 Word2Vec、GloVe 和 FastText 進行比較，並同時比較多層以及平面模型的差別，SVM 在多層及平面架構中都有較佳的表現。另外，值得注意的是，其所使用的三種詞向量生成方式皆與 BERT 不同，而 Zhang, P. and L. Huang (Zhang & Huang, 2020) 則是使用 BERT 預訓練模型對中文新聞文本 - THUCNews 進行意圖分類與比較，實驗結果表示，與循環神經網路(RNN)與卷積神經網絡(CNN)相比，BERT 提供更好的意圖分類準確率與召回率，因此我們期望將其應用在憂鬱症使用者對話意圖的領域中，也能獲得與之相當的效能。

在憂鬱症對話意圖方面，本研究參考 Mporas, Lyras et al. (Mporas et al., 2007) 對於對話意圖的 10 種分類，由於本研究主要目的為治療陪伴功能，因此整理出較適合本實驗的三種意圖，分別為詢問意圖、回應意圖、治療建議意圖。

2.3.4 主題分類

相較於情感分類以及意圖分類的研究，文本主題分類的相關研究就少了許多。文本主題分類會將輸入的文本分配到一組有限類別中，其被廣泛運用在文本搜索任務中，當使用者為了要尋找特定主題的資訊，會利用搜尋引擎中的主題搜索將資料限縮在特定主題中，例如 Google 所提供的進階搜索功能。



在主題分類的領域中，多數研究也以自然語言處理技術與深度學習模型的使用作為主要方法，在一關於電動車的消費者體驗的主題分類研究中，作者於論文中比較 4 種模型應用於主題分類的成效。消費者評論被分為八大類主題進行研究，並使用 BERT、XLNet、LSTM 和 CNN 等模型來實驗並評估結果，最終以深度學習為主的語言模型 BERT、XLNet 的準確率以及 F1-score 明顯優於其他二者 (Ha et al., 2021)。本研究會以憂鬱症病患所遇到的問題、困擾來建立主題類別，並以 BERT、RoBERTa (Liu et al., 2019) 等預訓練模型進行微調訓練，以確切了解對話中所含資訊。

在憂鬱症對話主題方面，本論文根據 Addis, Truax et al. (Addis et al., 1995) 的研究，目的是了解造成憂鬱症的主要原因，在他們的研究中，他們先認列了 93 種會被認為可能引發憂鬱症的原因，後來經過他們的研究分析後，發現有八項的表現最為突出，而本研究也沿用其研究成果，使用該八項原因作為憂鬱症患者之主要對話主題，分別為人際相處、性格、成就、親密關係、存在議題、童年、生心理問題。

2.4 BERT 預訓練模型

在自然語言處理的文本分類任務上，BERT 毫無疑問會是第一個被提起的，BERT 對 NLP 領域的影響眾所周知，當時超越了許多榜單上的神經網路模型(NN model)，而後基於 BERT 的預訓練模型也蓬勃發展，各種以不同任務為訓練目標的模型越來越多。在我們比較過多種預訓練模型後，本論文決定使用 RoBERTa 作為模型的主要架構，並會採用以 Whole Word Masking 的訓練流程來訓練的預訓練模型，以期能達到更加準確的效果。

2.4.1 RoBERTa

由 Liu, Yinhan 等人 (Liu et al., 2019) 提出的RoBERTa，是 BERT 的一種改進版本，其研究更改了 BERT 的預訓練流程，用更大的 batch (256-8000) 和更多的數據並以更長的時間來訓練模型。



除了上述流程的改動，RoBERTa 也有許多訓練方法的改變。首先，其研究使用 Dynamic MASK，原版的 BERT 的處理方法為在數據前處理期間執行一次 MASK，得到一個 Static MASK；而 RoBERTa 則是在每次向模型輸入一個序列時都生成新的 MASK；另外，RoBERTa 也取消了原本 Bert 的 Next Sentence Prediction (NSP) 任務。原本的 Bert 為了捕捉句子之間的關係，使用了 NSP 任務進行預訓練，NSP 任務會判斷兩個句子(這裡假設為 A 跟 B) 是否為連續的，因此，在 NSP 任務的訓練的數據中，50%的 B 是 A 的下一個句子，50%的 B 是隨機抽取的。而 RoBERTa 將 NSP 移除，其研究將方法改成每次輸入連續的多個句子，直到最大長度 512，並可以達到跨文章文本，也就是以 FULL - SENTENCES 方式去訓練，而原來的 Bert 每次只輸入兩個句子，以上的改進方向都使得 RoBERTa 在多個領域超越了 BERT，有著更出類拔萃的表現。最後要說明 RoBERTa 在多個 NLP 任務的表現成果，在 GLUE (General Language Understanding Evaluation) 基準的 NLP 任務上，RoBERTa 在四個任務上都取得了 state-of-the-art 的成果，分別是自然語言推理 (Multi-Genre Natural Language Inference, MNLI)、Question NLI、語義文本相似性基準 (Semantic Textual Similarity Benchmark, STS-B)，以及文字蘊含識別任務 (Recognizing Textual Entailment, RTE)，從這些成果也可以發現，在語意理解方面 RoBERTa 取得了非常卓越的表現，因此本研究認為在對話理解方面，RoBERTa 也會帶來同樣優秀的表現。

2.4.2 Chinese BERT with Whole Word Masking

雖然 google 在 2019 年已經提出了使用全詞 MASK (Whole Word Masking, 簡稱 wwm) 進行訓練的模型，在 wwm 中，如果一個完整的詞的部分被 mask，則同樣屬於該詞語的其他部分也會被 mask。那在中文的範疇中，由於 google 原版的 BERT-base-Chinese 並沒有考慮到中文的文字斷詞，依然是進行單個字的 MASK。因此，在 2021 年由 Y. Cui 等人(Cui et al., 2021)提出了一種使用 wwm 進行預訓練的中文模型，以下會說明 Mask 方式的不同：

說明	範例
原始文本	使用語言模型來預測下一個詞的 probability
斷詞文本	使用 語言 模型 來 預測 下 一 個 詞 的 probability
原始 BERT 輸入	使用語言 [MASK] 型來預測下一個詞的 pro [MASK]## lity
全詞 MASK 輸入	使用 語 言 [MASK] [MASK] 來 [MASK] [MASK] 下 一 個 詞 的 [MASK] [MASK] [MASK]

表 1、Whole word masking in BERT 範例 (Cui et al., 2021)

也能經由其研究的分析得知，在中文閱讀理解及中文文本分類任務上，Chinese-BERT-wwm(ext) 的表現都優於 BERT、ERNIE 等模型。因此使用 wwm 來進行訓練的模型為我們的首選。

2.5 基於 GPT-2 的對話生成

文本生成已經成為自然語言處理 (NLP) 中重要但最具挑戰性的任務之一，在文本生成的領域中，GPT-2 的表現非常優秀。GPT-2 的全名為 Generative Pre-trained Transformer 2 (Radford et al., 2019)，模型本身是基於 Transformer 的 Decoder 所建構而成，其訓練集為數量約八百萬的網頁，由研究人員從網路上爬取得到，大小共有 40 GB 的文字資料。對於其他預訓練模型來說，GPT-2 有著更大的突破，其在訓練時未針對特定下游任務，但卻在閱讀理解、機器翻譯、問答等下游任務上都有很好的表現，表示資料量與模型足夠大時，無監督訓練技術也能訓練出泛化能力高的模型。

Wei, Z. 等人 (Wei et al., 2018) 在對話語句生成任務中提出了一種基於強化學習的任務導向對話系統，用來進行自動醫療診斷，Yang 等人 (Yang et al., 2020) 則提出了一個可以提供 COVID-19 相關資訊的醫學對話系統，其利用多個基於 Transformer、GPT 和 BERT-GPT 的對話生成模型，並在任務上對模型進行微調，實驗結果表明以上述模型生成對話的質量是很高的，並可以提供豐富的醫療相關資訊給予使用者。因此本研究會以 GPT-2 預訓練模型來進行對話語句的生成，期望能在中文對話生成的領域有嶄新的研究發現。



2.6 翻譯評估

BLEU(Bilingual Evaluation Understudy) (Papineni et al., 2002) 指標常用來進行機器翻譯的評估，用來比較候選文本與多個參考翻譯的評價分數，同時也與人類給的評估高度相關；除了翻譯之外，BLEU 指標也可以應用於其他的語言生成問題，例如：語言生成、文本摘要、語音識別等。因此，本研究主要使用 BLEU 來測試哪一種翻譯方法有較好的成效。

2.7 評估方法

此節會以幾個常見的文本分類評估指標來進行說明，本研究會以準確率 (Accuracy)、F1 值(F1-score)作為主要的評估方式，並輔以精確率 (Precision)、召回率(Recall)、混淆矩陣(Confusion Matrix) 這三類來評估實驗結果，以下會詳細說明幾種方式的主要用法與使用時機：

1. 混淆矩陣 (Confusion Matrix)：可以清楚看出模型在每一個類別的個別分類結果。以下詳細說明各名詞定義，並於下圖 2 展示：
 - True Positive (TP)：真實情況是「有」，模型說「有」的個數。
 - True Negative (TN)：真實情況是「沒有」，模型說「沒有」的個數。
 - False Positive (FP)：真實情況是「沒有」，模型說「有」的個數。
 - False Negative (FN)：真實情況是「有」，模型說「沒有」的個數。

		True Condition	
		Positive	Negative
Predicted outcome	Total Population (I)		
	Positive	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

圖 2、混淆矩陣(Confusion Matrix)



2. 準確率 (Accuracy, F1- micro)：文本分類問題中最基本的評價指標，正確分類資料在全部測試資料中所佔的比例。由上面混淆矩陣說明公式如下：

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

3. 精確率 (Precision)：「模型預測」為真 (TP+FP)，而「真實資料」為真的比率。由上面混淆矩陣說明公式如下：

$$\text{Precision} = \frac{TP}{TP + FP}$$

4. 召回率 (Recall)：「真實資料」為真(TP+FN)，而「模型預測」為真的比率。由上面混淆矩陣說明公式如下：

$$\text{Recall} = \frac{TP}{TP + FN}$$

5. F 值 (F1-score)：是 precision 和 recall 的調和平均數 (harmonic mean) 可看作是該二指標的綜合指標，F-score 能綜合參考兩者，因此能更加全面地評斷模型的表現。F- score 的值分為 micro, macro 和 weighted 三種不同的算法，並只有在 Precision 和 Recall 都大的時候才會大。公式如下所示：

$$F - \text{measure} = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}}$$

2.8 文獻探討總結



根據文獻探討的結果，本研究整理了以下結論：

1. 因為中文資料的缺失，本論文使用英文資料集進行自殺預防研究

從過去的文獻中，我們發現在中文的自殺預防相關研究中，由於存在中西文化差異，中文世界的文化並不習慣面對遺書這類的文件而或是有較為開放的平台能讓患病者進行正向的交流，大多數人不願意將情緒表達出來。因此若要進行中文自殺預防的文字研究，必須借助英文資料來進行實驗。

2. 憂鬱症對話的情緒、主題與意圖類別

根據過去文獻，我們理解到哪些對話判斷類別適合進行憂鬱症對話分析。

- 情緒方面，我們使用中立情緒、開心正向、害怕擔憂、難過失望、噁心厭惡、生氣憤怒 6 種情緒進行分類。
- 意圖方面，我們使用詢問意圖、回應意圖、治療建議意圖 3 種意圖進行分類。
- 主題方面，我們使用人際相處、性格、成就、親密關係、存在議題、童年、生心理問題 8 種主題進行分類。

3. 本論文將使用 BERT 預訓練模型進行對話分類，期望能帶來良好的成果

過去有許多研究雖是採用自然語言處理的方法進行研究，但較多以邏輯式回歸、分類器等基本機器學習演算法來進行研究。近幾年對於文字模型的研究表現上都有突破性的進展，因此本研究主要以 BERT 預訓練模型來進行實驗。文獻中提到，BERT 模型在情緒、意圖、主題分類上都有許多良好的表現，期望能在本研究的多種分類任務上達成不錯的成效；同時，也能實驗以中文文本為基礎的分類任務是否能達到較優的表現，進而使生成方面更加穩定。

4. 對話分類及文本生成在醫療型聊天機器人領域還有待探索

過去許多在醫療型聊天機器人的研究顯示，對話分類以及自動文本生成還未在聊天機器人的領域應用廣泛。本研究將會以完整的對話系統做為最終目標，以對話判斷分類以及生成為基礎，期望可以在此領域上有更多的發現，並能對憂鬱症醫療協助做出貢獻。

第三章、研究方法



如同前述，在科技越來越發達的情況下，文明病也越來越普遍，憂鬱症患者的數量逐年上升中。在這個狀況下，我們發現憂鬱症患者時常無法得到及時的協助或是不願意取得他人協助，因此我們提出一個憂鬱症聊天機器人的架構，以與患者流暢的對話為目的，研究什麼樣的對話分類與生成方式適合與患者進行溝通對話。而憂鬱症患者的逐年增加也導致自殺人數的提高，這也是近幾年越趨嚴重的社會問題，本研究也會針對對話資料進行憂鬱程度的分類，在對話中區分出程度差異，進而能達到自殺預防的目的。

綜合上述，本章內容將主要針對自殺預防以及憂鬱症對話兩大主題進行研究的說明，下面小節將會依序說明研究的主要架構與方法、各資料集簡單介紹以及進行實驗所採用的 BERT 預訓練模型選擇。

3.1 方法概述

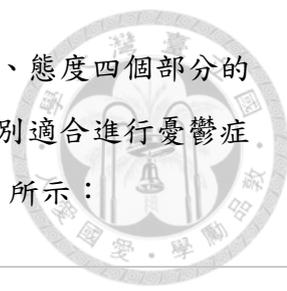
實驗主要分成三個階段來進行，如下表 2 所示。

階段	內容
第一階段	憂鬱程度分類 / 自殺預防
第二階段	對話標籤分類 (情緒、意圖、主題、態度)
第三階段	對話生成

表 2、實驗階段內容

第一階段我們會先將對話進行憂鬱程度的分類，主要分為一般日常對話、憂鬱傾向對話、自殺傾向對話三種。根據 Louise 等人的研究說明 (Louise et al., 2018)，至少有 90% 的人自殺身亡前是患有精神障礙疾病的，我們可以得知自殺傾向的患者有高機率為憂鬱症患者，因此本論文還會針對有自殺傾向的憂鬱症患者做更深入的研究。第二階段為使用者的對話判斷及分類，其中，我們參考 Microsoft 小冰 (Zhou et al., 2020) 的做法，小冰的架構中會判斷使用者對話中所提供的情緒 (Sentiment)、主題 (Topic)、意圖 (Dialogue Act)、態度 (Attitude)

四個方面。本論文沿用小冰的設定，使用情緒、主題、意圖、態度四個部分的分類作為目標，根據文獻內容，我們理解到哪些對話標籤類別適合進行憂鬱症對話分析，以下將各對話標籤類別以表格來整理，如下表 3 所示：



對話標籤	類別(Class)
對話情緒	中立情緒、開心正向、害怕擔憂、難過失望、噁心厭惡、生氣憤怒
對話主題	人際相處、性格、成就、親密關係、存在議題、童年、生心理問題
對話意圖	詢問、回應、治療建議
對話態度	正向、負向、中立

表 3、詳細分類標籤介紹

下圖 3 為本研究完整的流程與架構。本論文主要負責圖中第一、二階段的研究，我們會在第一階段判斷使用者的憂鬱程度，以此得知他們需要獲得幫助與否；若是使用者需要幫助，則會進入第二階段對話判斷的部分，藉由理解使用者對話中帶有的情緒與主題等，於第三階段生成適當的回覆，此部分則由研究室夥伴(Su, 2022)負責。

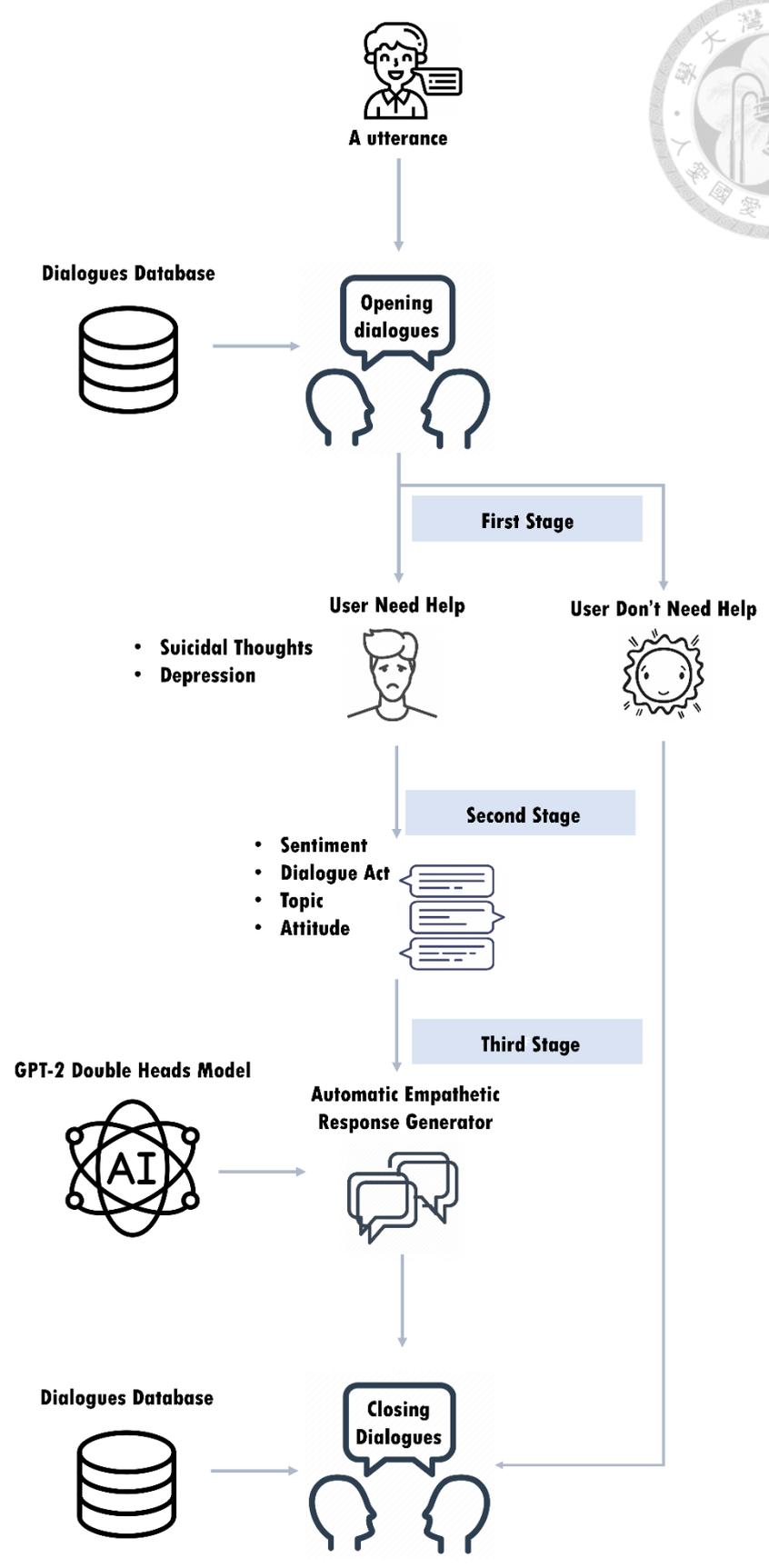


圖 3、實作方法與架構



3.2 資料集

本論文有兩個主要分類任務，分別為憂鬱程度分析以及對話類別分析，實驗也使用四個不同的資料集來完成分類任務，分別為 ChinesePsyQA、RSP Dataset、GSN Dataset 與 MTD Dataset，下表 4 為資料集內容簡介。

資料集	內容
<i>ChinesePsyQA</i>	中文心理問答資料集，主要進行憂鬱症對話分類與生成的資料集，取自壹心理平台
<i>Reddit Suicide Prevention (RSP)</i>	對話分為一般對話、憂鬱傾向對話、自殺傾向，目的為自殺行為預防，取自 Reddit
<i>Genuine Suicide Note (GSN)</i>	真實遺書資料集，取自 Kaggle
<i>Multi-Turn Dialogue (MTD)</i>	自殺傾向多輪對話資料集，取自 Reddit Suicide Watch subreddit

表 4、資料集說明

3.2.1 ChinesePsyQA

為中國壹心理平台中的問答資料，由於資料內容呈現較冗長與雜亂(原始資料請見附錄一)，我們會在資料前處理的階段中進行常用字替換以及文本長度的過濾。首先，本論文會先將非常用詞進行替換，例如：使用者時常在平台上將自己的男/女朋友稱作物件，因此我們會將此詞換成我們常用的對象一詞，以免造成失誤，其他部分會依照人工判斷將內容中部分大陸用語修正為較為口語、流暢的中文。第二，由於實驗的目的在進行對話的分類，所以本研究在篩選問句方面，會將過於冗長、細節的對話刪除，或是進行部分的修整，使對話的長度保留在一般對話的長度。最後，在問句回覆方面，原始資料集中，一個提問會有約 1-20 個回覆，由於文本數量過多、複雜且重複性高，同時為了減少每個類別的對話數量落差，本研究將回答篩選至 1-2 個，最終資料集所包含的問句與答句數量約為 1:2。

另外，使用此資料集最重要的在於文本標記的部分。該心理平台的使用者在上傳文章時，平台會提供類別供他們選取，但在觀察內容與類別的吻合度後，發現其類別跟文本無法進行配對，因此我們需要對文本的主題進行再分類。ChinesePsyQA 資料集將會由兩個人進行同一句子的標記，若出現標記的爭議，則會以第三人的來參與文本的標記，以達成文本標記之公正性。最後整理的結果，本研究總共標記的文本數量為 15,489 句問答，其中包含 5,607 個問句與 9,882 個根據問題的回覆。

本研究將使用此資料集來進行主題、情緒、意圖、態度四個部分的分類，在主題方面，本研究只使用使用者的發問來進行分類，一個問答題組統一會使用問題的主題作為整組對話的主題標籤；而在情緒、意圖、態度方面，則會使用標記資料集內所有的對話問答來進行。以下會列出本研究主題標記的準則與我們使用的標記資料集的範例，請參考表 5-7 的內容，而各標籤的類別分布請參考圖 4-7 的說明：

主題標籤	標籤介紹
人際相處	人與人之間的互動，包含爭執、友誼、社交相關等內容
性格	因為自身的性格、個性，從而在生活中遇到困擾、壓力等內容
成就	與個人成就、夢想、職涯選擇相關內容
親密關係	感情、婚姻經營等相關內容
存在議題	覺得生活沒有希望或是思考生命的意義相關內容
童年/家庭	小時候發生的事造成現在陰影、困擾，或是家庭問題對情緒造成影響等內容
生心理問題	大部分描述自己出現的心理或生理狀況，需要治療等相關內容

表 5、ChinesePsyQA 主題標記準則

文本	情緒	意圖	態度
總會想像人死後會去哪，害怕自己死掉，怎麼辦	難過失望	詢問	負向悲觀
解決你的個性問題，你要多出去走走，多參加些社交活動，你要懂得邁出第一步，先愛自己。	開心正向	治療建議	正向積極
我以後該怎麼對待欺騙我的人？	噁心厭惡	詢問	負向悲觀

表 6、ChinesePsyQA 回覆資料標記範例

內容

主題標籤

我不喜歡社交 覺得一個人可以生活得很好該怎麼辦？	人際相處
如何讓自己變得內心強大？	性格
畢業之後，不知道怎樣選擇自己的職業？	成就
喜歡上一個男生，但他有對象了，怎麼辦？	親密關係
失眠，抑鬱，沒有一件事順利，憑甚麼只有我這樣痛苦，偶爾會想著自殺	存在議題
我不理解為什麼很多人都說，不管自己的父母做錯什麼，他們始終都是你父母？	童年/家庭
睡眠不夠充足，失眠狀況嚴重，怎麼辦	生心理問題

表 7、ChinesePsyQA 主題標記資料範例

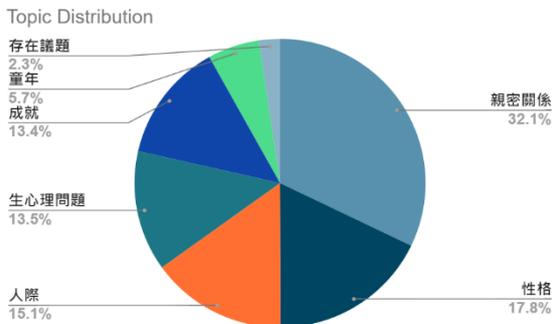


圖 4、ChinesePsyQA 主題分布情形

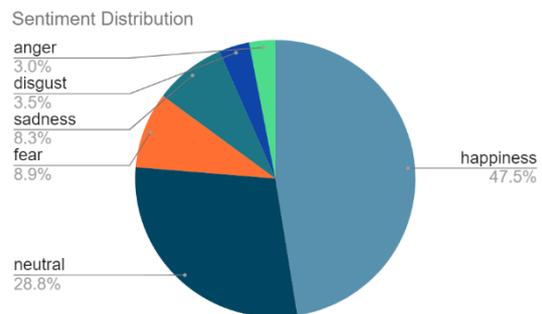


圖 5、ChinesePsyQA 情緒分布情形

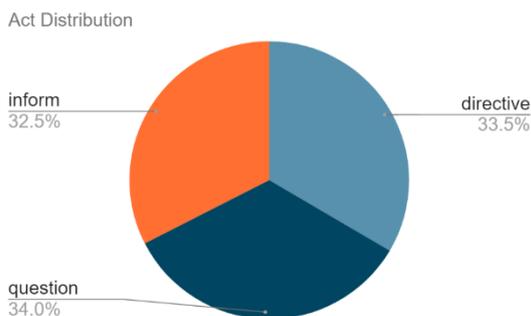


圖 6、ChinesePsyQA 意圖分布情形

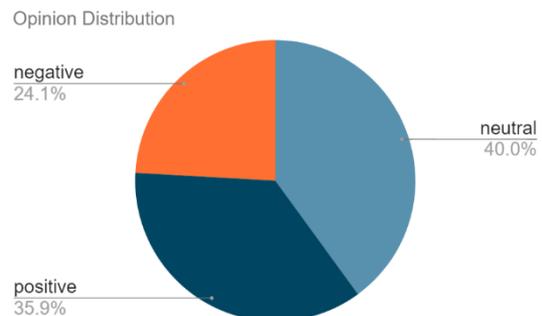


圖 7、ChinesePsyQA 態度分布情形

3.2.2 RSP (Reddit Suicide Prevention Dataset)

在本研究的第一階段，會將使用者的對話分為三種憂鬱程度，分別為一般日常、憂鬱傾向、自殺傾向三種，來進行憂鬱程度的分類。

RSP 由兩個部分組成，我們使用日常資料集 Daily Dialogue Dataset 作為一般日常對話，另外使用 Reddit 論壇上的 subreddit，r/depression 和 r/SuicideWatch 的部分發言來進行研究，尤其“SuicideWatch” subreddit 常被有自殺想法的人使用，可以為未發生的自殺行為提供重要信號。因此我們使用上述兩個 subreddit 的文章內容分別代表憂鬱以及自殺傾向的發言，並將其版上文章內容分別標記為憂鬱、自殺傾向。此資料集總共含有 3373 句對話，其對話類別分布如圖 8 所示：

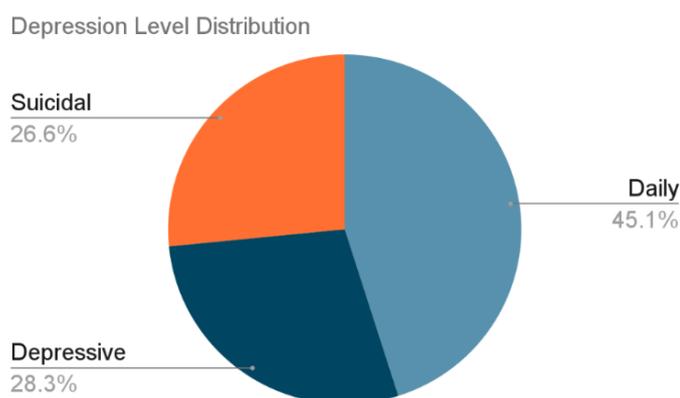


圖 8 、RSP Dataset 類別分布

而在實驗開始前，我們通過 3.3 節中的翻譯流程將英文資料翻譯成中文進行此分類任務。我們也將進行對話的過濾，將長度過短的對話剔除，例如，Oh!、Hi!、Ok!、Yeah 等單詞類的對話，最後，我們篩選出較為完整的日常對話組，取得部分作為本研究要使用的資料集。以下在表 8 展示 Daily Dialogue Dataset 原始資料集的一段對話，並在表 9、表 10 展示資料集英翻中的成果。

Daily Dialogue Dataset	
A:	Do you like cooking?
B:	Yes. I like cooking very much. I got this hobby when I was 12 years old.
A:	Why do you like it?
B:	I have no idea. I like cooking by myself. I like to taste delicious food.
A:	That's wonderful!
B:	And I love trying new recipes, which I usually test with my friends. You can come, too.

表 8 、Daily Dialogue Dataset 原始資料集對話翻譯

英文對話	中文內容
<i>A: What is the correct time, please?</i>	請問正確的時間是幾點？
<i>B: It's exactly twelve minutes past seven.</i>	時間正好是七點十二分
<i>A: When will the lecture begin?</i>	講座什麼時候開始？
<i>B: It'll begin at nine o'clock sharp.</i>	九點整開始

表 9 、Daily Dialogue Dataset 中英翻譯範例

英文內容	中文內容
<i>I want to die but most of all I think I want to feel pain.</i>	我想死，但最重要的是我想感到痛苦。
<i>Another period of apathy and emptiness setting in.</i>	又一個冷漠和空虛的時期開始了。這
<i>Is that better than feeling miserable all day?</i>	總比整天感到痛苦好嗎？

表 10 、Reddit 資料中英翻譯範例

本研究會將兩個翻譯後的資料集 Daily Dialogue Dataset、Reddit 爬蟲資料進行合併，是為 RSP Dataset，並以數字來做為憂鬱嚴重程度的標記，我們將一般日常對話標記為 0，憂鬱傾向對話標記為 1，自殺傾向對話標記為 2，以此來進行分類任務。表 11 列出 RSP Dataset 資料範例，分別有三個類別的文本資料：

Content	level
我五點半會到你家 生日快樂	0 (一般日常)
我厭倦了關心別人，但沒有人關心我。	1 (憂鬱傾向)
我覺得我的朋友並不真正在乎我，因為他們很難接觸。如果我屈服於抑鬱症，他們甚至會在乎我已經死了嗎？	2 (自殺傾向)

表 11、RSP Dataset 範例

3.2.3 GSN (Genuine Suicide Note)/ MTD (Multi-Turn Dialogue)

本研究第一階段，將另外使用此兩個資料集進行測試，如同 RSP 的處理方式，實驗前會先將資料集進行翻譯、調整後再進行標記，兩個資料集內的所有資料皆只含有自殺念頭的文本、對話，因此我們會將資料皆標記為 2 (自殺傾向)。GSN 內含有 418 筆資料，資料內容皆出自 Kaggle 的真實遺書資

料集；MTD 則包含 52 組三句以上對話，總共 180 句對話。以下在表 12 列出 MTD 資料範例，並將各組句數資料分布呈現於下圖 9：

Content	Level label
為什麼我要為從未有過的好生活如此努力奮鬥	2 (自殺傾向)
我一生中做了很多嘗試。我厭倦了為活著而戰，我厭倦了一切針對我的計畫。我不明白為什麼我必須如此努力地戰鬥才能來到這裡。	2 (自殺傾向)
我不想活著。關心我的人總是告訴我，我值得更好，我值得快樂。但我不明白一個應該活著的人，身邊除了災難什麼都不會發生。這一切只會更糟，活著只會更糟。	2 (自殺傾向)

表 12、MTD Dataset 範例

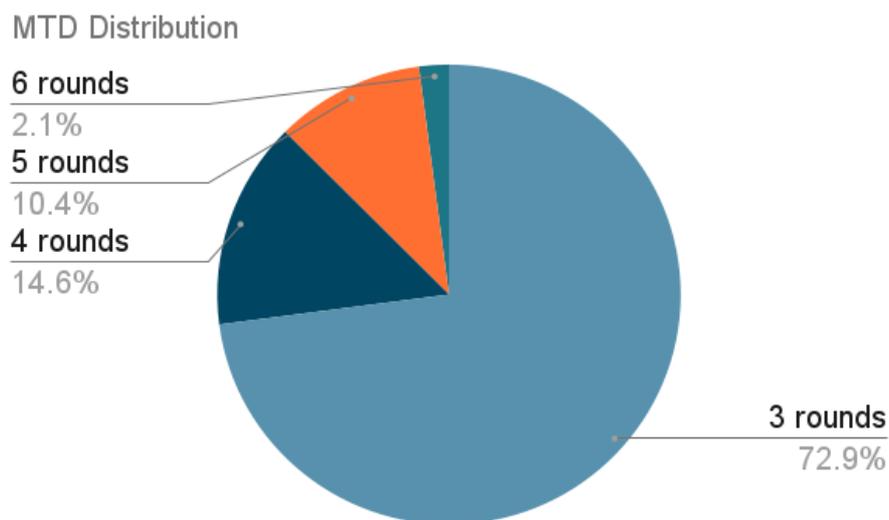


圖 9、MTD Dataset 各組句數分布

3.3 翻譯質量評估

在進行憂鬱程度分類時，由於中文實際資料數量的缺失，本研究在模型訓練時，會大量使用翻譯資料。為了使資料能快速翻譯並能進行模型的訓練，我們將會採用 Google Translate Web、Google Translate API 來進行大量英文文本的翻譯。若翻譯後資料能更精確的吻合中文文本的流暢度、易讀性，則會使模型訓練上有更好的效果，本研究採用機器翻譯領域的評估指標 BLEU 與人工評

估來測試哪一種方法有較好的翻譯成效。翻譯評估過程中，我們會先將語句進行人工翻譯，並以此作為標準，也就是 BLEU 評分中的參考語句，分數會介於 0.0-1.0 間，分數越高則代表翻譯的語句越接近中文口語。進行評估後，Google Translate Web 的平均得分為最高，語句最符合中文的邏輯，也較為通順，得分平均大約 0.53；而 Google Translate API 的平均得分為次高，稍微有不通順的地方，得分平均大約 0.27。

然而 Google Translate Web 無法一次進行大量資料的翻譯，因此我們使用 Google Translate API 作為主要翻譯工具，進行大量英文文本的翻譯，我們會在翻譯後進行翻譯完整性的評估，加上 Web 服務作為輔助，將翻譯不完整的句子加以調整。下圖 10 展示了完整的資料集翻譯流程：



圖 10 、資料集翻譯流程

經過上述翻譯流程的調整後，我們的翻譯結果平均的 bleu 評分可以達到 0.64，由此可以說明我們的翻譯結果還不錯，並可以讓後續實驗更好的進行。

3.4 模型挑選

在模型挑選與微調階段(fine-tuning)，根據前述 RoBERTa 模型優秀的表現，本論文在實驗中挑選的模型是以 RoBERTa 為主要架構的預訓練模型，並會優先採用以 Whole Word Masking 的訓練流程來訓練的預訓練模型。因此，本論文最終在憂鬱程度與對話分類的任務上會以 Chinese-Roberta-wwm-ext 、Chinese-Roberta-wwm-ext-large 預訓練模型作為主要模型進行模型微調(fine-tuning)，並以 BERT-Base、Chinese-BERT-wwm、Chinese-BERT-wwm-ext 為輔來進行結果的比較與分析。

第四章、實驗細節與實驗結果



本章節我們將針對第三章提出的方法進行實驗設定，並完整呈現本研究的實驗結果。首先，我們會說明所有參與實驗的模型以及參數設定的內容，接著會針對每項實驗的結果進行評估與討論，主要評估對話的分類與生成是否達到較佳的成果。

4.1 實驗設定

本研究總共會利用 7 種模型進行比較，分別為基本的 NN 模型 CNN 與 LSTM，以及 5 種 BERT 預訓練模型[7,18]：

1. **BERT Base**：基本的中文 BERT 模型。
2. **BERT-wwm**：利用 Whole Word Masking 進行訓練的中文 BERT 預訓練模型。
3. **BERT-wwm-ext**：屬於 BERT-wwm 的升級版，相比於 BERT-wwm 的進步是增加了訓練數據集，同時也增加了訓練步數。
4. **RoBERTa-wwm-ext**：利用 Whole Word Masking 進行訓練的中文 RoBERTa 預訓練模型，與 BERT-wwm-ext 相比，增加了 RoBERTa 的優點，動態 MASK 的替換與訓練中 Next Sentence Prediction (NSP) 任務的取消，這些也增進了此模型在自然語言任務上的成效。
5. **RoBERTa-wwm-ext-large**：中文 RoBERTa-wwm-ext-large 模型與先前的版本 RoBERTa-wwm-ext 訓練模式基本相同，唯一的差別在於參數數量的提升。

本研究所有模型皆採用 BERT Base Chinese 作為 tokenizer，並使用 Adam 作為 Optimizer。在 learning rate 的部分，我們在 $1e-6 \sim 1e-7$ 的區間內進行調整，以最適當的 learning rate 進行模型的 fine tune。另外，在模型進行 fine tuning 過後，主題、情緒、意圖、態度，我們都使用單層的 layer 作為主要架構；而進行憂鬱程度分類的 Classifier 則是採用 2 層 layer 進行



模型的訓練。

接著說明訓練集與測試集的切分。在實驗的第一階段，本研究使用 RSP、GSN 與 MTD 三個資料集，主要利用 RSP Dataset 進行訓練，並會以 RSP 的 15% 的資料、GSN 以及 MTD 來進行驗證與測試；而在第二階段，本研究會使用自行標記的 ChinesePsyQA，也以 15% 的資料來進行驗證與測試，兩個階段都會依照類別資料的數量進行平衡的資料集切割，若是訓練與測試的類別分布相差過多，會影響到模型訓練成效。

另外，本研究會選擇適合實驗的評估指標。由於我們的任務著重在對話的分類與分析上，因此我們會使用準確率 (Accuracy / F1-micro) 以及 F1-weighted 作為主要評斷的標準，除了整體準確率的評斷，本研究也會進行各類別的準確率評估。

4.2 對話標籤相關性人工評估

本研究資料集大多仰賴人工標記，因此會針對所有標籤進行人工評估，以確保人工標記的可信度。憂鬱程度的標籤是由原本 subreddit 的分類後再進行整理，因此不是完全由我們進行標籤，在標籤評估中，我們從網路上徵求到 25 位受測者來進行評估，每位受測者皆有大專以上的學歷，我們從 RSP 資料集選取包含憂鬱傾向、自殺傾向對話各半，共 50 句對話進行評估，一分為最不同意此標籤與對話相關，五分為非常同意此標籤與對話相關，本論文以相關性來表示此分數，最後得到的結果如下表 13 所示：

對話(數量)	相關性
憂鬱傾向對話(25)	4.021
自殺傾向對話(25)	3.752
所有評估對話(50)	3.887

表 13、憂鬱程度標籤相關性評估結果

根據上表得知，在憂鬱程度標籤的相關性評估中，大家較不認同自殺傾向對話中標籤與語句的相關性，我們推測原因為憂鬱情緒的表現較像是比較深的難過、失望情緒，但自殺傾向必須要讓受試者感受到更濃的絕望、痛苦，是一種很深刻的情緒，因此受測者較容易認同對話帶有憂鬱傾向。

而從評估分數來看，每一句憂鬱傾向對話的相關性評分皆超過 3 分，受試者認為憂鬱傾向標籤的相關性很高，但自殺傾向對話中，有 3 句話不被受試者所認可，評分低於 3 分，我們將相關性評分較低的結果呈現在下表 14。

我們發現憂鬱傾向對話中得分較低的，大部分是因為這些對話表現出說話者慢慢變好的狀態；而自殺傾向中得分低的對話都不會明確地提到自殺、死亡等詞彙，只有不停地抱怨生活，並說自己的存在沒有意義，與其他自殺傾向對話相比(見下表 14 編號 1 的內容)，他們並沒有在話語中表現出強烈的意圖，這樣就會造成無法人為判斷出他們的自殺意圖，也會造成模型判斷上的失誤。但整體來看，對話平均得分也有達到接近 3.9 分，問卷結果顯示受測者是認同此標籤對於語句的相關性的。

	內容	憂鬱程度	相關性
1	即使沒人在乎我，他媽的還有什麼意義？不管我在這裡與否，他們的生活都沒有區別。我想要痛苦，我只想死。	自殺傾向	4.3
2	我 14 歲那年開始傷害自己，現在 15 歲，我不知道該怎麼辦。我只有和同一堂課的女友在一起時，我才會很高興。	憂鬱傾向	3.1
3	我會沒事的，我可能不會，但是今晚我只能告訴自己。至少直到日出，我會沒事的。	憂鬱傾向	3.1
4	我太沮喪了，寫不出長長的遺書。我錯過了我的青少年時期，我從未被邀請參加舞會，也沒有參加過任何派對。(今年 8 月我將進入大四)我快成年了，還沒有取得任何成就。我開車也很糟糕，我一輩子都不會停車。最近失去了我的工作，所以就是這樣。生活不能比這更糟了，我的生活就是一團糟。	自殺傾向	2.7
5	我應該為自己所做的一切而受苦。我希望我繼續感到痛苦，以彌補自己造成的痛苦。	自殺傾向	2.7
6	我沒有毫無用處，我沒用。	自殺傾向	2.3

表 14、憂鬱程度標籤問卷資料範例

在對話標籤評估中，我們則是從網路上徵求 5 位受測者來進行內容標籤的評估，每位受測者皆有研究所以上的學歷。我們從 ChinesePsyQA 資料集中隨機選取 100 句對話進行評估，每句對話皆針對四個標籤進行評分，分別為情緒(Sentiment)、主題(Topic)、意圖(Dialogue Act)、態度(Attitude)，一分為最不同意此標籤與對話相關，五分為非常同意此標籤與對話相關，本論文以相關性來表示此分數，最後得到的結果如下表 15 所示：

對話標籤	相關性
對話主題	4.478
對話情緒	3.837
對話意圖	4.595
對話態度	4.1

表 15、對話標籤相關性評估結果

在對話標籤的相關性評估中，可以明顯看出在主題、情緒、態度三個方面我們都取得了 4 分以上的高分，代表受測者非常認可這些標籤的可信度。而在情緒方面，受測者較不認同此標籤與對話語句的相關性，大部分不被認可的語句多屬於開心正向 (Happiness) 與中立情緒 (Neutral)，詳細問卷資料說明如下表 16、17 所示：

文本	情緒	相關性
等有一天你足夠好，人生觀價值觀及行動都相對成熟的時候，這個問題就不是問題。	開心正向	3.0
如果時間允許可以嘗試一下去旅遊，多散散心，先嘗試一下讓自己的心境平靜下來。	開心正向	2.6
不知道未來的路在哪？	中立情緒	2.8
你這樣是不會有朋友的。	中立情緒	2.8

表 16、對話標籤相關性不佳範例



文本	情緒	相關性
失眠，抑鬱，沒有一件事順利，憑甚麼只有我這樣痛苦，偶爾會想著自殺	難過失望	4.4
總覺得會有看不順眼的事又不好跟當事人說，覺得說了會被討厭，怎麼辦？	害怕擔憂	4.7
我總是認為活著沒意思，要面對許多壓力，覺得死了就好，很鄙視自己。	噁心厭惡	4.6
老公外遇促使我們離婚了，他還多次要求和好，我現在該怎麼辦？	生氣憤怒	4.2

表 17、對話標籤相關性問卷負面情緒資料範例

上表 16 列出了一些情緒不被認可的對話，我們推測可能的原因在於憂鬱症對話資料集的正向情緒與一般帶有開心快樂情緒的對話語句中所帶有的情緒程度不相同，因此造成語句所透露出的情緒與大家所認知的不相同，受測者多會認為正向的句子較偏向中立情緒；而情緒中立的句子，卻多會讓受試者認為是偏向負面的情緒。但從所有對話平均結果來看，情緒的相關性平均依然有 3.8 的高分，這也說明，在資料集其他帶有負向情緒的語句中，如害怕擔憂(Fear)、難過失望(Sadness)等情緒標籤，還是被受試者所認同的，如上表 17 所示。

4.3 憂鬱程度分類結果

本研究於第一階段進行憂鬱程度的分類。在此階段中，會有四項分類任務要進行，並會使用 4.2.1 小節中說明的 5 種模型來進行分類及比較。在這個階段，本研究將進行四項任務；第一項任務，實驗將對話分為一般日常/憂鬱傾向/自殺傾向三種標籤，主要目的為進行日常對話與憂鬱對話的區別，並且在下表中我們稱此任務為 Task 1；第二項任務，僅使用去掉日常對話的 RSP Dataset 進行實驗，實驗將對話分為憂鬱傾向/自殺傾向兩種標籤，主要目的為在憂鬱對話中，分辨出有自殺想法的對話，進而能達到預防自殺的效果，在下表中我們稱此任務為 Task 2；而第三項與第四項任務為前兩項任務的延伸，我們會使用 GSN Dataset 及 MTD Dataset 來驗證 RSP Dataset，而 MTD 的測試方式比較不相同，是以整組多輪對話為單位進行分類，實驗

會經由整組多輪對話測試後的結果進行評估，若整組對話超過半數的句子分類正確，就將發起此對話之使用者認為有自殺傾向的使用者，因此MTD的結果表現會是以有幾組(sets)對話在 52 組中被認列來做為標準。又因為這兩個資料集皆只有自殺傾向對話這個類別，計算時此兩項任務的自殺傾向的類別 Recall 皆為 1，算出來的 F1 數值會偏高，因此會以 Accuracy 作為主要評估標準。

Task4 因為評估方式不同，因此不計算 F1，僅列出有幾組(sets)對話判斷準確。本論文於表 18 簡列出各項任務內容並將實驗結果呈現於表 19，下表中的評估指標以 Acc.代表準確率(Accuracy, F1-micro)，以 F1 代表 F1-score 的 Weighted Average (F1-weighted)：

Task 1	使用 Daily Dialogue + RSP 進行 3 個類別分類，切出部份資料做為測試集
Task 2	使用 RSP 進行 2 個類別分類，切出部份資料做為測試集
Task 3	使用 RSP 進行訓練，並用 GSN 進行測試
Task 4	使用 RSP 進行訓練，並用 MTD 進行測試

表 18、憂鬱程度各項任務說明

Model	Task 1		Task 2		Task 3		Task 4	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	sets
BERT Base	86.6	86.6	73.4	73.1	67.2	80.4	55.8	(29)
BERT wwm	86.2	86.1	83.5	83.4	71.1	83.1	75	(39)
BERT wwm ext	88.5	88.6	82.7	82.7	73.2	84.5	84.6	(44)
RoBERTa wwm ext	88.9	88.9	85.6	85.6	74.6	85.5	84.6	(44)
RoBERTa wwm ext large	89.7	89.7	87.1	87.1	76.6	86.6	86.5	(45)

表 19、憂鬱程度分類結果

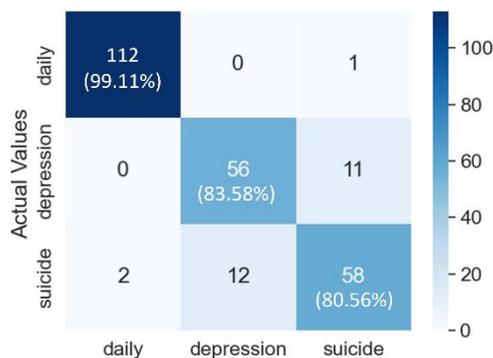


圖 11、Task1 Confusion Matrix (per-class recall)

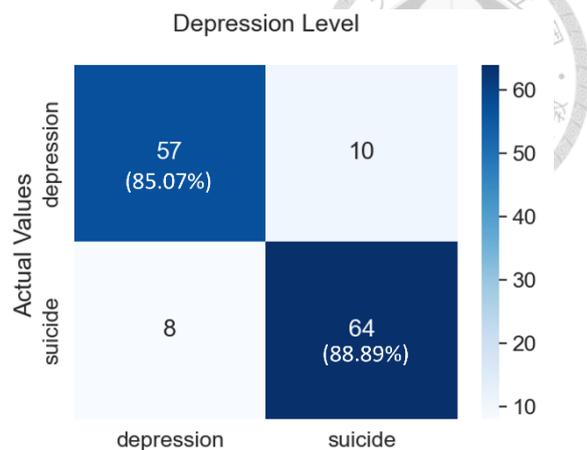


圖 12、Task2 Confusion Matrix (per-class recall)

從表 19 的結果可以發現，Task1 與 Task2 的分類結果上都是 RoBERTa wwm ext large 模型獨占鰲頭，分別達到 89.7%與 87.1%的準確率，與過去相似的分類任務結果相比，有很好的表現。接著從圖 11、12 的混淆矩陣來看，可以看出在一般日常對話上這個類別上，Recall 非常高，也代表日常性對話與憂鬱症患者的對話其實還是存在著很大的差異。而在憂鬱傾向與自殺傾向的分類上，Task1 在類別的 Recall 上皆低於 Task2，這也說明 Task2 能找出較多存在自殺想法語句，而兩項任務在自殺傾向的類別中也都達到 8 成以上的 Recall。

另外，說明 Task3 與 Task4，兩項任務分別達到 76.6%與 86.5%的準確率，雖與前兩項任務相比有一些差距，但 Task4 的 MTD Dataset 表現超過 85%，對未來發展多輪對話很有幫助。

4.4 對話分類結果

本研究的第二階段為對話主題、情緒、意圖、態度四個不同分類任務的結果，我們使用 ChinesePsyQA 進行此階段的任務。在此階段中，我們會使用 4.1 小節中說明的所有 7 種模型來進行分類及比較，並將實驗結果呈現於表 20。而下表中的評估指標以 Acc.代表準確率(Accuracy, F1-micro)，以 F1 代表 F1-score 的 Weighted Average (F1-weighted)：

Model	Topic		Sentiment		Act		Attitude	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CNN	67.5	-	65.4	-	74	-	69.9	-
LSTM	62.9	-	67.2	-	79.2	-	68.6	-
BERT Base	81.7	80.8	73	72.7	86.3	86.5	81.0	81.1
BERT wwm	81.9	81.1	72.6	72.7	86.5	86.7	81.1	81.1
BERT wwm ext	87.2	87.0	74.7	74.4	87.2	87.3	81.3	81.4
RoBERTa wwm ext	86.9	87.2	74.4	74.3	87.3	87.4	81.9	82.0
RoBERTa wwm ext large	89.8	90.3	76.9	76.8	87.6	87.7	82.8	82.8

表 20、對話分類結果

本研究於實驗結果有以下發現。首先，在本研究的 CNN 與 LSTM 的模型評估中，由於主要以 BERT 預訓練模型為主，這兩個模型的效果明顯較其它結果差。從結果中，可以發現 RoBERTa wwm ext large 模型的結果從所有結果中脫穎而出，於每個標籤的分類中都有最好的表現，但也可以發現在類別數比較少的 Act 與 Attitude 分類中，結果都非常相近；這同時也顯示出若需要分類的類別數越多，參數量較大的模型 RoBERTa wwm ext large，相比其他模型就會有明顯的優勢。

在情緒的部分，可以看出其表現沒有其他標籤好，只有 76.9% 的準確率，同時也從 4.2 節的人工評估結果來看，情緒的標籤在受試者的評分中，也取得了最低的平均分。我們推測原因在於憂鬱症對話中的正向情緒表現較不明顯，這也使開心正向與中立情緒的情緒界線過於模糊，因而會對人工評估以及實驗結果產生誤差。由於上述原因，本研究於生成階段不將情緒標籤輸入模型進行回覆生成。

除了情緒以外，其餘三個標籤的分類準確率都達到 8 成以上，甚至在 Topic 與 Act 的分類上都有接近 9 成的準確率，在相似的分類任務中，已經是非常突出的表現。在分類模型都有不錯的表現下，在接下來的生成部分就會使生成內容更加完整、準確。



4.5 對話系統串接結果

本研究的第三階段我們嘗試將本論文的分類階段與對話生成階段進行串接，由此了解兩個部分串接起來的結果。本論文會先將問題進行主題、意圖與態度的分類，再交由(Su, 2022)進行回覆生成。

在回覆生成的方面，由於實驗目的是跟患者進行對話，因此本研究皆是使用正向態度、指導意圖兩個類別進行回覆，讓回覆的對話中不會出現過於負面的字眼，影響到使用者的情緒。本論文於表 21 呈現以分類結果為成就主題的對話問題與答覆，其他主題的成果展示於附錄二：

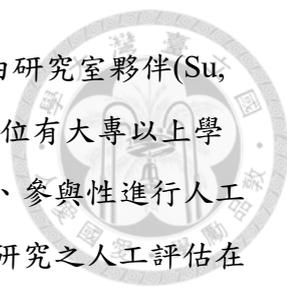
Question	Topic	Act	Attitude
對現在的工作不滿意，我該換嗎？	Achievement	Question	Neutral
Attitude	Act	Response	
Positive	Directive	你可以試著找到自己的興趣愛好，去做一些職業生涯規劃。	

表 21、對話系統串接範例(Achievement)

對話內容	對話標籤	相關性
對現在的工作不滿意，我該換嗎？	對話主題	4.4
	對話意圖	5
	對話態度	4.2

表 22、對話問題人工評估結果

首先，從問題本身來看，在分類上，成就主題代表個人成就與職場問題等，可以看到問題的確屬於成就主題的範疇中，且在本論文的人工評估中，這個問題的所有標籤都獲得了4分以上的高分，如表 22 說明。可以證明問句的分類結果是具有可信度的，而正確的分類結果可以使回覆內容更加準確。



接著，本研究會在回覆上進行人工評估，此部分由研究室夥伴(Su, 2022)負責，其論文從資料集隨機抽 100 個問題，徵求五位有大專以上學歷的受試者，並利用 3 種指標，包含相關性、情緒理解、參與性進行人工評估，一分為最不符合指標，五分為最符合指標。他的研究之人工評估在三項指標都取得了 4 分以上的好表現，由此可以確定在受試者的理解中，大部分回覆都是屬於正向態度以及指導意圖的。

最後，從上述結果中也可以發現主題、態度、意圖都是實驗中不可或缺의 標籤，也說明本論文將語句進行分類、標籤後，是有效幫助模型進行語句生成的。

第五章、結論



5.1 研究成果

本研究希望能以聊天機器人架構為基礎，進行對話分析與對話生成兩個部分，本論文以對話分析為主進行研究。在研究第一階段，我們使用英文資料集進行資料集的翻譯、合併與整理，再針對文字內容進行憂鬱程度的分類；第二階段，我們進行資料集的標記與整理，並進行主題、情緒、意圖、態度四個標籤的分類。從研究結果中發現，使用 BERT 預訓練模型，RoBERTa 模型在大部分的時候都會表現比 BERT 更加優秀，而使用 whole word masking 的模型，更加符合中文斷詞的邏輯，也表現得很好。透過表現優秀的預訓練模型，在第一階段，模型能判斷出帶有自殺想法的文字與帶有憂鬱傾向文字的差別，若配合完善的對話系統，則能達成自殺預防的目的；第二階段，模型也在各個標籤上都能有很好的準確率，在對話系統中，能成功地將對話歸類，尤其是在主題上有接近九成的準確率，機器對語句的了解，使得在對話語句生成方面能更加完善。未來在建構一個完整的對話系統上，對話分類越準確，可以使對話表現更加流暢及完整，對應用發展上有很大的幫助。

5.2 研究貢獻

本研究的主要研究貢獻包含以下幾點：

1. 對話中的自殺預防：

本研究試圖從文字中尋找隱含在對話中自殺意圖，建構了 MTD 資料集，以多輪對話的判斷為目的，能在多句來自同一位使用者的對話中判斷是否有自殺意圖。在此任務上，透過預訓練模型優秀的表現，本研究取得了不錯的成果，在未來的應用中，若是能有更多相關的中文資料，相信更能完善自殺預防的流程。



2. ChinesePsyQA 的標記與整理：

過往研究中，心理健康研究中的中文資料甚少，本研究與(Su, 2022)建構了 ChinesePsyQA 資料集，裡面含有總計約一萬五千句的對話，並標記了主題、情緒、意圖、態度四個標籤，以此來進行憂鬱症對話的研究。本研究採用 ChinesePsyQA 在對話分析與生成任務中皆獲得了不錯的表現，在分類上有接近九成的準確率，生成的語句不僅可以滿足憂鬱症患者基本聊天需求，也能進行完美的對答。

3. 對話標籤分類準確：

本論文針對對話的四個標籤進行了分類，我們使用的模型 RoBERTa wwm ext large 在主題、意圖與態度標籤部分都獲得了良好的效果，跟其他相似的任務相比，我們獲得了相當好的準確率，也同時在實驗中證實了情緒標籤對於此系統來說較不適用。除此之外，我們從最終結果得知，分類的成效好壞會影響到後續生成的品質，因此可以確認這三個標籤對於憂鬱症對話系統來說都是不可或缺的，而分類的品質佔了很重要的一部分。

5.3 研究限制

本研究主要有以下限制：

1. 中文資料缺乏

本研究在進行對話語句的分類與生成皆需要使用大量中文資料集，然而目前相關研究中所提出的中文資料集的質量不均且數量稀少，多數都是未限定領域、未經標籤的閒聊資料集。而在中文資料集中，與心理健康相關及遺書相關的公開資料相對來說更稀少，因此在研究過程中，本研究花了很多的時間進行資料的蒐集、標記、翻譯以及整理。本研究在進行第一階段時，也因為中文資料的缺乏，使用英文資料進行翻譯，可能會存在些許語意流失以及語句不通順的問題，這些問題可能會導致研究與實際應用情境不符，這也是本研究在未來待改善的部分。

2. 對話資料少

接續上一點，由於資料集多由本研究自行建構，無法完成大量資料的整理，目前只有約 5600 組對話。雖然在研究中，已經能達到很好的表現，但在真實的對話系統應用情境下，資料量仍需再擴增，否則可能無法應對使用者的所有問題。

5.4 未來研究方向

本研究目前能完成的有限，未來我們認為可以將研究的架構擴大，可以解決包含更多數據、更多不同種類的醫療保健問題，在心理健康方面可以與心理諮商機構以及自殺專線等配合，能更有效地提供幫助。另一個方向，可以選擇考慮不單單只使用社群的資料進行研究，若與警察局、諮商機構合作，就能以更貼切的方式來分析對話，打造一個更加完善的對話系統。

在分類任務上，若是出現許多可以讓患者討論的論壇，並能蒐集到更多的中文資料，相信還有很大的改進空間，我們也能使用如 GPT-2、ERNIE 等其他文字模型來研究，或許會有更好的成果。

附錄



附錄一、壹心理平台原始問答範例



感觉无法面对真实的自己，也许困惑太多，也许是别的？

19个回答

08-11 1201 阅读

伴随着做心理咨询（十年抑郁）去年今天的这个时候，我走出抑郁！现在在做成长型咨询。困惑挺多的，从记事起不断的有各种各样的困惑，而自己也深受困惑的影响（现阶段，因为要面对与抑郁症斗争落下的很多问题，个人成长、生活、工作的很多东西不懂，人情世故也懂不娴熟，经常会因为别人说我欠缺什么，或者什么问题，而深受影响并沉浸其中（其实，很多事确实不是事，但，我会沉浸其中，各种纠结）还有工作等与他人的差距.....）是有很多痛苦，无法投入真实的当下，咨询师告诉我，是我没有办法面对真实的自己，我也知道这个问题，也告知了我一些方法，（我也能够觉察到我的感受，不想让我面对真实的自己）不知道该怎么做，不想让自己那么痛苦，该怎么打破这个僵局？让自己更好的投入的生活当中（说起来惭愧，之前设立的目标，因遇到事情自己沉浸负面东西当中，耽误闲置！自己也说了很多大话，以为自己能够坚持做下去，其实生活变化很多）麻烦大家帮我指点一下

楼主悬赏 ¥73.00

收起

悬赏截止日期2022年08月25日 17时47分, [点击查看如何可以瓜分奖金](#)

成长

自我成长

工作学习

自我接纳



1星优质答主 累计帮助了18人



题主，你好！

反复看了好几遍你的描述，感觉你目前还是有些抑郁情绪的，我有几点困惑，之前有服药治疗吗？抑郁症的好转会有一个清晰的节点吗？还是说那个日子是治疗的转折点呢？

我读完总结起来感觉你的困惑分为两个层面：

第一是现实层面的问题。长期抑郁导致自己应对生活的模式会有点不尽如人意。

第二层是认知层面的问题。会特别在意别人的评价，感觉与现实脱节，不愿接受目前的自己。

现实问题大多是能力的问题，是能力就可以培养，可以练习，就像其他老师说的，把目标订到最小，可量化，可评估，并在达成后给自己奖励。可以是一朵花，一点好吃的，一件衣服等等，或者干脆肆无忌惮地发呆玩手机而不自责，都可以是奖励。

这也是认知行为疗法的宗旨，要相信这个看起来不起眼的办法，这是个自我赋能建立自信的过程。

那么认知层面的问题，可以依托咨询师继续成长，最好是认知流派的老师，一起揪出那些不合理信念，建立新的认知模式。

自己练习的时候，可以试试“情绪释放技术”。百度或者直接各大视频网站直接搜索这几个字就可以看到具体操作。很简单，全程几分钟的时间就可以完成。

很抱歉，我能做到的就这么多。隔空抱抱的话我就不说了，我知道自己的局限，我隔空给不到你需要的温度，我码了这么多字，依然很可能帮不到你，我接受自己的局限，你也别太为难自己。

祝好！

赏金截止时获赞11个



关注



有用 11

评论 1

08-12

楼主反馈 1

举报

详情

附錄二、在對話系統串接的部分，我們在論文正文中只展現「成就」主題的回覆生成結果，以下依序展現所有 7 種主題的對答結果，分別為成就、存在議題、生心理問題、人際、親密關係、性格、童年：



主題：〈成就〉		
問題	態度	意圖
對現在的工作不滿意，我該換嗎？	中立	詢問
回覆		
你可以試著找到自己的興趣愛好，去做一些職業生涯規劃。	正向	指導
主題：〈存在議題〉		
問題	態度	意圖
什麼都失敗，不想活了。	負向	告知
回覆		
你可以找到自己想要的的生活方式，也許你會有新的發現。	正向	指導
主題：〈生心理問題〉		
問題	態度	意圖
怎麼才能知道自己是否有心理疾病，比如說憂鬱症？	中立	詢問
回覆		
建議你去醫院精神科檢查一下，如果發現有問題，可以尋求專業人士幫助。	正向	指導
主題：〈人際〉		
問題	態度	意圖
如何與朋友交往？	中立	詢問
回覆		
多多溝通，讓彼此瞭解，瞭解內心的想法和感受。	正向	指導



主題：〈親密關係〉		
問題	態度	意圖
被告白了，但我只想普通朋友，我應該怎樣做，怕拒絕他會破壞關係？	中立	詢問
回覆		
不要刻意排斥他，讓他知道你的想法，試著去解決他的問題。	正向	指導
主題：〈性格〉		
問題	態度	意圖
遇到每件事情，總是怕事情做不好，怎麼辦？	負向	詢問
回覆		
你可以試著去了解自己的內心，不要太在意別人的看法，這樣你才能更好地接受你的想法。	正向	指導
主題：〈童年〉		
問題	態度	意圖
家裡出了問題，不想回家了，怎麼辦？	負向	詢問
回覆		
你可以試著和家人溝通一下，讓他們了解你的想法。	正向	指導

Reference



- [1] Adamopoulou, E., & Moussiades, L. (2020). An overview of chatbot technology. IFIP International Conference on Artificial Intelligence Applications and Innovations,
- [2] Addis, M. E., Truax, P., & Jacobson, N. S. (1995). Why do people think they are depressed?: The Reasons For Depression Questionnaire. *Psychotherapy: Theory, Research, Practice, Training*, 32(3), 476.
- [3] Athota, L., Shukla, V. K., Pandey, N., & Rana, A. (2020). Chatbot for Healthcare System Using Artificial Intelligence. 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO),
- [4] Belfin, R., Shobana, A., Manilal, M., Mathew, A. A., & Babu, B. (2019). A graph-based chatbot for cancer patients. 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS),
- [5] Cameron, G., Cameron, D., Megaw, G., Bond, R., Mulvenna, M., O'Neill, S., Armour, C., & McTear, M. (2017). Towards a chatbot for digital counselling. Proceedings of the 31st International BCS Human Computer Interaction Conference (HCI 2017) 31,
- [6] Cui, Y., Che, W., Liu, T., Qin, B., & Yang, Z. (2021). Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3504-3514.
- [7] Dejan, T. I., & MARKOVIÄ, B. M. (2016). Possibilities, Limitations And Economic Aspects Of Artificial Intelligence Applications In Healthcare. *Ecoforum Journal*, 5(1).
- [8] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, [abs/1810.04805](https://arxiv.org/abs/1810.04805). <http://arxiv.org/abs/1810.04805>
- [9] Dharnapuri, C. M., Agarwal, A., Anwer, F., & Mahor, J. (2022). AI Chatbot: Application in Psychiatric Treatment and Suicide Prevention. 2022 International Mobile and Embedded Technology Conference (MECON),
- [10] Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. Nebraska symposium on motivation,
- [11] Fadhil, A., & Gabrielli, S. (2017). Addressing challenges in promoting healthy lifestyles: the al-chatbot approach. Proceedings of the 11th EAI international conference on pervasive computing technologies for healthcare,
- [12] Fitzpatrick, K. K., Darcy, A., & Vierhile, M. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. *JMIR mental health*, 4(2), e7785.
- [13] Ha, S., Marchetto, D. J., Dharur, S., & Asensio, O. I. (2021). Topic classification of electric vehicle consumer experiences with transformer-based deep learning. *Patterns*, 2(2), 100195. <https://doi.org/https://doi.org/10.1016/j.patter.2020.100195>
- [14] Isometsä, E. T., Henriksson, M. M., Aro, H. M., Heikkinen, M. E., Kuoppasalmi, K. I., & Lönnqvist, J. K. (1994). Suicide in major depression. *The American*

- journal of psychiatry.*
- [15] Jain, P., Srinivas, K. R., & Vichare, A. (2022). Depression and Suicide Analysis Using Machine Learning and NLP. *Journal of Physics: Conference Series*,
- [16] Liddy, E. D. (2001). Natural language processing.
- [17] 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 preprint arXiv:1907.11692*.
- [18] Lokman, A. S., Zain, J. M., Komputer, F., & Perisian, K. (2009). Designing a Chatbot for diabetic patients. *International Conference on Software Engineering & Computer Systems (ICSECS'09)*,
- [19] Louise, S., Fitzpatrick, M., Strauss, C., Rossell, S. L., & Thomas, N. (2018). Mindfulness-and acceptance-based interventions for psychosis: Our current understanding and a meta-analysis. *Schizophrenia Research*, 192, 57-63.
- [20] Mporas, I., Lyras, D. P., Sgarbas, K. N., & Fakotakis, N. (2007). Detection of dialogue acts using perplexity-based word clustering. *International Conference on Text, Speech and Dialogue*,
- [21] Pang, B., Lee, L., & Vaithyanathan, S. (2002). *Thumbs up? sentiment classification using machine learning techniques* Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, <https://doi.org/10.3115/1118693.1118704>
- [22] Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*,
- [23] Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., & Leenaars, A. (2010). Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3, BII. S4706.
- [24] Pham, X. L., Pham, T., Nguyen, Q. M., Nguyen, T. H., & Cao, T. T. H. (2018). Chatbot as an intelligent personal assistant for mobile language learning. *Proceedings of the 2018 2nd International Conference on Education and E-Learning*,
- [25] Piau, A., Crissey, R., Brechemier, D., Balardy, L., & Nourhashemi, F. (2019). A smartphone Chatbot application to optimize monitoring of older patients with cancer. *International journal of medical informatics*, 128, 18-23.
- [26] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- [27] Radziwill, N. M., & Benton, M. C. (2017). Evaluating quality of chatbots and intelligent conversational agents. *arXiv preprint arXiv:1704.04579*.
- [28] Rajalakshmi, S., Asha, S., & Pazhaniraja, N. (2017). *A comprehensive survey on sentiment analysis*. <https://doi.org/10.1109/ICSCSN.2017.8085673>
- [29] Ravitz, P. (2004). The interpersonal fulcrum: interpersonal therapy for treatment of depression. *Can J Psychiatry Bull*, 15-19.
- [30] Reyes, A., & Rosso, P. (2012). Making objective decisions from subjective data: Detecting irony in customer reviews. *Decision Support Systems*, 53(4), 754-760. <https://doi.org/https://doi.org/10.1016/j.dss.2012.05.027>
- [31] Rizzo, A., Shilling, R., Forbell, E., Scherer, S., Gratch, J., & Morency, L.-P. (2016). Autonomous virtual human agents for healthcare information support and clinical interviewing. In *Artificial intelligence in behavioral and mental health care* (pp. 53-79). Elsevier.
- [32] Robertson, S. S. M. (2006). 憂鬱症最新療法-人際心理治療的理論與實務 (唐

- 子俊、唐慧芳、黃詩殷、戴谷霖, Ed.).
- [33] Schuurmans, J., & Frasincar, F. (2020). Intent Classification for Dialogue Utterances. *IEEE Intelligent Systems*, 35(1), 82-88.
<https://doi.org/10.1109/MIS.2019.2954966>
- [34] Sousa, M. G., Sakiyama, K., Rodrigues, L. d. S., Moraes, P. H., Fernandes, E. R., & Matsubara, E. T. (2019, 4-6 Nov. 2019). BERT for Stock Market Sentiment Analysis. 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI),
- [35] Stewart, W. F., Ricci, J. A., Chee, E., Hahn, S. R., & Morganstein, D. (2003). Cost of lost productive work time among US workers with depression. *Jama*, 289(23), 3135-3144.
- [36] Su, C.-E. (2022). Multi-label Dialogue Generation for Depression Patients.
- [37] Sun, C., Qiu, X., Xu, Y., & Huang, X. (2019). How to Fine-Tune BERT for Text Classification? In M. Sun, X. Huang, H. Ji, Z. Liu, & Y. Liu, *Chinese Computational Linguistics* Cham.
- [38] Wei, Z., Liu, Q., Peng, B., Tou, H., Chen, T., Huang, X.-J., Wong, K.-F., & Dai, X. (2018). Task-oriented dialogue system for automatic diagnosis. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),
- [39] Weizenbaum, J. (1976). Computer power and human reason: From judgment to calculation.
- [40] WHO. (2021). *WHO depression statistics*.
- [41] Yang, W., Zeng, G., Tan, B., Ju, Z., Chakravorty, S., He, X., Chen, S., Yang, X., Wu, Q., & Yu, Z. (2020). On the generation of medical dialogues for COVID-19. *arXiv preprint arXiv:2005.05442*.
- [42] Zhang, P., & Huang, L. (2020). *Research on Chinese Intent Recognition Based on BERT pre-trained model* Proceedings of the 2020 5th International Conference on Mathematics and Artificial Intelligence, Chengdu, China.
<https://doi.org/10.1145/3395260.3395274>
- [43] Zhou, L., Gao, J., Li, D., & Shum, H.-Y. (2020). The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1), 53-93.
- [44] 張家銘. (2006). 正視憂鬱症造成的社會經濟負擔.