

長庚大學資訊工程學系

碩士論文

Graduate Institute of Computer Science and Information Engineering

Chang Gung University

Master Thesis

透過多模型集成的方式整合使用者 MBTI 以提升聊天
機器人個人化的回覆效果

Enhancing the personalized response capability of
chatbots by integrating user MBTI through a multi-model
ensemble approach

研究生: 黃建銘

Graduate Student: Chien-Ming Huang

指導教授: 張賢宗 教授

Advisor: Hsien-Tsung Chang Ph.D.

中華民國 113 年 1 月

January, 2024

長庚大學碩（博）士學位論文 指導教授推薦書

資訊工程學 系（所） 黃建銘 君所提之論文

透過多模型集成的方式整合使用者 MBTI 以提升聊天機器人個人化的

回覆效果

係由本人指導撰述，同意提付審查。

指導教授 張國強 (簽名)

113 年 1 月 10 日

長庚大學碩士學位論文口試委員會審定書

資訊工程學系 研究所 黃建銘 君所提之論文

透過多模型集成的方式整合使用者 MBTI 以提升聊天機器人個人化的

回覆效果

Enhancing the personalized response capability of chatbots by integrating user MBTI through a multi-model ensemble approach

係本委員會審議，認為符合碩（博）士資格標準。

論文口試委員會

召集人 李博暉 (簽名)

委員 李博暉

張雲峰

系主任、所長

陳一輝

中華民國 113 年 1 月 29 日

致謝

感謝 Kaggle 的 (MBTI) Myers-Briggs Personality Type Dataset 競賽提供資料集作為本實驗的訓練及驗證資料。感謝 16Personalities 網站之測定結果作為本實驗的對照人格標準。感謝所有幫助我完成測試問卷的人，沒有你們的參與，我是很難在有限的時間內蒐集到這些的數據，同時也感謝在問卷中提供建議的人，因為有這些回饋讓我在實驗修改以及結果分析的撰寫有很大的幫助。接著我要特別感謝幾位在本篇論文中對我影響深遠的人，首先是我的指導教授張賢宗教授，不只是提供了我研究用的所有設備，還包含了實驗耗材、API 等等的所有費用，在我每一次的進度報告中，總是能點出我沒有注意到的要點，並耐心的解釋，且當我遇上問題的時候，他的建議總能醍醐灌頂，使遇到的困難迎刃而解。因為有他的指導，成為了我推進研究的最大助力。再來是就讀長庚大學人工智慧研究所的張如億同學，因為有她的研究報告才會讓我作為發想的起點產出現在這篇論文，並且也感謝她陸續提供相關的論文賞析作為我的參考資料。最後是就讀台北科技大學資訊工程研究所的官慶恩同學，感謝他在最終測試的部分給予我許多寶貴的意見，讓我能及時的修正錯誤。最後我要感謝我的家人們，在研究的這段期間，我遇上了不少的難關和低谷，對我的情緒和心態有莫大的打擊，但因為有你們的包容和鼓勵，讓我也能夠重新振作，再次面對並且戰勝一層又一層的挑戰，最終完成這篇研究。謝謝你們，有你們作為我的後盾，我很幸福。

摘要

聊天機器人得力於大型語言模型的開發，在生成上已經有高質量的結果。儘管如此，若是談及針對使用者的個人化需求，卻難以在對話中自動地展現。對此本篇論文研究的重點將聚焦於實現讓機器人在對話中自動理解使用者，以改變回覆的風格。具體來說，我們透過提出的方法來剖析使用者的對話，在得到他的 MBTI 人格後，將其轉為資訊套用於對話機器人上，來增加使用者的對話滿意度。為了實現目標，我們將工作分為 MBTI 的預測任務以及對話機器人的設計，對於前者我們首先依據不同的資料形式微調了兩個 BERT 模型，並以其一為基礎修改了資料集的特徵，用於後續三個機器學習模型的訓練。對於後者我們以 GPT-3.5 turbo 作為機器人的對話生成模型，在與使用者對話時，系統會將對話記錄依序傳送給各個預測 MBTI 的模型，由於輸出的結果不只一個，我們採用多數決投票機制來集成最終的結果，並以此作為使用者的 MBTI 人格套用至對話機器人上。就結果而言，在模型性能上，我們提出的兩個方法分別得到了 66.25% 和 98.76% 的準確率，超過了以往的各項 baseline，而在實測成果中，有 64% 的受測者認為回覆的語意更好理解，76% 的受測者認為整體的對話感受更加舒適，證實我們的方法能有效地應對使用者改善機器人回覆。最後，我們也分享了我們在實驗中做出的各種嘗試，並且闡述了這項研究中值得深思及探討的幾個要點。

關鍵字：MBTI 預測、聊天機器人、集成學習、GPT-3.5、BERT

Abstract

This paper introduces a novel method to improve chatbot personalization by incorporating MBTI personality type prediction into interactions. It aims to automatically discern users' MBTI types from dialogues and tailor responses to individual needs. The methodology involves two primary tasks: the prediction of users' MBTI personality types using fine-tuned BERT models and the development of a chatbot using GPT-3.5 turbo for dialogue generation. The MBTI prediction involves processing dialogue records through the models and employing a majority voting mechanism to ascertain the user's personality type. This information is then used to tailor the chatbot's responses, aiming to increase user satisfaction. The proposed methods achieved significant accuracies of 66.25% and 98.76%, surpassing existing baselines. User tests demonstrated the effectiveness of the approach, with 64% of participants reporting improved understanding of responses and 76% experiencing a more comfortable conversation. These results highlight the potential of MBTI-based personalization in enhancing chatbot interactions. The paper also discusses various experimental insights and key areas for further research in chatbot personalization.

Keywords: MBTI prediction, Chatbot, Ensemble Learning, GPT-3.5, BERT

目錄

	Page
致謝	i
摘要	ii
Abstract	ii
目錄	iii
圖目錄	vi
表目錄	vii
第一章 緒論	1
1.1 研究背景	2
1.2 研究動機	2
1.3 研究目的	4
第二章 文獻探討	6
2.1 MBTI 預測任務	6
2.2 對話機器人	9
第三章 研究方法與評估	11
3.1 資料預處理	11
3.1.1 資料集	12
3.1.2 資料清洗	13

3.1.3 處理未平衡資料	13
3.2 訓練階段	15
3.2.1 BERT	15
3.2.2 將序列由多句轉成單句	16
3.2.3 特徵提取	17
3.2.4 將 MBTI 文本標籤序列轉換成比例分布序列	17
3.2.5 以 MBTI 比例序列進行分類模型的訓練	18
3.2.5.1 SVM	18
3.2.5.2 ELM	18
3.2.5.3 Random Forest	19
3.2.6 GPT-3.5 turbo	19
3.3 實際使用	19
3.3.1 多數決投票機制	22
3.4 評估指標	22
3.4.1 模型性能	23
3.4.1.1 Accuracy	23
3.4.1.2 Precision	23
3.4.1.3 Recall	23
3.4.1.4 F1 score	24
3.4.1.5 MBTI 的 4 個向度準確率	24
3.4.2 系統實測	24
3.4.2.1 使用者問卷	24
3.4.2.2 同類型測驗結果比較	25
第四章 實驗結果與討論	26
4.1 實驗環境	26
4.2 模型性能之結果展示	27
4.3 與現有方法之數據比較	29

4.4 實際測試之結果展示	30
4.4.1 MBTI 人格測定的準確度結果	31
4.4.2 前後選擇的結果 & 在對話中語意及感受上的評分	32
4.5 結果討論	35
4.5.1 初始測試的 12 題問答的效果以及準確度	35
4.5.2 為何要以 BERT-multi 和 BERT-Combo 的結果來做最終的集成投票	37
4.5.3 第一階段投票選擇 3 個方法的論點探討	38
4.6 實驗嘗試	38
4.6.1 以 BERT-multi 改寫資料集後訓練的模型	39
4.6.2 資料集依 n-gram 重組	40
4.6.3 轉換為 MBTI 比例序列之差別	40
4.6.4 在 MBTI 比例序列中刪除某些人格指標	41
第五章 結果與未來展望	42
5.1 結論	42
5.2 未來展望	43
5.2.1 使用規模更龐大的模型來實作	43
5.2.2 和機器人運行的平台中的其他元素結合	44
5.2.3 準備更大規模 & 多元 & 多語言的資料集	44
5.2.4 結合針對「連續對話」而建立的訓練資料集	45
5.2.5 以最大化使用者對話中的人格特徵為目標的改良	46
參考文獻	47
附錄 A — 使用者問卷形式	52

圖目錄

1.1 MBTI 組成示意圖	3
1.2 MBTI 的 16 種不同人格示意圖	4
1.3 系統架構示意圖	5
3.1 資料預處理流程圖	11
3.2 資料分割後的數量圖	12
3.3 資料平衡示意圖	13
3.4 資料平衡後的數量圖	14
3.5 訓練階段流程圖	15
3.6 序列由多句轉成單句示意圖	16
3.7 特徵提取示意圖	17
3.8 系統實際運行架構流程圖	20
3.9 3 種機器學習模型的集成方法流程圖	20
3.10 多數決投票機制示意圖	21
4.1 受測者結果於四個向度 (E.I、S.N、T.F、J.P) 指標中的正確數量占比圖	32
4.2 受測者根據第一輪前、後對話選擇的占比圖	33
4.3 受測者在對話中語意及感受上的評分圖	34
4.4 三位受測者在與套用與否 MBTI 人格的機器人聊天內容範例	34
4.5 初始測驗 12 題的題目測試示意圖	36
4.6 第一輪對話二階段投票的 3 種方法之正確數	36
4.7 初始問答 & 第一 & 第二輪對話之正確數	36
4.8 刪除人格指標的兩個方法示意圖	41
A.1 使用者問卷-階段一	52
A.2 使用者問卷-階段二	53

表目錄

4.1 硬體實驗環境說明表	27
4.2 軟體實驗環境說明表	27
4.3 實驗中重要的函式庫說明表	27
4.4 實驗中的方法之 Accuracy、Precision、Recall 和 F1 Score 的分數	28
4.5 實驗中的方法之結果以四個向度 (E.I、S.N、T.F、J.P) 的 準確率	28
4.6 和相關研究中同樣使用 Kaggle 資料集並且將成果分成四 個向度 (E.I、S.N、T.F、J.P) 來比較表 [1][2][3][4][5][6][7]	29
4.7 和參考文獻 [8] 中使用已過濾資料集進行訓練的方法比 較 J/P 之準確率	30
4.8 以實際測試的結果展示本研究使用方法於四個向度 (E.I、 S.N、T.F、J.P) 的準確度	37
4.9 以 BERT-muti 改寫資料集後訓練的方法之 Accuracy、 Precision、Recall 和 F1 Score 的分數	39
4.10 以 BERT-muti 改寫資料集後訓練的方法以四個向度 (E.I、 S.N、T.F、J.P) 的準確率	39

第一章 緒論

隨著自然語言處理技術的飛速發展，聊天對話型機器人已從最初的規則式系統演進為今日的人工智慧驅動模型。這些機器人的應用範圍極為廣泛，不僅涵蓋了多項專業領域，同時也深入到我們日常生活的各個層面，對我們的互動方式和行為模式產生了深遠的影響。特別是自 2022 年 10 月推出的 ChatGPT[9]，憑藉其深度學習和大數據分析的強大結合，提供了高度精確且豐富的對話輸出，迅速成為人們在徵詢意見、探索信息或面臨挑戰時第一個會想到的應對方案。

儘管 ChatGPT 在對話生成方面的表現令人印象深刻，但當涉及到滿足用戶對個性化對話的需求時，它的侷限性也開始顯現。目前的技術難以全面理解和反映用戶的人格和情緒特質，這是因為每個人都具有獨特的人格和思考模式。而且，當前的系統還無法自動化地根據用戶的對話進行人格分析，這進一步限制了其在個性化交流上的能力。正因如此，本研究旨在探索一種新方法，將 MBTI 人格分類指標應用於聊天機器人，以提升其在個性化對話方面的效用。

本研究將開發一個創新的系統，該系統融合了自然語言處理技術和 MBTI 人格理論。利用 gpt-3.5 turbo 作為聊天機器人的生成核心，配合 BERT 模型對用戶對話進行分析，從而預測出相應的 MBTI 指標。這些指標將通過多個機器學習模型進行預測，最終透過投票機制進行集成來確定用戶的 MBTI 人格類型，作為聊天機器人對話生成的參考資訊，目標在於實現更精確、個性化的回應。

1.1 研究背景

聊天對話型機器人在近年來透過技術不斷的進步，也漸漸的茁壯了起來，從原先預設對話的規則式聊天機器人，到現今運用複雜的自然語言工程，來實現開放性對話的人工智慧聊天機器人，這些技術的進步不只促進了聊天機器人的應用範疇，從客服支援到教育輔導，再到醫療諮詢，還加深了它們在日常生活中的影響力。而現在，一提到聊天對話機器人，大部分的人第一個聯想到的，無非是在 2022 年 10 月提出的 ChatGPT 了，結合了深度學習和大規模數據分析，展現出卓越的對話理解和生成能力。其擁有的技能不單單只是做到擬人化的對話互動，更令人震驚的是其對於使用者對話所展現的超高質量輸出，這樣的對話架構不僅符合人類的對話邏輯，其內容更是融合了相關知識所產生的精華。對眾多的使用者而言，輸出的成果若能符合心中的期望，就稱得上是滿意答覆了，也因此 ChatGPT 成為了許多人在尋求建議、獲取資訊或解決問題時的首選參考依據。

1.2 研究動機

雖然 ChatGPT 生成的結果從理性的角度上來看，都會讓人讚嘆。但是如果使用者對於對話機器人的任務不再只是注重輸出的內容，而是同樣關注對於回話者的個人化需求，例如：回覆風格、角色要素……等等，那麼現階段的 ChatGPT 將不再是最優秀的選擇。固然有著強大的語言模型來進行語意的理解和內容的生成，但對於機器人而言還是難以自動的去理解使用者的人格和情緒，如果你希望機器人能夠針對你現階段的人格來做回覆，則你必須要在對話中將這些文字作為情報輸入給機器人，他才會根據你的對話來去修正。

你可能會很好奇，理解使用者的人格再去做回應對機器人而言有很大的影響嗎？老實說，對機器人的影響只在於其生成時所需要額外參考的資訊，但對於使用者來說，讓機器人生成的一句話根據自己的特質而變成自己最理想的對話，那就別具意義了。同樣的一句話，每個人的解讀可能不同，因為每個人都是獨立的個體，有著不同的人格、思考模式，因此，如果我們能夠嘗試讓機器人透過理解使用者的人格資訊，來將原先要輸出的對話內容進行調整，改變成讓使用者閱讀起來最能夠理解和接受的文句，那我相信這會讓聊天機器人的前景再向前邁進一步。



Figure 1.1: MBTI 組成示意圖

儘管過往的研究中，並沒有明確表示關於聊天機器人對話風格的界定依據，但如果是以人的對話來進行人格判斷，卻有著相當豐富的研究歷史，其中最代表性的正是邁爾斯-布里格斯性格分類指標 (Myers-Briggs Type Indicator，簡稱 MBTI)，是人格類型學中一種內省的自我報告問卷，它表明人們在感知和決策的方式上存在不同心理偏好。該評量為以下四個向度分別指定一個傾向 (可參考 Figure 1.1)：內向 (Introversion) 或外向 (Extraversion)、實感 (Sensing) 或直覺 (Intuition)、思考 (Thinking) 或情感 (Feeling)、判斷 (Judging)

ISTJ	INFJ	INTJ	ENFJ
ISTP	ESJF	INFP	ESPF
ENFP	ESTP	ESTJ	ENTJ
INTP	ISFJ	ENTP	ISFP

Figure 1.2: MBTI 的 16 種不同人格示意圖

或感知（Perceiving），並取各傾向的英文名稱的字首來組成 4 個字母的評量結果，以區分 8 大指標，總共能夠組合出 16 種不同的人格取向（可參考 Figure 1.2）。對於 MBTI 這個指標而言，我認為其能夠充分的顯示人們對於事物的判斷與偏好，透過指標代表的特性，能更簡單地去理解一個人的性格，以及他的行動所表達的涵義，而這一點若是套用在 ChatGPT 的回覆上，將能彌補機器人對於個別使用者在對話個性上認知的不足。

1.3 研究目的

在本研究中，我們旨在開發一個創新的系統（可參考 Figure 1.3），該系統結合了先進的自然語言處理技術和心理學理論，以精確預測和應用 MBTI 人格類型於聊天機器人的對話過程中，期望透過這樣的設計實現自動地個人化聊天回覆。就研究細項來探討，我們會依據對話任務的特性修改資料集來訓練數個模型作為 MBTI 預測的工具，並且實作出一個以 MBTI 人格套用於聊天機器人對話的系統。具體而言，我們以 gpt-3.5 turbo 作為聊天機器人的對話生成核心，透

過 BERT 模型來對使用者的對話進行判斷，並生成對應的 MBTI 指標，然後再將這些指標的序列轉化為比例序列傳送至 ELM、SVM 和 Random Forest 模型進行預測，三者預測出來的結果將會進行投票，並由最高票的結果和 BERT 輸出的結果以及使用者原本的 MBTI 人格，三者再進行一輪的投票判斷，最終得到的 MBTI 人格將作為使用者的回覆人格，提供 gpt-3.5 turbo 進行對話生成時需參考的額外資訊。

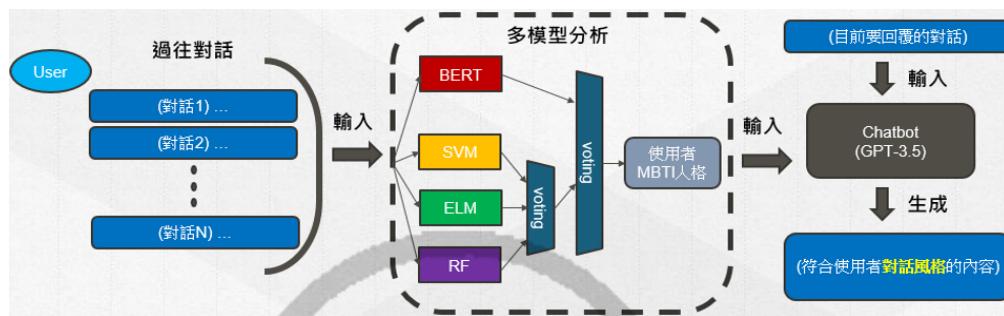


Figure 1.3: 系統架構示意圖

本研究提出的貢獻主要如下：

1. 整合 MBTI 人格分析於聊天機器人：我們提出一種新穎的方法，將傳統的 MBTI 人格測試與現代的自然語言處理技術結合，用於提升聊天機器人對話的個性化和準確性。
2. 複合模型預測系統：通過融合多種機器學習模型和投票機制，我們的系統能夠更全面地分析和預測使用者的 MBTI 類型，從而提供更精準的個性化對話體驗。

第二章 文獻探討

本章節主要以兩個主題來探討於本研究的相關文獻。首先，我們將回顧關於 MBTI 預測任務的研究，深入瞭解先前的工作在此領域中所取得的成果與挑戰。接著，我們將探討有關對話機器人的相關研究，以了解現有的對話系統發展和相關技術的最新進展。這兩個主題的研究將有助於我們更全面地理解研究背景並奠定本研究的基礎。

2.1 MBTI 預測任務

MBTI 性格分類法早在 1956 年就由 Katharine Cook Briggs 和她的女兒 Isabel Briggs Myers¹提出，目的是發展一種能夠幫助人們更好地了解自己和他人的工具，當時測定的方法是使用一系列的問題和情境描述，要求受試者根據他們的偏好和傾向作出選擇。而近年，電腦資訊逐漸發達，人工智慧、機器學習等等的演算法出現，將許多原先繁瑣的工作由先進的技術代為執行，而 MBTI 的預測也不例外，透過大數據分析以及類神經網路模型的訓練，讓這項任務能夠經由科技來實現。為了進行自動化的 MBTI 人格偵測任務，Plank 和 Hovy (2015)[10] 率先使用了 Twitter 上的資料來作為訓練使用的資料集，他們一共蒐集了 120 萬條的英文推文，其中蒐集的要點為此人必須明

¹Katharine Cook Briggs 和她的女兒 Isabel Briggs Myers 是 Myers-Briggs Type Indicator (MBTI) 人格分類系統的創始人。

確的標記自身屬於 16 種 MBTI 標籤中的其中一種，此外，除了對話人所屬的 MBTI 標籤之外，還包含了性別標籤。最終，每個被選定的使用者都蒐集至少 100 則推文來作為資料數據。而作者也透過特徵和模型演算法進行訓練，並以四個不同向度 (E.I、S.N、T.F、J.P) 來進行檢測，就結果而言，E.I、T.F 的結果較其他的 baseline 還要優秀。而與之相比較為不同的是 Verhoeven 等人 (2016)[11] 提出的一種用來對文本提取 MBTI 的策略，他從 Twitter 上蒐集使用者的推文，並且根據其中一些關鍵字來判斷這個人的 MBTI 人格，其實踐的方法是找出文中寫道” I am” 或是” I have” 等文字，並在後續的文字為 16 個人格標籤的其中一項來判定，而在標記完成後，作者也透過手動檢查上下文來檢測是否符合，最終他蒐集了 1362 位使用者，每一位有 200 條推文，且除了 MBTI 標籤外也包含了性別資訊，在驗證方面同樣是以四個不同向度來區分，和其他的 baseline 相比，同樣是在 E.I、T.F 的結果更加出色。

關於預測 MBTI 使用的分類模型，包含了基本分類模型和類神經網路模型，從 Raje 和 Singh (2017)[12] 研究中提出的證明可得，對於此類型的文本分類任務，基本分類模型能夠展現較好的性能。常見的基本分類模型包含了樸素貝葉斯 (NB)；支援向量機 (SVM)；決策樹 (J48)；多層感知器 (MLP)；和 K 最近鄰 (KNN)，在 Bharadwaj 等人 (2018)[1] 的 MBTI 預測成果中，使用了 SVM、樸素貝葉斯和類神經網路分類器，最終是 SVM 在三個不同的資料集上奪下佳績。而在 Lima 等人 (2019)[13] 的氣質 (使用由 David Keirsey 提出的 Keirsey Temperament Sorter 來進行) 與心理 (使用 MBTI 來進行) 推斷實驗中，以上述提及的五種基本分類模型演算法來進行評估，從結果來說，由決策樹衍生的隨機森林 (random forest) 在兩個類型的預測中都取得了最佳的性能。

往後的研究也都針對分類模型進行改良，並且根據過去的 baseline 之結果來進行應對，Choong 和 Varathan (2021)[8] 就採用了由 Ke 等人 (2017)[14] 提出的 LightGBM 來做為實驗的其中一種分類器，其主要原因在於之前的 baseline 在 MBTI 的四個向度中，除了 J.P 之外表現都相當穩定，而 J.P 的判斷是以某人在做出決策時如何跟外界進行互動來評估，這個部分的特徵相較其他更加稀疏，因此作者認為可以借助 LightGBM 這種梯度增強演算法來增加辨識的成功率，而最終結果也表明了這點。Amirhosseini 等人 (2020)[2] 以及 Nisha 等人 (2022)[15] 則是採用了基於決策樹的機器學習整合演算法-XGBoost 來做為研究主軸，在複數的決策樹中透過權重給定及調整來預測結果，並且整合這些樹形成更具效率及規模的模型，與之前的 baseline 相比，四個向度的平均準確率更高。Fernau 等人 (2022)[3] 為了要捕捉使用者的個人化需求，而決定將 MBTI 的概念引入對話機器人當中。而作為 MBTI 的分類方法，與以往不同的是他們採用了 BERT 等大型語言模型來做為他們的主要方法，原因在於任務的需求必須要讓辨識用的模型具備強大的泛化性，來克服複雜的語意理解和資源調節的困難。最終，他們比較了數個由 BERT 變化而來的模型，並且在和過去使用大型預訓練模型來實作的方法中取得了最先進的結果。Wang 等人 (2023)[16] 設計了一種面試型的評估架構來為角色扮演的機器人進行評估，包含了 MBTI 以及 Big-Five Personality 兩大人格指標，有別於以往封閉式人格測驗問答，他們透過開放式的問題來得到每個帶入角色的機器人生成出來的應答，透過這樣的方法，能從中分析並擷取到更加細緻的人格特點和對話習慣，在最後的性類型評比中和標準答案有高度的一致性，這也表明了在豐富的對談中所隱藏的個人訊息量是相當可觀且不可忽略的。

2.2 對話機器人

對話機器人的應用範疇相當的廣泛，客服支援、電子商務、醫療保健、社群媒體等等都能夠見到它活躍的身影。然而以往的對話機器人，能勝任的任務卻不如現今普遍，主要和自然語言處理與生成技術的成熟與否有強烈的關係。最早期的對話機器人使用的是有限狀態機和規則引擎來實現，在和使用者的對談上有諸多限制，而後是透過統計模型以及機器學習的方法，並且整合初期的自然語言處理技術，來提升機器人對於對話的理解能力。直到深度學習的興起，拜深度神經學習網路所賜，在長文本的對話上展現了不同以往的對話記憶性和專注性，其結合了大數據分析帶來的豐富知識性後，讓整體言語處理的能力有飛躍式的提升。

而在 2018 年，由 OpenAI[17] 的 Alec、Kartheek 等人 (2018)[18] 提出的 GPT-1 橫空出世，以預訓練模型在自然語言處理中的優越性能為核心，通過 transformer 架構、無監督式學習和大規模的參數設置等特點，成為了當時通用的自然語言處理模型。隨後，由 Alec 等人 (2019)[19] 提出 GPT-2，擁有更多的參數量，在生成能力上更為強大，並於各類型的自然語言生成任務上取得佳績，甚至在當時引發了對其生成能力過強而被濫用的隱憂。2020 年，Brown、Alec 等人 (2020)[20] 提出的 GPT-3，以更龐大的參數量為主打，成為目前最為通用的自然語言生成模型。回覆機器人至此在生成回覆訊息時將保有高質量且具上下文語意的內容，然作為回覆端，若是能夠在生成回覆訊息時根據使用者以及當下環境進行個人化的編輯，那機器人將更加符合擬人化的設定。Daniel(2023)[21] 根據心理學上的依據，認為回覆機器人若要根據上下文回覆並且具備人性情感，就需要建立三層的狀態層來輔助 GPT-3 生成，三層狀態層分別是第一層：記錄使用者當下對話；第二層：將第一層對話作為示例來生成對話補全；

第三層：將以往的對話紀錄統整並區分，以因應不同的對話任務需求。這些層會在與使用者對話時作為 prompt 納予 GPT-3 額外資訊，使其生成更具個人色彩的回覆。

而時至今日，若是提起最受人矚目的回覆型機器人，我想所有人的答案都將是由 openAI 於 2022 年底發布的 chatgpt 吧，作為目前最接近人類的對話機器人，不單單是高質量的生成文本，更為人震驚的是極致的擬人化對談，並且能夠根據使用者的需求動態調整。如此強大的模型，原先是由 Ouyang 等人 (2022)[22] 提出的 InstructGPT 模型，他們認為理解使用者意圖是作為回覆機器人的首要任務，而這件事不單是依靠更大的語言模型就能實現，於是他們展示了一種透過根據使用者回覆進行微調，來讓語言模型和使用者保持一致意圖的方法，這使得機器人的回覆變得和使用者高度同步，大大降低了期望落空的可能性。



第三章 研究方法與評估

本章節將研究方法分為資料預處理、訓練階段及實際使用。資料預處理的部分將提及本研究的使用資料以及過濾規則，訓練階段描述的是我們提出的方法在訓練時的各步驟細項，實際使用則是介紹本系統的使用流程以及運作機制。而評估的部分則分為模型性能及系統實測，前者是我們的方法在和 baselines 比較時會用到的評估指標，後者則是受測者使用我們的系統後，會依照使用者感受和預測成果對比這兩者進行評分。

3.1 資料預處理

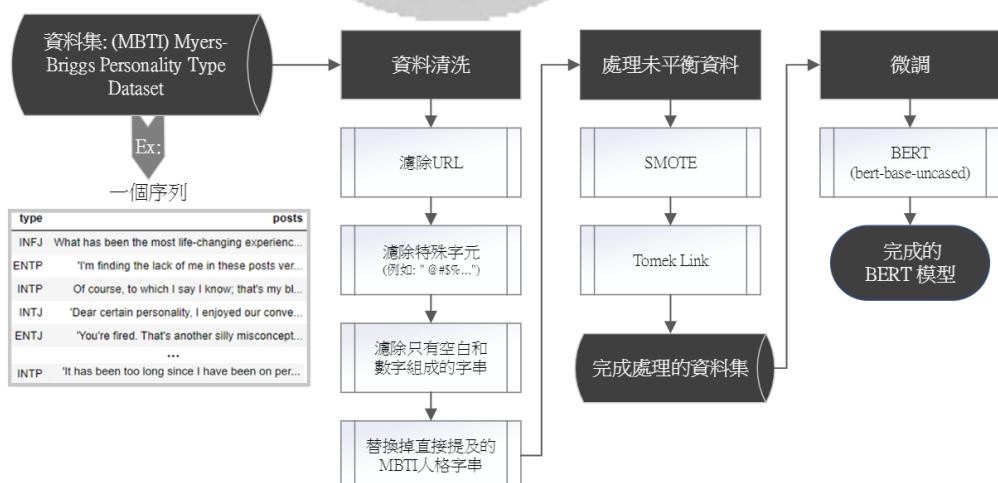


Figure 3.1: 資料預處理流程圖

在此簡述資料預處理的流程 (可參考 Figure 3.1)，其中的各個部分都會在後續有詳細介紹。在 Figure 3.1 中，先將文本資料的 Dataset

的資料進行預處理，包含數據清洗以及數據平衡，接著輸入至 BERT 模型進行微調。

3.1.1 資料集

type	label	data_type	counts	type	label	data_type	counts	type	label	data_type	counts
ENFJ	5	test	28	ESTJ	14	test	6	ISFJ	10	test	25
		train	133			train	27			train	116
		val	29			val	6			val	25
ENFP	7	test	102	ESTP	12	test	13	ISFP	8	test	41
		train	472			train	62			train	190
		val	101			val	14			val	40
ENTJ	4	test	34	INFJ	0	test	221	ISTJ	11	test	30
		train	162			train	1028			train	144
		val	35			val	220			val	31
ENTP	1	test	103	INFP	6	test	275	ISTP	9	test	51
		train	479			train	1282			train	236
		val	103			val	275			val	50
ESFJ	15	test	6	INTJ	3	test	164				
		train	29			train	764				
		val	7			val	163				
ESFP	13	test	7	INTP	2	test	196				
		train	34			train	913				
		val	7			val	195				

Figure 3.2: 資料分割後的數量圖

(MBTI) Myers-Briggs Personality Type Dataset[23]，在 Kaggle 上可以找到，此資料集共有 8675 筆數據，每一筆可分為 type 和 posts，分別代表一個人的 MBTI 人格以及他在 PersonalityCafe[24] 論壇上面發布的貼文或留言，每人最多會蒐集 50 句，並且透過“|||”符號分隔。由於後續的模型訓練及微調都是以此資料集為基礎，因此資料分割統一比例為 [70:15:15]，分別用於訓練、驗證以及測試，關於數據分割的數目可以參考 Fugure 3.2。

3.1.2 資料清洗

前面提及的 Dataset 由於是直接從論壇上抓下來的文句，因此內容很容易出現參差不齊的情況，在此我們參考了文獻 ([2], [3]) 中提出的過濾規則，濾除包括 URL、一些特殊字元 (如 @、\$..... 等等)、只有數字與空格組成的字串、以及將直接提及的 MBTI 人格進行替換，這麼做的目的在於以正常情況下來說，對話中不太會大量穿插 MBTI 人格標籤，但由於這個資料集的來源是和人格相關的論壇中使用者的對話內容，因此在提及 MBTI 人格的頻率就會遠高於普通對話，而以這樣的內容作為資料集素材，會導致後續在訓練及評估時對結果造成影響，對此我們才將這類字串進行替換 (換成' certain personality')，來規避掉這個負面效果。

3.1.3 處理未平衡資料

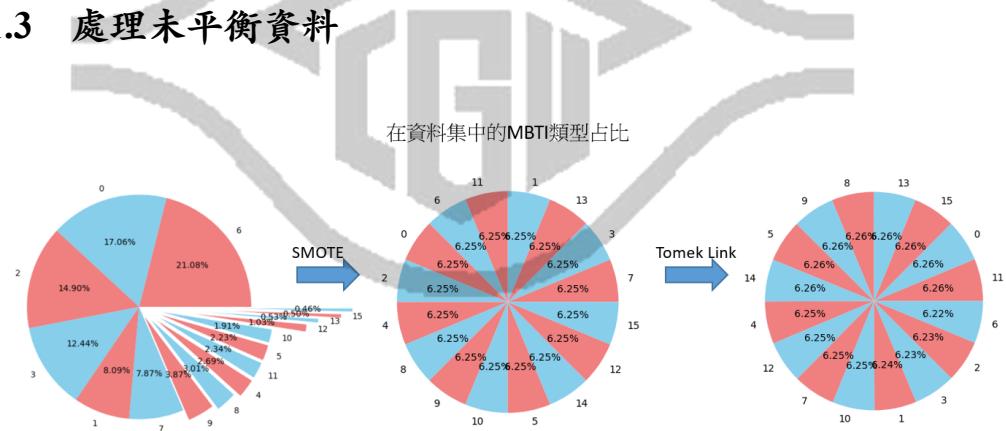


Figure 3.3: 資料平衡示意圖

從 Figure 3.3 左邊的圓餅圖可以發現，資料集分布的極不平均，占比最高的 MBTI 指標數量超過 1800 個，但占比最低的卻不到 100 個，如此大的落差導致的不平衡將大大地影響訓練的結果，在參考文獻 ([3]) 中也有提及這樣的情況，而他們採用了 nlpaug[25] 數據增強庫來擴增樣本數來提高模型泛化能力。對此我們採用上採樣 SMOTE-(Synthetic Minority Over-sampling Technique) 與下採樣 Tomek Link 的方法 (可參考 Figure 3.3)。SMOTE 處理數據的具體流程在於，

type	label	data_type	counts	type	label	data_type	counts	type	label	data_type	counts
ENFJ	5	test	215	ESTJ	14	test	210	ISFJ	10	test	223
		train	1276			train	1279			train	1274
		val	197			val	228			val	210
ENFP	7	test	201	ESTP	12	test	230	ISFP	8	test	200
		train	1263			train	1280			train	1275
		val	207			val	213			val	207
ENTJ	4	test	205	INFJ	0	test	275	ISTJ	11	test	199
		train	1272			train	1282			train	1273
		val	199			val	275			val	211
ENTP	1	test	206	INFP	6	test	201	ISTP	9	test	215
		train	1274			train	1262			train	1275
		val	203			val	188			val	205
ESFJ	15	test	210	INTJ	3	test	201				
		train	1274			train	1271				
		val	221			val	209				
ESFP	13	test	207	INTP	2	test	192				
		train	1279			train	1264				
		val	196			val	194				

Figure 3.4: 資料平衡後的數量圖

其會先隨機選擇少數類別的樣本，然後找此樣本的 K 個最鄰近的樣本，並在選定的樣本和鄰居之間生成新的樣本，之後不斷重複這個步驟來達到數據平衡。而 Tomek Link 用於去除重疊樣本，具體的流程為，先識別互為相鄰的兩個樣本，假若兩者屬於不同類別，則移除其中一個樣本(通常選擇多數類別)。在本研究中，我們期望先以 SMOTE 對少數樣本進行數據合成來增加數目，再透過 Tomek Link 找出類別之間的雜訊，將這些會影響到分類性能的不利因子去除，使模型能更簡單的辨識出樣本真正的類別。兩者交互使用後，得到的數據集比例相較以往更加平衡，更有利於後續的訓練任務，關於數據平衡的數目可以參考 Figure 3.4。為何我們使用 SMOTE 和 Tomek Link 而非採用參考文獻 ([3]) 的方法是在於我們認為解決數據不平衡是改善後續模型訓練的首要目標，和擴增整體樣本數的數據增強相比，我們提出的方法更能有效的解決根本問題。

3.2 訓練階段

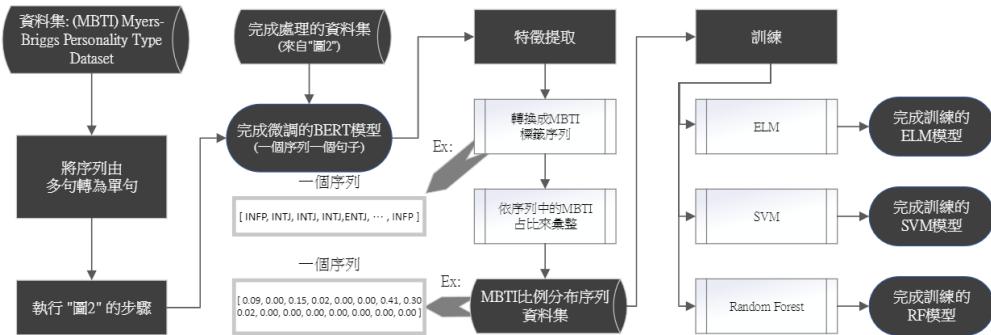


Figure 3.5: 訓練階段流程圖

在此簡述訓練階段時的流程 (可參考 Figure 3.5)，其中的各個部分都會在後續有詳細介紹。由前述的資料預處理完成後，輸入至 BERT 模型進行微調，這裡我們微調了兩個 BERT 模型，第一個完全參照 Figure 3.1 的流程，第二個則是在一開始先將資料集由原先的”一個序列多個句子”的形式轉換成”一個序列一個句子”的形式 (可參考 Figure 3.5 前段)，再繼續後續的流程。完成後我們以第二個 BERT 模型將原先的資料集當中的每一句對話轉換成 MBTI 的人格序列，再進行統計轉換成 MBTI 比例序列 (可參考 Figure 3.5 中、後段)，這麼做的目的就是為了將這個由新的特徵組成的資料集輸出給 SVM、ELM、Random Forest 三個方法進行訓練，最終將完成的模型及參數進行保存，以供後續系統使用。訓練時各階段詳情將在之後一一呈現。

3.2.1 BERT

BERT 採用 transformer[26] 模型中 Encoder 的部分，經由自注意力機制 (Self-Attention Mechanism)，讓模型在處理文本序列時能全局的思考所有字詞的位置，且不同於單向的循環神經模型，雙向的架構使其能在輸入時同步納入左側及右側的內容來考慮，透過以上

的組合能讓模型完整的理解文本中的上下文信息，再配合上擁有大規模語料庫預訓練所獲得的大量知識，讓我們能夠過微調就能夠進行 MBTI 人格的分類，在本研究中，經測試後選用的 BERT 模型為 bert-base-uncased¹[27]，batch-size 為 12、learning rate 為 1e-5，在訓練及評估時使用早停策略以取得最佳結果。特別說明，在整個模型訓練的流程（可參考 Figure 3.1、Figure 3.5）內，總共有兩個 BERT 的模型，使用的預訓練模型是相同的，唯一的區別在於輸入的資料形式不同。其一是用原本”一個序列多句對話”的數據集來做微調，為的是能夠全面的分析所有的對話再進行預測，另一個則是改用”一個序列一句對話”的數據集來微調，為的是聚焦於單句話的涵義，用於後續轉換資料集以進行模型的訓練。

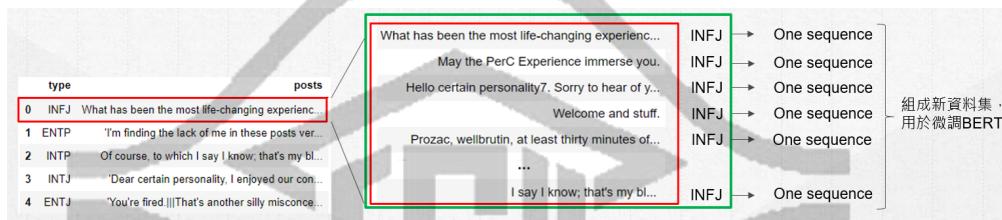


Figure 3.6: 序列由多句轉成單句示意圖

3.2.2 將序列由多句轉成單句

原先的資料集是一個序列中包含一個 MBTI type 加上一個 posts 組成，其中 posts 裡面記錄了這筆資料中的所有對話。為了因應後續的模型訓練，我們需要一個方法來將原資料中 posts 的每句話改成對應的 MBTI，因此我們決定訓練一個以”一個序列只有一句話作為 posts 的資料集”來當作輸入的 BERT 模型做為執行方案。將 posts 中的每一句話分隔出來，對應上原本的 type，各自作為一個序列，組成上述的資料集，就是這個段落的工作（可參考 Figure 3.6）。

¹Model #params Language
bert-base-uncased 110M English

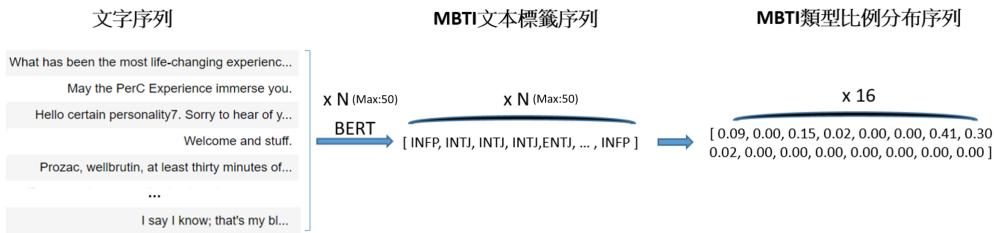


Figure 3.7: 特徵提取示意圖

3.2.3 特徵提取

為了在後續用於訓練對文字較不敏銳的基本方法，這裡選擇將原先的資料集進行改寫。我們透過前一個段落得到的資料集來微調 BERT 模型，並用於將每筆資料中的每一句話轉換成 MBTI 標籤後，再依照整筆資料的 MBTI 標籤占比，輸出成後續要使用的訓練資料集，詳細的說明在下一個段落。(可參考 Figure 3.5、3.7)。

3.2.4 將 MBTI 文本標籤序列轉換成比例分布序列

從 BERT 改寫後輸出的 MBTI 序列為文本的標籤序列，且根據原先一筆數據中有多少句話，序列就會有多少個元素，對於大部分都超過 40 個元素且長度並不固定的長序列而言，後續要進行辨識的模型在特徵提取上相較於固定長度的短序列肯定要更加困難，對此我們透過資料處理將文字序列彙整，透過統計來取得一個序列內 16 種 MBTI 的分布，再將這個分布轉化為數字套用在新的序列上，也就是說，原本的文字序列將變成乘載 MBTI 比例的數據序列，且由於 MBTI 的種類數目固定，因此序列的長度也會固定，這有益於後續模型進行特徵提取。

3.2.5 以 MBTI 比例序列進行分類模型的訓練

為了以更合適的特徵進一步提高 MBTI 的辨識的準確率，將選擇模型來為前述的 MBTI 比例序列進行後續的訓練，由於目前的訓練資料已經從原本的文字轉變成數據化的比例，並且在維持穩定的輸出之外還需要兼顧對話的即時性，因此這裡選擇的模型將針對數字的敏感性和輸出的效率來做選擇，在此我們選出以下三個方法作為實驗的代表。

3.2.5.1 SVM

SVM (Support Vector Machine, 又稱為支援向量機) 模型在高維度的數據空間中有出色的表現，且能夠透過找到最大間隔分離來提高整體泛化能力，在小樣本的任務中效果更加明顯。而對於本次實驗任務的多類別分類，SVM 模型採用的 kernel 為 rbf，對多分類任務有良好的適性，並且透過交叉驗證得到最佳的 C 為 0.5，C 代表懲罰參數或正則化參數，用於調整 SVM 模型的錯誤樣本容忍度和決策邊界的優化。

3.2.5.2 ELM

ELM (Extreme Learning Machines, 又稱為極限學習機) 模型由於不需要迭代優化，因此能快速訓練模型，且模型本身不用做太多超參數調整也不容易過擬合，相對而言是個具備良好效率的模型。訓練時同樣是透過交叉驗證，得到的最佳 hidden size 為 25，並透過 sigmoid 激勵函數來穩定輸出的準確性。

3.2.5.3 Random Forest

隨機森林演算法透過多個決策樹來選擇最終的結果，這讓其有良好的穩定性，並透過隨機特徵選擇來降低過擬合的風險。我們同樣透過交叉驗證來決定參數的設置包含了樹的數量、最大深度以及最小樣本的分裂樹，最終的是採用了 [10, 200, 2] 的組合來進行實驗。

3.2.6 GPT-3.5 turbo

GPT (Generative Pre-trained Transformer) 是一種透過大量文本訓練和深度學習，模仿人類對話方式的自然語言處理模型。在最近推出的版本 GPT-3.5 turbo[28] 中，新增了 prompt 來增強文本對模型的上下文關係。在這其中可分為 3 個腳色類別，user、assistant 和 system，user 表示使用者的對話，assistant 表示機器人的回覆，而 system 則表示特定的場域及對話資訊來幫助模型對對話的生成。因此對本研究而言，主要的目的在於將之前一系列處理後得到的使用者之對應 MBTI 人格輸入至 system 來做為使用者的對話人格，讓模型能以此作為回覆的參考，並根據 assistant 和 user 交織的對話記錄來進行生成。

3.3 實際使用

詳細的流程於 Figure 3.8、Figure 3.9，使用者在一開始使用時，會進行引導式問答，題庫來源是 MBTI-M[29] 版本的改編量表²，透過這個將引出使用者的初始人格，然後透過這個人格來賦予 GPT-3.5 turbo 模型進行生成，而隨著對話的增加，對話記錄也將回傳給

²(Version: 1998-N) developed by the psychology department at East Carolina University in the United States.

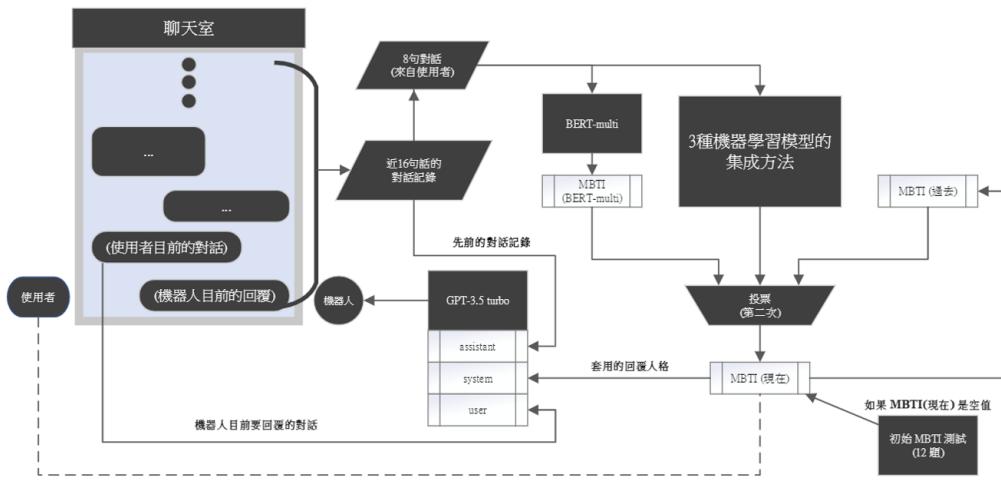


Figure 3.8: 系統實際運行架構流程圖

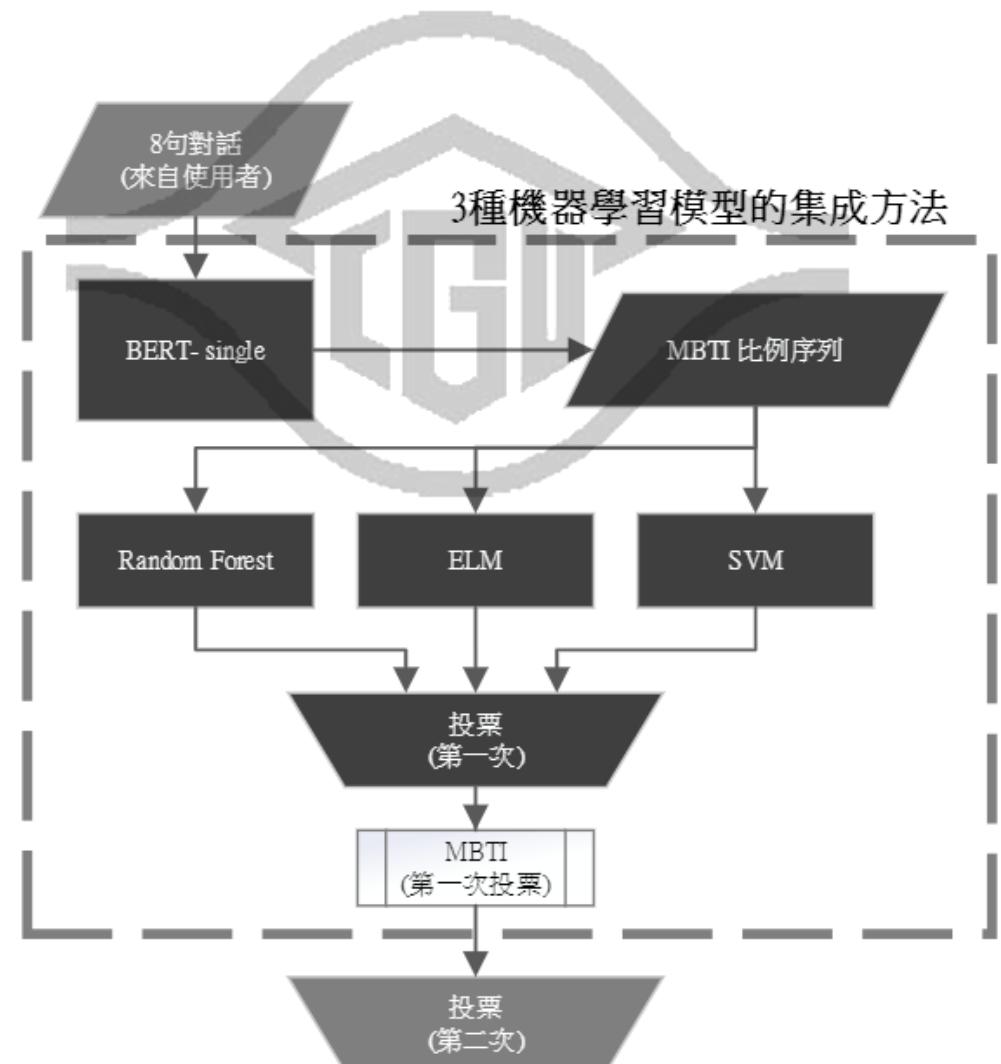


Figure 3.9: 3種機器學習模型的集成方法流程圖

GPT-3.5 turbo 和兩個 BERT(為了方便後續說明，將”一個序列多句對話”微調的 BERT 稱做 BERT-multi，”一個序列一句對話”微調的 BERT 稱做 BERT-single)，GPT-3.5 turbo 將擷取近 8 句對話來作為近期的對話記憶，成為回覆的示例。而兩個 BERT 會將分別這些對話進行辨識，BERT-multi 會一次性辨識所有的對話，然後再作預測。而 BERT-single 則會對每一句進行 MBTI 的分類任務 (可參考 Figure 3.9)，並轉換成 MBTI 的比例序列傳至後續的三個模型方法，SVM、ELM 以及 Random Forest 進行判別，三者判別出的成果會逐一進行比對，以 MBTI 的 4 個向度 E.I、S.N、T.F、J.P 分別進行投票，票高的指標將被保留，最終得到的將是三個模型相互評比的結果。再將此結果與 BERT-multi 的結果以及現階段的人格進行三方的綜合評比，此評比同樣是以 4 個向度來進行多數決投票，完成後的結果除了會更新原本的人格之外，也將作為嶄新的人格輸出至 GPT-3.5 turbo，供後續的對話輸出，期望透過不斷的短期對話記憶和長期的對話分析產生的對話人格，讓機器人有更多機會去理解使用者，創造更多符合本人取向的回覆。

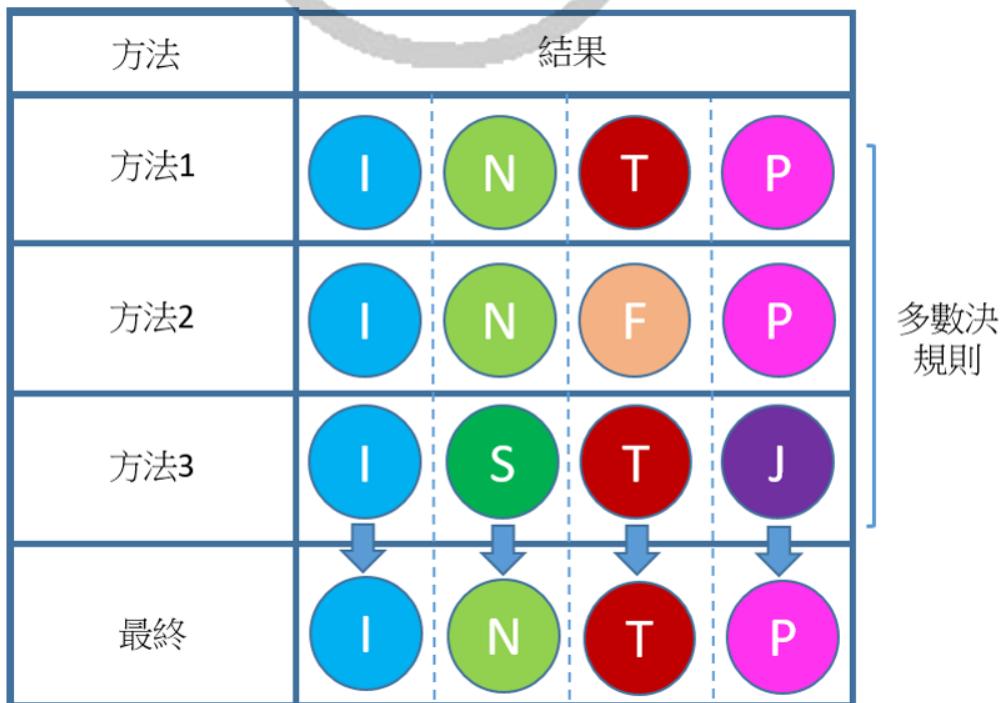


Figure 3.10: 多數決投票機制示意圖

3.3.1 多數決投票機制

可參考 Figure 3.10，在實際使用時，系統總共會進行兩次的多數決投票，第一次是 SVM、ELM 和 Random Forest 的結果投票，第二次則是 BERT、原本的人格記錄和第一次投票的結果投票，同樣都是以 MBTI 的 4 個向度分別投票並將最終結果組合。而兩次採用多數決投票進行決策，其實有不同的考量定位。第一次的使用主要的作用在於能夠提高整體系統的穩定性，由於每個模型在判斷上都有自己的規則，因此在不同的樣本上容易有不同的表現，為此我們期望通過多數表決，讓整體選擇更正確的類別，不會過度依賴單一方法而造成不確定性。而第二次的投票是希望能以不同的角度來判斷使用者的 MBTI 人格，儘管 BERT-multi 和三個基本方法的組合都是針對使用者對話，但不同的是前者是剖析了所有對話的語意後再進行辨別，後者則是根據每一句對話偏好的人格統計來判斷。除了兩個使用模型方法之外，採用使用者原先的 MBTI 人格一同做判斷也是本次研究的考量，其原因在於即使 MBTI 人格容易受到互動的影響而改變 4 個字母的排列，但就如原作者 Carl Jung [30] 所述：「每個人會運用不同的態度、認知、判斷和執行方式，但都會偏重其中幾種，就如同我們兩隻手都能使用，但隨時間進展也會有偏好的慣用手一樣。」因此我們認為連同使用者過去的 MBTI 人格一同判斷才不會去忽略掉這個重要的特性。

3.4 評估指標

評估的指標包含了模型性能以及系統實測兩者，前者使用的指標分別為 Accuracy、Precision、Recall、F1 Score 以及 MBTI4 個向度的準確率，用於和相關的 baseline 進行比較。後者會透過使用者問卷以

及同類型測驗結果比較來驗證整體系統的成效及泛化程度。

3.4.1 模型性能

以下為本實驗用於計算模型性能的五個方法。

3.4.1.1 Accuracy

準確度是模型正確分類的樣本數與總樣本數之比，體現了模型在所有 MBTI 人格類別上的總體性能。具體的公式如下：

$$Accuracy = \frac{Correct\ Predictions}{Total\ Predictions} * 100\%$$

3.4.1.2 Precision

精確度表示模型在預測為某一類別時，實際屬於該類別的比例，體現了模型的準確性。具體公式如下：

$$Precision = \frac{TruePositives}{(TruePositives + FalsePositives)} * 100\%$$

3.4.1.3 Recall

召回率表示模型在所有實際屬於某一類別的樣本中，成功預測為該類別的比例，體現了模型的覆蓋程度。具體公式如下：

$$Recall = \frac{TruePositives}{(FalseNegatives + TruePositives)} * 100\%$$

3.4.1.4 F1 score

F1 分數是精確度和召回率的調和平均值，它綜合考慮了模型的準確性和覆蓋程度。在類別分布不平衡的情況下，能夠去平衡 Precision 和 Recall 的權衡。具體公式如下：

$$F1Score = 2 * Precision * Recall / (Precision + Recall) * 100\%$$

3.4.1.5 MBTI 的 4 個向度準確率

透過外向 (E) 與內向 (I)、感覺 (S) 與直覺 (N)、思考 (T) 與感覺 (F)、判斷 (J) 與知覺 (P) 等 4 個向度來作為評斷基準，除了藉由每個向度代表不同性格特質來提供更精細的分析之外，在特定的情況下，特定向度的準確率可以更直接的去反應受測者在某個面向的價值，以及更好地去診斷模型的弱點，並讓研究者能針對此來進行改良。

3.4.2 系統實測

以下為本實驗用於展現系統實測成果的兩個方法。

3.4.2.1 使用者問卷

讓使用者分別以有 MBTI 指標作為參考資訊的機器人和沒有的兩者互相比較，再依語意、體感差別將結果分為 5 個等級，於問卷上讓使用者選擇。實測時會先以沒有套用 MBTI 人格的機器人先和使用者完成 8 次的對話，之後再套用 MBTI 人格並清除對話記錄，以同樣的對話進行一次，讓使用者比較前後差異 (測驗中並不會告知使用者前後為非套用及套用 MBTI 人格的機器人)。語意的比較重點在於兩者之間生成的結果是否有詞不達意、情境理解錯誤、和使用者

對答偏題等狀況；而體感的部分比較的重點在於兩者之間的結果在語氣和用字上是否讓使用者更加舒適，更符合期望。

3.4.2.2 同類型測驗結果比較

請使用者先至 16Personality[31] 進行測驗，再用本系統進行 1 次的初始人格測驗和 2 個輪次（共 16 句）的對話，最終產生的結果兩者之間互相比較，以查看準確度。



第四章 實驗結果與討論

本章節將闡述本次實驗的研究成果，主要可分為兩個部分，第一部分為模型方法的性能比較，會將本實驗提出的方法¹與各個參考文獻中的 baseline 之結果進行對比。第二部分則是運用於對話機器人上的效用，我們會請受測者實際使用我們設計的聊天機器人系統，在體驗完成後根據整個流程的感受以及測得的 MBTI 人格準確性來作為評估的手段。值得一提的是，在第一部分的測試中，根據參考文獻 ([8]) 的結果評比，針對來自 kaggle 的這個資料集，區分成了 kaggle 和 kaggle-filter 的版本，而本實驗只以 kaggle-filter 作為依據，原因在於本實驗的成果將套用於實際的對話機器人上，訓練用的對話資料必須盡可能的符合普通人之間的對話標準，而 kaggle-filter 相較於 kaggle 濾除了會影響的要素，因此本實驗使用的資料集將會採用過濾完成的版本，而不是最原始的版本，為的就是不波及後續對話機器人測試時的公平性。

4.1 實驗環境

本研究的硬體實驗環境說明如 Table 4.1 所示，軟體實驗環境說明如 Table 4.2 所示，實驗中重要的函示庫說明如 Table 4.3 所示。

¹ 實驗內容中出現 BERT-Combo 表示是指以 BERT 做特徵提取後訓練的 SVM+ELM+Random Forest 之集成組合。

Table 4.1: 硬體實驗環境說明表

硬體項目	說明
CPU	12th Gen Intel(R) Core(TM) i7 - 12700
RAM	16GB
GPU	NVIDIA GeForce RTX 2080 Ti
OS	Windows 10

Table 4.2: 軟體實驗環境說明表

軟體項目	說明
Python	3.10
Jupyter Notebook	6.5.2
Conda	23.1.0
CUDA	11.8
GPU Driver	536.23

4.2 模型性能之結果展示

Table 4.4以各項評估指標來呈現性能，從中可以發現，不論是 SVM、ELM、Random Forest，還是它們的集成 BERT-Combo，都取得了非常出色的成績。與之相對的 BERT-multi 表現並不突出，儘管是透過複數句話來進行判斷，能獲取更多元的對話特性，但仍舊落於人後。不過值得注意的是，三個機器學習的方法有如此高的準確度，和它們的訓練序列有很大的關聯，我們分析了數筆預測正確資料，發現在 MBTI 比例序列中，最大的數字都是這筆資料的 MBTI type，因此它們就能很輕易的判斷出答案，而為什麼這個現象會出現在絕大部分的測試資料呢？我想這就我們使用的原始資料來源有關了，由於我們的資料來源是 PersonalityCafe[24] 蒐集而來的貼文，在這個論壇中的使用者聊天的主題都是以個性、人格作為啟發，因此

Table 4.3: 實驗中重要的函式庫說明表

函示庫項目	說明
torch	1.13.1+cu117
transformers	4.24.0
scikit-learn	1.2.1
imbalanced-learn	0.10.1

Table 4.4: 實驗中的方法之 Accuracy、Precision、Recall 和 F1 Score 的分數

Method	Metrics			
	Accuracy(%)	Precision(%)	Recall(%)	F1 score(%)
BERT-multi	66.25	66	66	65
SVM	97.84	98	98	98
ELM	98.46	98	98	98
Random Forest	98.92	99	99	99
BERT-Combo	98.76	99	99	99

Table 4.5: 實驗中的方法之結果以四個向度 (E.I、S.N、T.F、J.P) 的準確率

Method	Metrics			
	E/I(%)	S/N(%)	T/F(%)	J/P(%)
BERT-multi	88.38	90.92	87.24	81.22
SVM	97.76	99.00	98.61	99.07
ELM	99.23	99.07	99.23	99.30
Random Forest	99.76	99.61	99.53	99.38
BERT-Combo	99.71	99.38	99.79	99.42

絕大部分的對話都具有明顯的人格傾向，這也就導致了 BERT-single 能夠簡單的將對話轉成對應的 MBTI 人格，讓序列不斷累積正確人格的數量，以最高比例的 type 讓機器學習得以辨識。而這樣的結論，也讓人擔心是否會有泛化性的問題，對此我們將在後續的實際測試中看到結果。

Table 4.5則拆分成四個不同向度，更細緻的察看每個方法對於 MBTI 人格指標的感度。和 Table 4.4的狀況相同，機器學習的三個方法以及集成的 BERT-Combo 都取得了非常高的準確度，其原因就如前一段所言，這裡就不多加贅述。而值得關注的是，BERT 的 J.P 指標表現都不如其他三個指標，這也和參考文獻中 ([8]) 提到的“J.P 的預測相較其他指標更加困難”相互吻合。在參考文獻 ([4]) 中也表示 J.P 的預測困難可能不只是涉及了文字的意義，還跟人們的行動和行為有密切關聯，因此或許針對這個部分來選擇合適特徵，預測的結

Table 4.6: 和相關研究中同樣使用 Kaggle 資料集並且將成果分成四個向度 (E.I、S.N、T.F、J.P) 來比較表 [1][2][3][4][5][6][7]

Method	From	Metrics				
		E/I(%)	S/N(%)	T/F(%)	J/P(%)	Average
LSTM	Cui & Qi, 2018[5]	89.51	89.85	69.10	67.65	79.03
SVM	Bharadwaj, 2018[1]	84.90	88.40	87.00	78.80	84.78
K Nearest Neighbor	Li, 2018[4]	90.00	90.00	91.25	76.25	86.88
BERT + MLP	Mehta, 2020[7]	78.80	86.30	76.10	67.20	77.10
RoBERTa - Distil	Vásquez & Ochoa-Luna, 2021[6]	88.46	92.68	88.46	84.91	88.63
XGBoost	Amirhosseini & Kazemian, 2020[2]	78.17	86.06	71.78	65.70	75.43
SimCSE-BERT/RoBERTa	Fernau & Hillmann, 2022[3]	69.17	76.44	63.94	63.40	68.24
BERT-multi	-	88.38	90.92	87.24	81.22	86.94
BERT-Combo	-	99.71	99.38	99.79	99.42	99.50

果才會有所成效。

4.3 與現有方法之數據比較

由 Table 4.6的結果中可以看到，在方法的表格內有許多條分隔線相隔，主要的用意在於雖然所有的方法採用的資料集都是來自 Kaggle，但第一條分隔線以上是使用非濾除過的資料集去做處理，而使用這樣的資料集是會對結果造成影響的，原因在於非過濾資料集中的對話內容有殘存著不少的 MBTI 指標文字，這會使得模型在辨識時直接被這些文字影響而更容易得到答案，導致驗證失去公正性，因此可以看到在在過往的方法中(前兩大格)，上方數據明顯較下方更加優秀。而我們實驗採用的數據集是過濾後的數據(可參考

Table 4.7: 和參考文獻 [8] 中使用已過濾資料集進行訓練的方法比較 J/P 之準確率

Method	From	J/P Accuracy (%)
Complement Naïve Bayes		61.22
LightGBM		66.26
Logistic regression	Choong EJ, Varathan KD. 2021. '[8]'	59.28
Random Forest		62.01
Support Vector Machine		64.04
BERT-multi	-	81.22
BERT-Combo		99.42

Method 中的 data cleaning)，因此比較的對象應該以過往方法下方的對象為主，而從這樣的結果來看，我們的方法不論是 BERT-multi 還是 BERT-Combo 都比以往的方法更加優秀，即使從四個向度分開觀察，我們的方法相對於原本表現較好的參考文獻 ([2]) 的結果都要有所提升。

在 Table 4.7 中我們比較了參考文獻 ([8]) 裡嘗試的 5 種方法，主要是因為它有註明使用了已過濾的 Kaggle 資料集，並且比較的指標選擇了浮動差別較大的 J.P 準確率。從結果可以看出，同樣都是使用機器學習模型的方法來實作，我們的方法取得了更好的成果，我們認為關鍵除了以合適的資料特徵來處理之外，透過三個方法的集成能夠以更廣泛的角度來搜索文句之間的關聯。

4.4 實際測試之結果展示

本次測試總共邀請到了 25 位受測者，基本流程如下：

1. 受測者至 16Personality 網站測試。
2. 受測者進入系統，先進行 12 題的 MBTI 測試得到初始人格。
3. 使用者和未套用 MBTI 人格的機器人進行 8 句話的對話。
4. 完成後將機器人套用 MBTI 人格，以同樣的 8 句話再與機器人對

話一次，並比較前後差異。

5. 完成後會得到新的 MBTI 人格，然後再與機器人進行 8 句的對話。
6. 完成後會得到最終的 MBTI 人格，之後就根據表單內的問題進行填寫。

(附註：使用者於步驟 3、4 階段時是不知道機器人是否套用 MBTI 人格的，只知道其中一次有套用，另一次沒套用，而在我們的實驗預設中，是設定前一次為無套用，後一次為有套用的。在後續的使用者問卷中將透過比較並選擇前、後結果來進行答覆。我們會這麼做的目的就是為了實踐雙盲測試，以確立實驗的公正性。)

表單內大致分為兩部份，第一部分記錄了 16Personality 網站的結果以及使用對話機器人最終得到的結果，會以前者作為受測者的標準 MBTI 人格去與後者進行對比。第二部分則是先讓受測者依照第一輪對話前、後兩次的區別來選擇哪一個機器人（無套用及有套用 MBTI 人格）的回覆更好，接著再讓其根據語意以及感受上的不同來作評分，分數可分為 1-5，1、2 表示變差，而 1 變差的程度更甚於 2，3 表示沒有差別，4、5 表示變好，而 5 變好的程度更甚於 4。

4.4.1 MBTI 人格測定的準確度結果

Figure 4.1 為受測者結果於四個向度指標 (I/E, S/N, T/F, J/P) 中的正確數量占比。從結果中可以看出，在 4 個向度上，我們的方法對應到由 16Personality 測定的結果，兩者有高度的相合，在 S.N 有 76% 的準確率，E.I 和 J.P 更是有 84% 的準確率，唯一表現比較不好的 T.F 則是 52% 的準確率，我們分析了在 T.F 中兩者不同的案例，其中發現大多數都是 F 被判為 T，和 T 被判成 F 相比，比率是 3:1。F 和 T 分別代表感性和理性，通常 F 傾向的人在處理事情時會更直接的去思考自己的情緒、感受，以及對其他事物的影響，而 T 傾向的

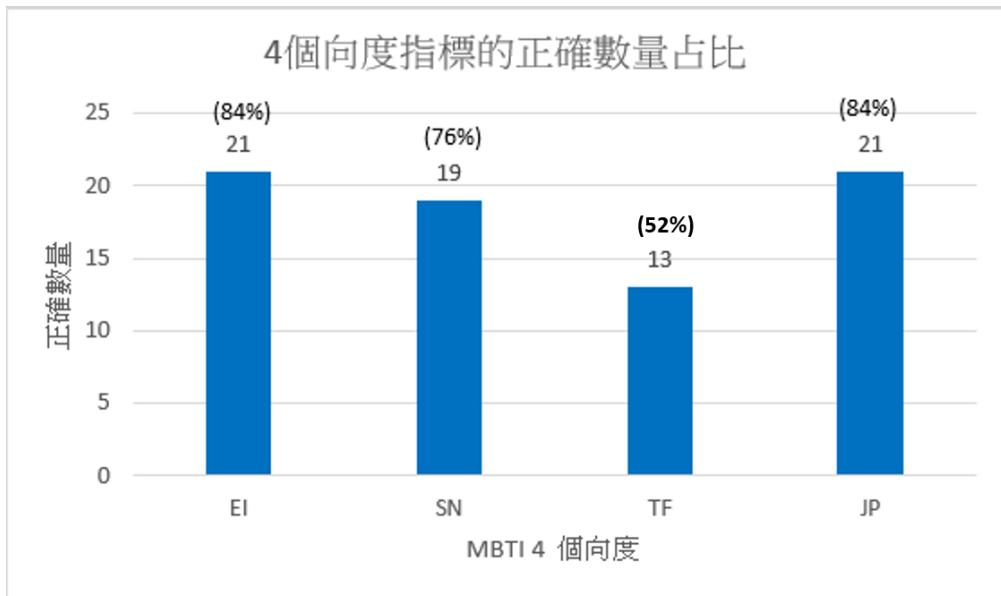


Figure 4.1: 受測者結果於四個向度 (E.I、S.N、T.F、J.P) 指標中的正確數量占比圖

人則是會以事情的運作原理為準則，並且根據邏輯來進行決策。我們進一步觀察了這些被判為 T 傾向的受測者的對話，發現有個共通的關鍵在於他們總是以問題來作為回覆的對話，而這些問題中往往只包含了他們對某件事情的疑問，卻沒有他們自身對這件事情的看法或感受，這導致我們的模型很容易將它判為 T 傾向。與幾位在 T.F 判別錯誤的受測者在測試後訪談，他們表示由於和機器人的對答中，機器人給出的回答太過飽滿，導致他們很難再回到剛才的話題去給出自己的想法，而是會自然地去想另一個問題來提問，而一旦陷入這個對話循環，就很難跳脫出去，系統也就容易給出 T 傾向的判定。當然我們也有觀察 F 傾向被判正確的案例，我們發現即使他們在提問時，也會在問題前後表達自己的看法，這個動作就能增加模型判別為 F 傾向的可能性。

4.4.2 前後選擇的結果 & 在對話中語意及感受上的評分

Figure 4.2 表示了使用者在第一輪前、後對話中，認為哪一邊更符合自身偏好的數量占比，其中有 80% 的受測者認為後一次的回覆更

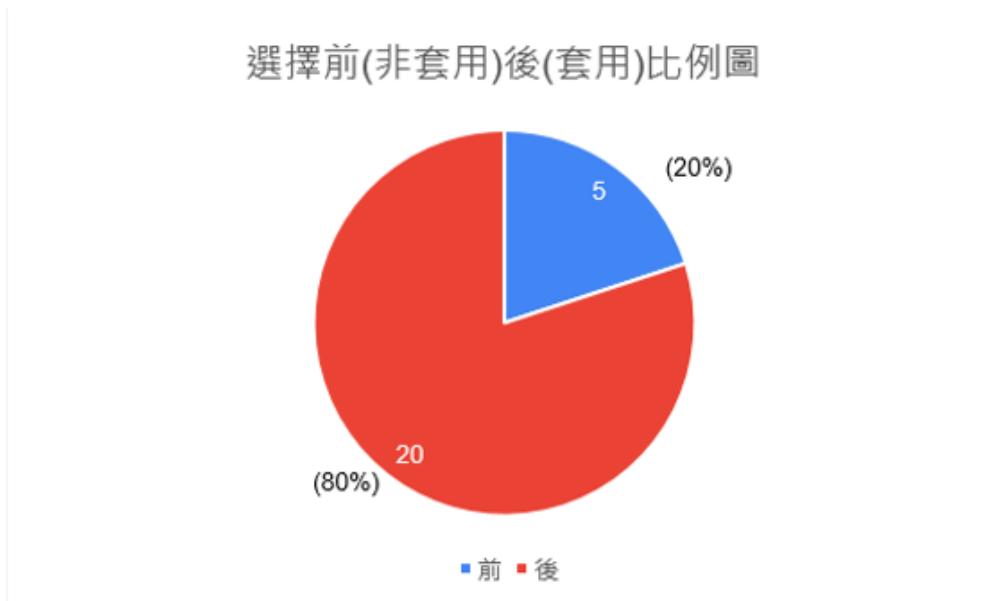


Figure 4.2: 受測者根據第一輪前、後對話選擇的占比圖

好，而僅有 20% 的受測者選擇前一次的回覆，這也就說明了套用了 MBTI 人格的聊天機器人更適用於多數人在聊天上的期許。後續也將讓使用者以語意及感受兩個不同面向來更細緻地闡述對於套用與非套用 MBTI 人格的差別。

Figure 4.3為受測者在對話中語意(上)及感受(下)的評分，分為 1-5 分，越高表示越好。從結果可以看出，不論是在語意或是在感受上的評分，正面的評價都是高於負面，這顯示著套用 MBTI 人格在對話上是具有正向影響的。從語意方面的數據中可得知，平均給分為 4 分，標準差為 0.94，正面的評價 (4+5) 共佔了 64% 的比重，超過了負面以及中立的評價，和幾位給出了正面評價的受測者訪談，他們表示在套用 MBTI 人格後給出的對話更加充實，能從這些內容中得到更多的資訊，更符合自己內心期望的結果，也讓他們更明確地去銜接前一句的對話。而在感受方面的數據中可得知，平均給分為 3.96 分，標準差為 0.96，正面的評價 (4+5) 共佔了 76% 的比重，同樣的也超過了負面以及中立的評價，不過負面評價 (1+2) 的佔比來到 12%，於是我們詢問了給出負面評價的受測者他們的想法，有人認為在套用 MBTI 人格後的機器人給出的回覆過於詳細，但在他的認

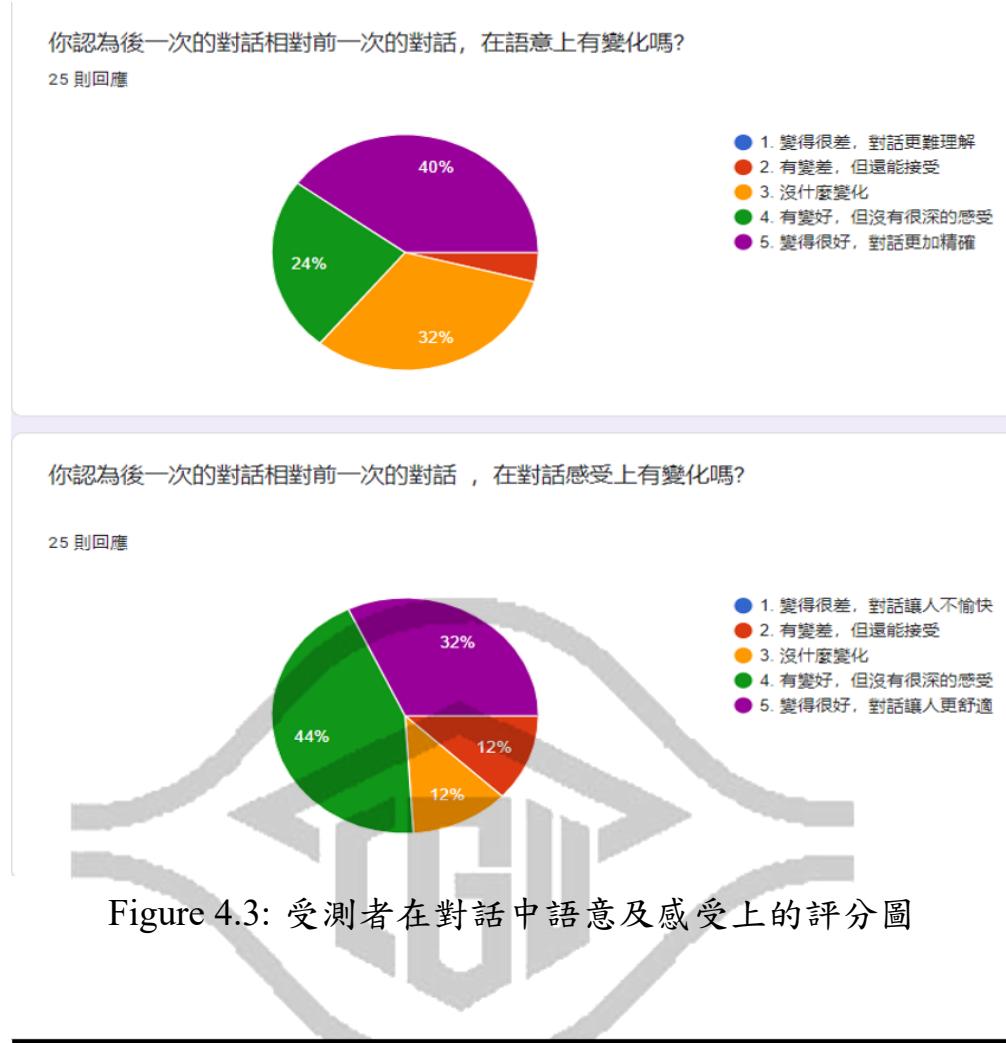


Figure 4.3: 受測者在對話中語意及感受上的評分圖

ISFP	交往好累，我想分手!
Normal	或許可以好好談談，看看是否有其他解決方式。
Apply personality	我理解你的感受。分手是個艱難的決定，但有時候是必要的。請勇敢地追求自己的幸福。
ISTJ	當今天有個跨年烤肉活動，裡面有我討厭的人，我該參加嗎？
Normal	這是你自己的決定，但如果你覺得和討厭的人一起會讓你不舒服或影響到活動的愉快，或許考慮另尋一個舒適的方式度過跨年。
Apply personality	如果你覺得討厭的人會影響你的心情和活動的樂趣，或者會給你帶來壓力，你可以考慮不參加。重要的是保護自己的舒適度和享受活動的樂趣。
ENFP	我覺得眉毛連載一起的人最有男子氣概了。
Normal	每個人對於男子氣概的定義都不同，但眉毛的形狀確實可以給人留下深刻的印象。每個人都有自己獨特的魅力和特點！
Apply personality	男子氣概並不僅限於眉毛的形狀或外貌特徵，更重要的是內在的品格和行為。男子氣概體現在自信、堅強、負責的態度，以及對正義和真理的追求。外在的特徵只是冰山一角，重要的是發揮你的內在價值和魅力。

Figure 4.4: 三位受測者在與套用與否 MBTI 人格的機器人聊天內容範例

知上，聊天對話不應該是這麼龐大的字數，這會使得他們覺得對話頗有壓力，也造就了他們認為在感受上不如沒套用 MBTI 人格的主因。同樣地我們也和幾位給出正向評價的受測者訪談，他們表示和沒有套用 MBTI 的對話相比，套用的明顯感受到對話上的溫度，整體的對話感受程度有明顯的提升，給出的回覆更能讓人接受，不論是在語氣或是用字上。整體來說，我們認為在語意及感受兩個方面都有這樣正面的影響正是 MBTI 人格給出的提示去調整了生成模型的輸出，讓產出的對話對受測者而言更能被接受，除了更加精確，增進對話的理解能力之外，也更加的舒適，對應個人的對話偏好。Figure 4.4 中記錄了三位不同的受測者對話時，機器人在套用 (Apply personality) 與不套用 (Normal) MBTI 人格之間所做出的回應，從中可以觀察出兩者之間給人的差異。

4.5 結果討論

本節將深入剖析實驗的整體過程。這包括對實驗策略所引起的變化進行分析，以驗證其對結果的具體影響。此外，我們也同樣關注實驗過程中所產生的疑問與不同的觀點，為了能夠釐清實驗結果，我們將提供更全面透徹的研究解讀，來做為探討的根基。

4.5.1 初始測試的 12 題問答的效果以及準確度

12 題的初始測驗是為了在一開始完全沒有使用者資訊時，能夠先一步取得的手段。測驗的每一題都是二選一的題目 (可參考 Figure 4.5)，能讓使用者快速地完成並進行後續的機器人對話。而這樣的測試就實際結果來說究竟對本實驗有多大的影響，可以看到 Figure 4.6 顯示了第一輪對話第二階段投票的三個方法在 4 個向度的正確數

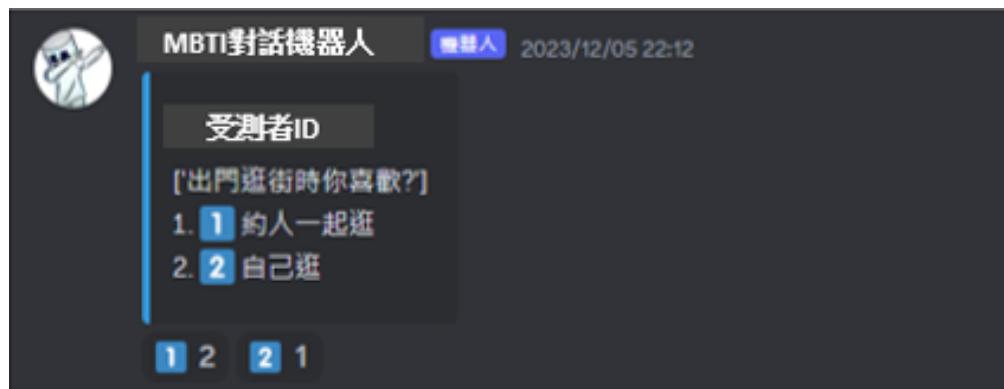


Figure 4.5: 初始測驗 12 題的題目測試示意圖

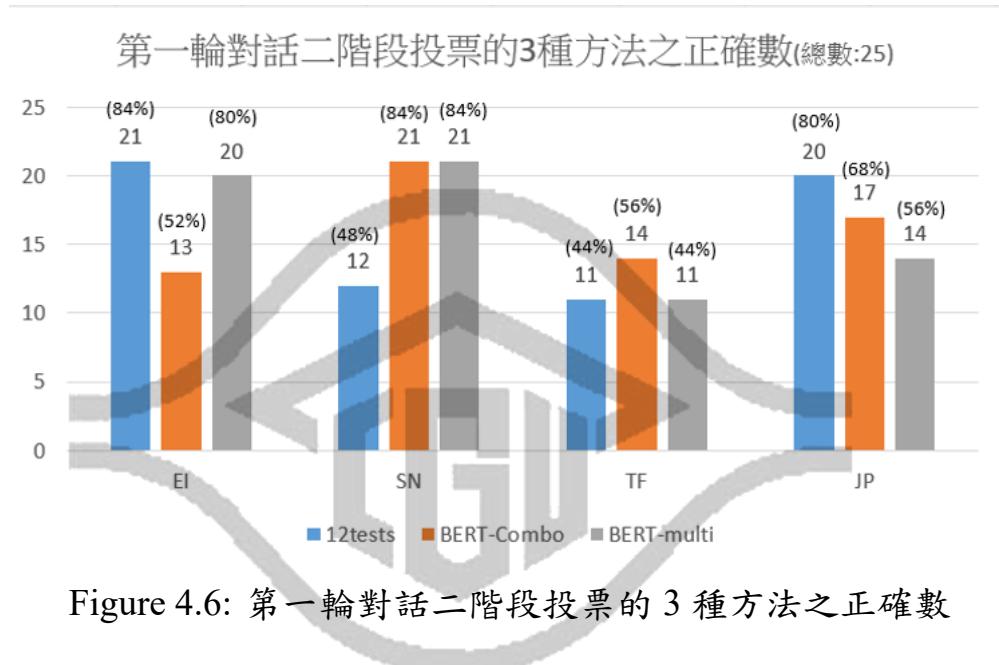


Figure 4.6: 第一輪對話二階段投票的 3 種方法之正確數

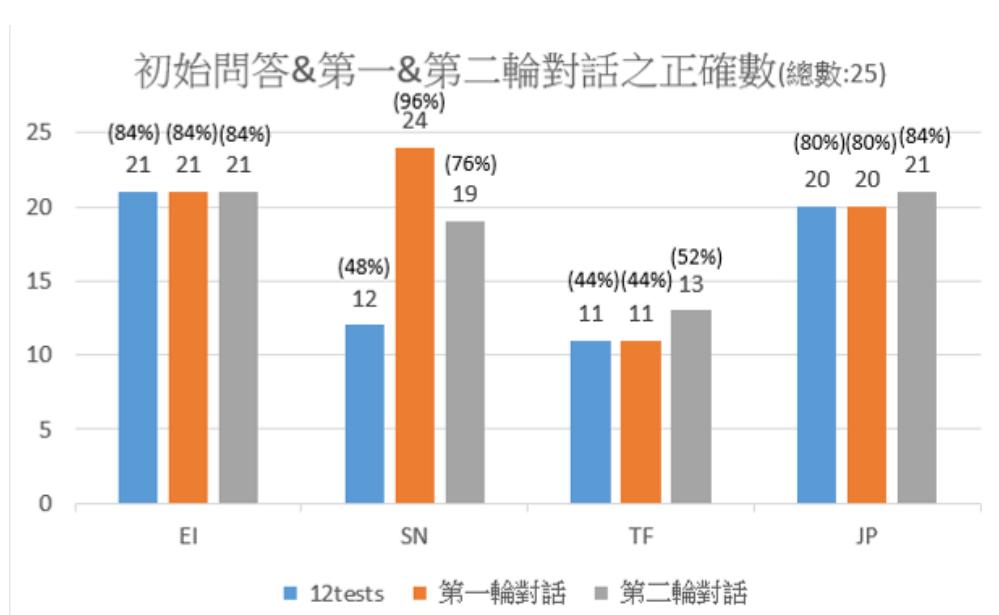


Figure 4.7: 初始問答 & 第一 & 第二輪對話之正確數

Table 4.8: 以實際測試的結果展示本研究使用方法於四個向度 (E.I、S.N、T.F、J.P) 的準確度

Method	Metrics			
	E/I(%)	S/N(%)	T/F(%)	J/P(%)
SVM	64	66	60	56
ELM	60	74	48	70
Random Forest	60	74	54	64
BERT-Combo	62	74	58	64
BERT-multi	76	76	40	64

量，當中可以看到 12 題初始測驗在 E.I 以及 J.P 的表現非常的好，分別有 84% 和 80% 的準確率，相比於另外兩個方法，在這兩個向度上能更好地讓投票結果符合期待。另外我們也提供了第一論和第二輪對話投票後的正確數量圖 (可參考 Figure 4.7)，可以從中觀察兩個輪次的對話相對於初始問答在 4 個向度上的成效。

4.5.2 為何要以 BERT-multi 和 BERT-Combo 的結果來做最終的集成投票

應該有讀者會有所疑問，為何在 Result 看到 SVM、ELM、Random Forest 的準確率明顯高於 BERT-multi 還要高，但最終 BERT-multi 却是要跟三者集成之後的 BERT-Combo 來做投票，會這麼做的原因在於，首先 BERT-multi 和其他方法的訓練資料集就不同，BERT-multi 的方式是以所有的對話來做為輸入，以全局的角度去做 MBTI 的預測，而其他的方法則是以每一句的對話所對應的 MBTI 人格轉變成比例序列的形式來輸入，相對前者而言更加關注句子本身字詞，而忽略了句子之間的關聯。從 Table 4.8可以看到，就結果而言，除了 BERT-multi 之外的所有方法，相比於 Result 的準確度都要下降許多，這表示以單句序列作為訓練來源的方法在泛化程度上並不理想。儘管如此，這些方法在實測時的表現也還是有可取之處，BERT-Combo 作為 3 個方法集成的結果，在 4 個指標上都有穩定的

發揮，和 BERT-multi 相比更可以發現，BERT-Combo 在 T.F 上的表現優於 BERT-multi 許多，而 BERT-multi 則是在 E.I 表現上相對出色，這顯示著這兩個方法存在著互補性，透過不同的角度去分析使用者的文句，得到更全面且準確的結果，這是我們認為要採用這兩個方法來作為最終集成的主要原因。

4.5.3 第一階段投票選擇 3 個方法的論點探討

在第一階段的投票中，我們選用了 3 個機器學習模型的方法來做為投票的角色，這麼做的目的除了是 3 個模型在性能上的表現符合預期之外，取 3 個模型來投票更能夠實現相對多數決的投票，不會出現打平的狀況。而對此可能會有人提出質疑，如果以更多的方法來參與投票並且賦予各個模型投票權重，那是否會比現階段的成效還要好。就我們的角度而言，首先我們認為選擇多個模型來進行投票是可行的，只是模型之間的性能不能差距過大，若差距過大就會有失水準。而如果要賦予權重來避免打平的狀況，權重的訂定規則又會是一個無法界定的標準，且這種以對話來進行辨識的任務，資料之間的變化很大，如果胡亂設置權重很可能會影響模型在不同場域要素上的表現，要評估的話將會有太多的不確定性因素，因此我們認為權重的設定應該盡可能避免，若真的有需要則要詳細的分析與記錄，以確立整體的成效。

4.6 實驗嘗試

本節將介紹在研究過程中所進行的各項實驗性探索。儘管這些嘗試最終未被納入主要結果，但它們對於研究方向的探索與調整起了關鍵作用。透過詳細記錄這些過程，不僅能夠為本研究的完整性增

Table 4.9: 以 BERT-multi 改寫資料集後訓練的方法之 Accuracy、Precision、Recall 和 F1 Score 的分數

Method	Format of training data	Metrics			
		Accuracy(%)	Precision(%)	Recall(%)	F1 score(%)
ANN		26.26	15	26	17
SVM+ANN	MBTI label sequence	38.52	28	39	30
LSTM		25.58	13	26	15
ANN		35.02	23	35	25
XGBoost		42.08	38	42	39
LightGBM		42.24	38	42	39
SVM	MBTI distribution sequence	43.24	37	43	39
ELM		43.47	40	43	40
Random Forest		42.32	36	42	39
BERT-Combo		43.55	37	44	40

Table 4.10: 以 BERT-multi 改寫資料集後訓練的方法以四個向度 (E.I、S.N、T.F、J.P) 的準確率

Method	Format of training data	Metrics			
		E/I(%)	S/N(%)	T/F(%)	J/P(%)
ANN		78.64	87.48	57.42	61.90
SVM+ANN	MBTI label sequence	78.95	87.48	78.65	66.12
LSTM		77.69	86.23	56.74	55.17
ANN		78.76	87.48	76.58	60.25
XGBoost		77.88	87.86	79.34	67.43
LightGBM		78.11	87.94	79.19	67.67
SVM	MBTI distribution sequence	78.73	87.86	79.57	68.74
ELM		79.80	88.40	78.11	69.43
Random Forest		78.80	87.71	79.26	67.74
BERT-Combo		78.88	87.94	79.26	69.05

添一筆，也能提供未來研究中更加全面的理解。

4.6.1 以 BERT-multi 改寫資料集後訓練的模型

原先我們是以 BERT-multi 來做現在 BERT-single 的工作，也就是將資料集轉成 MBTI 序列的任務。從 Figure 4.9、4.10 可以分別查看它們的成果，和現在使用的方法之準確率等性能相比能看出明顯的差別，這也表明了任務的輸入資料形式對於作用模型在性能上的影

響是巨大的。雖然當初的表現就現在來說並不理想，但我們仍然對其做出許多的嘗試來改變其性能，以下的段落就是以這些方法進行改革後得到的回饋與心得，這些數據對於幫助我們找到目前的最佳方法來說功不可沒。

4.6.2 資料集依 n-gram 重組

在原先的實驗中，BERT 模型輸出的 MBTI 序列並不是以比例的方式呈現，而是一個接一個的 MBTI 指標，當時是直接將這些序列送入後續的模型來進行訓練，不過訓練出來的結果並不理想，於是我們採用了 n-gram 的方式來改進資料集，期望透過增進指標之間的關聯性，讓模型更容易捕捉，以取得更好的成果。不過測試後發現效果都較原本的更差，以當時用來測試的 LSTM 模型為例，沒有啟用 n-gram 的平均準確率約在 25% 左右，而啟用後，最優秀的準確率為 21.34%，且隨著 n 值變小，準確率也隨之下降。從這個實驗結果來說，我們認為 n-gram 並沒有辦法激發出 MBTI 指標與指標之間的關聯性，更可能創造出了多餘的噪音來影響模型的判斷和輸出。

4.6.3 轉換為 MBTI 比例序列之差別

如前段所述，原先的訓練資料是一個個的 MBTI 指標，因為結果不理想而採用 n-gram 進行修改，但仍然未解，於是我們剖析了資料集的特性，發現其中的語句之間並沒有連貫性，對於時序性較敏感的 LSTM 模型可能會發揮不良，因此我們將 LSTM 模型換成 ANN 模型，並且思考既然資料之間不存在連貫，那是否有方法能簡化資料及讓模型能更好的去擷取特徵，最終我們將 MBTI 序列轉化成了現在使用的比例序列，從更改後的效果來看，不只將原先的序列由字符串轉成數字，更重要的是將原先最長可達 57 個單位長度的序列縮減

為全體一致的 16 個單位，這樣的資料特徵改變也將體現於成果上，原先 ANN 在使用非比例序列時的結果為 26.26%，而換成比例序列則一口氣來到了 35.02%，4 個指標中成長最多的是 T.F，從 57.42% 上升到 76.58%，這樣的進步顯示著特徵優化對於模型訓練有著不可忽視的成效。



Figure 4.8: 刪除人格指標的兩個方法示意圖

4.6.4 在 MBTI 比例序列中刪除某些人格指標

刪除的任務主要目的在於讓資料集透過減少被視為噪音的資料，讓模型能夠將特徵提取集中在被保留的資料上。刪除的任務分為兩種(可參考 Figure 4.8)，一種是將 MBTI 指標依照 4 個向度(E.I、S.N、T.F、J.P)來拆分，如果這個指標跟目標指標的 4 個向度都是不同的，則刪除掉這個指標的比例。另一種則是倘若某個指標低於整體平均值($1/16$)，就進行刪除。以 SVM 模型進行測試，前者測得的準確率為 42.63%，後者為 43.16%，而完全不使用刪除規則的 SVM 模型為 43.24%，就實驗結果來說，三者的差距不到 1%，非常的小，不過最優好的結果仍為不使用刪除規則，推測原因在於刪除的這個行為並不會讓模型更好的去聚焦特徵，反而會破壞資料之間的關聯，而刪除規則的後者優於前者，我們認為是後者的規則刪除的比例實在少，跟原本的資料集相比改變不多，所以測得的結果才會較前者的更好。

第五章 結果與未來展望

本章節將統整研究的始末，集各個部份的重點於此闡述。簡而言之，本研究探索了將 MBTI 人格標籤與機器人對話生成結合的新領域。我們不僅採用了深度學習模型以及傳統機器學習模型，更以集成技術結合兩者之精華，以期達到更精確的人格分析，來獲得更具個人色彩的生成回覆。以下將概述我們的研究發現和對未來發展的啟示。

5.1 結論

本研究主旨 在於將 MBTI 人格標籤套用於機器人的聊天回覆生成任務，我們期望透過訓練的模型來對使用者的對話進行辨識，分析出最符合它的 MBTI 人格，再將這個標籤作為於機器人的生成回覆提示，來測試成效。在模型訓練方面，和以往 baseline 不同的地方在於我們同時採用了深度學習模型的 BERT 以及基本的機器學習模型 SVM、ELM 和 Random Forest。我們在完成資料預處理工作後，分別使用不同資料形式微調了兩個 BERT 模型，稱作為 BERT-multi 和 BERT-single。之後我們以 BERT-single 將文本資料中的每一句話轉化成一個 MBTI 標籤，最後將整個序列轉換成 MBTI 比例序列，並分別給三個基本的機器學習模型進行訓練。在聊天系統執行上，我們以 GPT-3.5 turbo 作為機器人的對話生成模型，並將 MBTI 標籤當作

對話提示，系統運作期間會蒐集使用者與機器人之間的對話，在累積一定的數量後交由模型進行辨識，這裡我們會透過兩次的集成投票來決定最終的 MBTI 人格，第一次是由三個基本模型投票的結果，第二次則是由第一次的結果和 BERT-multi 辨識的結果以及使用者當前的 MBTI 人格三者進行投票的結果。

在評測的部分，我們提出的方法和過去的 baseline 相比取得了較好的成績，證實了在 MBTI 分類任務上我們有更好的性能。而實際測試的效果也同樣出色，不論是在辨識的準確度上，亦或是使用者對機器人回覆的感想，這也代表著分析並套用 MBTI 人格的機器人在對話上是有所成效的。

5.2 未來展望

經由本研究的各項成果，我們從其中的幾項趨勢作為啟發，提出了一些未來可能可以改進的方法，期望能夠成為個人化的生成式對話機器人之相關研究的助力，共同構築未來這個領域的發展藍圖。

5.2.1 使用規模更龐大的模型來實作

本研究中使用的 BERT 模型為 bert-base-uncased，為 Google 推出的同名系列中預訓練規模最小的模型。比它大的同類模型包括了 bert-large-uncased[32] 以及它的變體 RoBERTa[33] (A Robustly Optimized BERT Pretraining Approach)，在自然語言處理的任務中，大部分的研究都顯示這兩個模型的性能將優於 bert-base-uncased，前者的原因在於更加龐大的數據量以及模型參數量，後者則是更加嚴格的訓練過程以及參數的設置，兩者對於任務訓練所需要花費的資源都大於 bert-base-uncased 許多。而本研究最終選擇 bert-base-

uncased 的原因在於實驗時所使用的設備之運算資源並不足以負荷最佳化情況的 bert-large-uncased 或是 RoBERTa，若是為了強行使用而去修改理想的參數設定，則最終測得的性能也比不上 bert-base-uncased，因此我們才決定以 bert-base-uncased 作為實驗使用的模型。

5.2.2 和機器人運行的平台中的其他元素結合

本實驗的運行平台為 Discord[34]，在此平台中的有個標籤設定，稱作身分組，它能夠賦予在同一個伺服器的帳號共用的標籤，例如本研究中的 MBTI 人格，透過這些標籤，機器人就能夠讀取並且作為額外的資訊來增進回覆地質量，滿足使用者個人化需求。而除了 MBTI 標籤之外，我們認為也可以增加其它的元素，例如情緒、年齡、喜好……等等，來增加機器人對於使用者認知的多元性。另外，我們也認為，在現階段判定使用者 MBTI 人格之後，除了賦予對應的 MBTI 標籤身分組之外，也可以根據這個人格來賦予機器人 MBTI 的標籤，進而增進聊天時的舒適程度。只不過關於兩個對應的 MBTI 對聊天的影響，目前雖然有相關的資料，但是一直沒有一個很確切的說法，因此現階段我們只有做到以網路上整理出來的 MBTI 適性人格對應表中，隨機選出其中一種適性較高的，來作為機器人的人格進行回覆。

5.2.3 準備更大規模 & 多元 & 多語言的資料集

現階段的實驗都是透過 (MBTI) Myers-Briggs Personality Type Dataset 來進行實作，主要原因在於此資料集在 Kaggle 平台上已經存在六年之久，不只是有許多高手參與競賽，更有學者以這個資料集為基礎來發表研究結果，因此本研究會選用這個資料集，除

了追求資料的質量之外，也包含了它所具備的公信力。然而若是要對研究更加精進，那我認為數量及涵蓋領域都有增進的空間。以目前 8675 筆的數據量來說，若能夠在確保質量的情況下增加數據量，我想對於未來使用更大的模型來訓練或是延伸任務型態都將會有所幫助。而根據參考文獻 ([8]) 中提到此數據集由於僅從一個論壇中提取，因此很容易因為焦點在同一主題上而導致數據偏差，進而去影響模型訓練後的泛化能力。對此有研究人員嘗試從其他主題多樣性更豐富的網站進行資料提取，但是對於 MBTI 標籤分布的極度不平衡以及標註 MBTI 系統的不一致性，讓數據集在使用上有諸多限制，為此若能夠提出一個詳盡且完成驗證的 MBTI 數據資料集，將會很大程度的幫助在這個領域上的研究。而本次實驗中，受測者使用的對話語言基本上都是以中文為主，不過由於訓練數據集是由英文作為使用語言，因此為了要讓模型能夠辨識使用者的對話，會將使用者對話透過模組轉成英文，才能夠分類 MBTI 標籤，但這也導致一個問題，就是某些中文用語直接轉成英文會沒辦法完整表達原先的涵義，例如某些俗語、諧音梗、同音不同意的詞等等，這會使得我們的模型接收到錯誤的資訊而誤判了原先的意思，最後給出不符合期望的 MBTI 標籤。對此如果能夠蒐集各個語言的資料集來進行訓練，就應該能夠解決因為不同語系的隔閡而造成理解落差的問題，進而增加實際使用時在 MBTI 標籤辨識上的準確度。

5.2.4 結合針對「連續對話」而建立的訓練資料集

目前在進行 MBTI 人格預測方面的資料集基本上都是來自各個論壇中使用者的推文或留言蒐集而成，由於只要是該使用者發布的文字訊息就有被納入的可能，因此這些數據之間通常並不會有所關聯。然而如果是像本研究中應用於對話系統上，輸入的內容是使用者的連續對話，那這其中就必然會存在從論壇非連貫文句中得不到的關

聯，例如文句之間的時序性等。我們認為缺乏這點對於辨識對話會有所的影響，就訓練而言，連續的對話資料能提供更多的上下文來讓模型更好的理解對話流程與背景，而不連續的對話資料則提供更廣泛的語境來讓模型適應各種不同的用語及風格，兩者各有優勢，因此我們認為應該綜合的使用這兩種資料集，才能讓模型在理解整體上下文語意的同時也能因應不同情境下展現出強大的泛化能力。

5.2.5 以最大化使用者對話中的人格特徵為目標的改良

從套用了 MBTI 的對話機器人實驗結果來看，雖然數據上的表現有好有壞，但細看這些對話結果可以發現，在大部分的測試資料中，倘若對話中出現了明顯不含人格意義的字句，例如：你好、再見，就很容易讓模型方法在判斷時，將結果判別為同一個人格類型 (IN-)，而一旦這樣的句子在對話中占據一定的比例，就很容易將整體的結果引導到特定的人格，而無法呈現受測者真正代表的人格。以這個問題做為發想，我們認為有兩個可行的方法能夠改善，第一點是設立機器人與使用者的聊天主題，我們發現使用者會輸入不具人格特徵的話，有很大的因素在於他們與機器人之間沒有一個能夠共鳴的話題，這導致使用者容易詞窮或是不知道該聊什麼時，用不具人格特徵的短句帶過，倘若能透過某些手段獲取使用者的個人資訊，來決定聊天主題供使用者參考，我們認為能有效避免掉這個狀況。第二點則是讓系統能夠自行篩選使用者的文句中富含人格特徵的語句作為辨識目標，由於普通人在聊天時，即使是談論要事，或多或少也會在其中參雜不具意義的字句，假如我們能夠透過模型來識別出兩者的不同，先剔除沒有意義的部分再輸入，那對於最終的人格辨識之準確應該會有所成效。

參考文獻

- [1] Bharadwaj S, Sridhar S, Choudhary R, and Srinath R. Persona traits identification based on myers-briggs type indicator(mbti) - a text classification approach. In 2018 international conference on advances in computing, communications and informatics (ICACCI), 2018.
- [2] M. H. Amirhosseini and H. Kazemian. Machine learning approach to personality type prediction based on the myers–briggs type indicator®. Multimodal Technologies and Interaction, 4(1):9, 2020.
- [3] D. Fernau, S. Hillmann, N. Feldhus, and T. Polzehl. Towards automated dialog personalization using mbti personality indicators. In Proc. Interspeech, pages 1968–1972, 2022.
- [4] Charles Li et al. Feature extraction from social media posts for psychometric typing of participants. In Augmented Cognition: Intelligent Technologies: 12th International Conference, AC 2018, Held as Part of HCI International 2018, pages July 15–20, 2018, Las Vegas, NV, USA, 2018. Springer International Publishing.
- [5] Brandon Cui and Calvin Qi. Survey analysis of machine learning methods for natural language processing for mbti personality type prediction. Google Scholar Google Scholar Reference 1, 2017.
- [6] R. L. Vásquez and J. Ochoa-Luna. Transformer-based approaches

for personality detection using the mbti model. In 2021 XLVII Latin American Computing Conference (CLEI), pages 1–7, Cartago, Costa Rica, 2021. doi: 10.1109/CLEI53233.2021.9640012.

- [7] Y. Mehta, S. Fatehi, A. Kazameini, C. Stachl, E. Cambria, and S. Eetemadi. Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. In 2020 IEEE International Conference on Data Mining (ICDM), page 1184–1189, 2020.
- [8] Choong EJ and Varathan KD. Predicting judging-perceiving of myers-briggs type indicator (mbti) in online social forum. PeerJ, 9:e11382, 2021. doi: 10.7717/peerj.11382.
- [9] ChatGPT. Chatgpt. <https://chat.openai.com/>, n.d. Accessed: [2023.12].
- [10] Barbara Plank and Dirk Hovy. Personality traits on twitter—or—how to get 1,500 personality tests in a week. In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 92–98, Lisbon, Portugal, 2015. Association for Computational Linguistics.
- [11] Ben Verhoeven, Walter Daelemans, and Barbara Plank. Twisty: a multilingual twitter stylometry corpus for gender and personality profiling. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 1632–1637, Portoroz, Slovenia, 2016. European Language Resources Association (ELRA).
- [12] Raje MS and Singh A. Personality detection by analysis of twitter profiles. In Proceedings of the eighth international conference on soft

- computing and pattern recognition (SoCPaR 2016), pages 667–675, 2017. doi: 10.1007/978-3-319-60618-7_65.
- [13] Lima ACES and Castro LND. Tecla: a temperament and psychological type prediction framework from twitter data. *PLOS ONE*, 14(3):e0212844, 2019. doi: 10.1371/journal.pone.0212844.
- [14] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, and T.Y. Liu. Lightgbm: a highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154, 2017.
- [15] K.A. Nisha, U. Kulsum, S. Rahman, M.F. Hossain, P. Chakraborty, and T. Choudhury. A comparative analysis of machine learning approaches in personality prediction using mbti. In A.K. Das, J. Nayak, B. Naik, S. Dutta, and D. Pelusi, editors, *Computational Intelligence in Pattern Recognition*, volume 1349 of *Advances in Intelligent Systems and Computing*, Singapore, 2022. Springer.
- [16] Xintao Wang et al. Does role-playing chatbots capture the character personalities? assessing personality traits for role-playing chatbots. *arXiv preprint arXiv:2310.17976*, 2023.
- [17] OpenAI. Openai. <https://openai.com/>, n.d. Accessed: [2023.12].
- [18] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. 2018.
- [19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

- [20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pages 1877–1901, 2020.
- [21] Daniel S. Goldman. A stateful multi-context aware design using openai’s gpt (towards digital sentience). 2023.
- [22] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744, 2022.
- [23] MBTI myers-briggs personality type dataset. <https://www.kaggle.com/datasets/datasnaek/mbti-type>, n.d. Accessed: [2023.12].
- [24] Personality cafe. <https://www.personalitycafe.com/>, n.d. Accessed: [2023.12].
- [25] Nlpaug. <https://pypi.org/project/nlpaug/>, n.d. Accessed: [2023.12].
- [26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.
- [27] Bert-base-uncased. <https://huggingface.co/bert-base-uncased>, n.d. Accessed: [2023.12].
- [28] Models-openai api. <https://platform.openai.com/docs/models/gpt-3-5>, n.d. Accessed: [2023.12].
- [29] I. B. Myers, M. H. McCaulley, N. L. Quenk, and A. L. Hammer. Mbti

- manual: A guide to the development and use of the myers-briggs type indicator. In Consulting Psychologists Press, 1998.
- [30] Carl G. Jung. Personality types. 1971.
- [31] Free personality test, type descriptions, relationship and career advice. <https://www.16personalities.com/>, n.d. Accessed: [2023.12].
- [32] Bert-large-uncased. <https://huggingface.co/bert-large-uncased>, n.d. Accessed: [2023.12].
- [33] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. CoRR, abs/1907.11692, 2019.
- [34] Discord. <https://discord.com/>, n.d. Accessed: [2023.12].



附錄 A — 使用者問卷形式

MBTI套用機器人之測試用表單

此表單為“套用於聊天機器人之MBTI人格指標研究”之回饋使用，請在完成Discord中的一系列指示後，再填寫這份表單，感謝您的配合！

[登入 Google 即可儲存進度。瞭解詳情](#)

* 表示必填問題

請填寫16Personality網站的結果 *

您的回答

請填寫你的初始MBTI人格(\$MBTI_test的結果)*

您的回答

請填寫你最終的MBTI人格(2輪對話之後的結果) *

您的回答

你認為在第一輪對話中，前一次的對話比較好，還是後一次的對話比較好？*

選擇

繼續 清除表單

Figure A.1: 使用者問卷-階段一

後兩題請以"後"一次相對"前"一次的狀況作答

你認為後一次的對話相對前一次的對話，在語意上有變化嗎？*

- 1. 變得很差，對話更難理解
- 2. 有變差，但還能接受
- 3. 沒什麼變化
- 4. 有變好，但沒有很深的感受
- 5. 變得很好，對話更加精確

你認為後一次的對話相對前一次的對話，在對話感受上有變化嗎？*

- 1. 變得很差，對話讓人不愉快
- 2. 有變差，但還能接受
- 3. 沒什麼變化
- 4. 有變好，但沒有很深的感受
- 5. 變得很好，對話讓人更舒適

對於將MBTI人格指標套用在聊天機器人上，您有什麼建議或看法嗎？

您的回答

返回

提交

清除表單

Figure A.2: 使用者問卷-階段二