

天主教輔仁大學統計資訊學系

第二十二屆專題研究成果報告

指導教授：杜逸寧博士

聊癒療鬱 Chatting your style healing your soul

——以語言風格匹配及情感分析療癒具有負面情緒傾向者
的自然語言系統

學生：陳琪鈞、黃經圖、蔡方寧、李逸詩

、陳雅柔、李品律、吳家欣 撰

中華民國 111 年 12 月

目 錄

第壹章 緒論	14
第一節 研究背景	14
第二節 研究動機	14
第三節 研究目的	15
第貳章 文獻探討	16
第一節 語文探索與字詞計算的發展	16
(一)LIWC	16
(二)LSM	17
第二節 Word2Vec 演算法	19
(一)Word2Vec	19
(二)CBOW	20
(三)Skip-gram	21
(四)總結	22
第三節 中文斷詞系統	22
(一)CKIP	23
(二)Jieba	23
(三)CKIP、Jieba 比較	25
第四節 K-Nearest Neighbors 演算法	25
第五節 邁爾斯—布里格斯性格分類指標	26
第參章 研究方法	28

第一節 研究架構.....	28
(一)Word Embedding.....	28
(二)句子結構模型	28
(三)word2vec 調整句子向量詞性權重	28
第二節 研究流程.....	29
第三節 資料預處理.....	30
(α)一般答覆資料集	30
(β)負面情緒資料集	31
(γ)聊天答覆資料集	31
(δ)S2VecQ 資料集.....	32
第四節 訓練語言風格模型.....	34
(a)語言風格語料庫	35
(b)N-gram 模型.....	35
(c)Word Embedding 模型.....	36
(d)句子結構模型	37
(e) 語言風格模型	37
(f)透過不同語言風格挑整 DatasetA.....	39
(g)多種語言風格之聊天答覆資料集 LsmData	40
第五節 聊天及產生回覆語句	40
(1-1)計 算 S2VecQ 與 Input 間的 餘 弦 相 似 度	40

(1-2)回覆使用者	45
(2-2)使用 LIWC 辭典分析使用者的情緒	47
(2-3)聊天機器人主動關心使用者	48
第肆章 實驗結果	49
第一節 資料收集	49
(一)一般答覆資料集	49
(二)負面情緒資料集	49
(三)使用者 MBTI	50
第二節 語言風格模型	51
(一)句子結構模型的多種語言風格語料庫	51
(二)語言風格模型參數調整	57
第三節 挑選適當語句回覆使用者	59
第四節 實驗設計與驗證	61
(一)驗證 MBTI 做為因素是否更適合使用者	61
(二)使用者是否偏好語言風格模型轉換過後的語句	62
第伍章 系統介面	64
第陸章 結論與建議	66
第一節 研究結果	66
第二節 研究限制與未來發展	66
(一)研究限制	67
(二)未來發展	67

參考文獻.....	69
-----------	----

附錄一 蒐集之網路文章來源.....	72
--------------------	----

附錄二 蒐集之網路作者文章來源.....	73
----------------------	----

表 目 錄

表 2-1 中文語文風格匹配度(LSM)所涵蓋之功能詞	18
表 2-2 測試文檔.....	25
表 2-3 斷詞結果比較	25
表 3-1 一般答覆資料集之部分資料	30
表 3-2 原始聊天機器人交流群資料集部分資料	30
表 3-3 修改後聊天機器人交流群資料集部分資料	30
表 3-4 壓力來源為學業面向之問題	31
表 3-5 將 Q 改以不同方式詢問但保留原有語意之範例	32
表 3-6 Input 與 DatasetQ 相似度對應示意表	32
表 3-7 不同分群方法群集間餘弦相似度數值	34
表 3-8 使用者 Input 及 Jieba 斷詞結果.....	40
表 3-9 兩字詞間的相似度註記	40
表 3-10 不同維度下的 Pearson 相關係數.....	42
表 3-11 使用者輸入之聊天語句	46
表 3-12 聊天答覆資料集之部分資料	46
表 3-13 NegQ23 資料集之部分資料.....	46
表 3-14 使用者資訊轉換為虛擬變數 dummy variable 型態	47
表 3-15 使用者資訊轉換為 dummy variable	47
表 3-16 使用者與各問卷填寫者之歐幾里德距離	47
表 3-17 計算負面情緒詞比例之範例	48

表 4-1 一般答覆資料集分類	49
表 4-2 負面情緒資料集各問題收集筆數	49
表 4-3 將目標變數 YES 放大 100 倍之評測指標	52
表 4-4 蒐集之網路作者名單	54
表 4-5 語言風格不相似之兩兩作者	55
表 4-6 最終挑選出之作者語料庫	57
表 4-7 語言風格模型參數定義	57
表 4-8 500 組組合的屬性值範例	57
表 4-9 決策樹表現	58
表 4-10 語言風格模型在 riceandshine 語言風格的表現	58
表 4-11 使用者 Input 及 Jieba 斷詞結果示意表	60
表 4-12 各 MBTI 人格偏好相似或互補之比例	61
表 4-13 相同 S2VecQ 產出 LsmA 之作者的各種組合	62
表 4-14 MBTI 各面向之語言風格偏好程度	63

圖 目 錄

圖 1-1 三階段之問卷	15
圖 2-1 Word2Vec 向量空間	19
圖 3-1 研究流程圖	29
圖 3-2 階層式分群之分群結果	33
圖 3-3 Tri-gram 舉例	36
圖 3-4 句子結構示意圖	37
圖 3-5 語言風格模型架構圖	37
圖 4-1 問卷回覆比例與實際 MBTI 人格類型之分佈比例比較	51
圖 4-2 各 MBTI 人格類型的回覆數	51
圖 4-3 將目標變數 YES 放大 100 倍之評測指標作圖	53
圖 4-4 挑選適當語句回覆使用者流程圖	60
圖 5-1 登入頁面	64
圖 5-2 註冊頁面	64
圖 5-3 忘記密碼	64
圖 5-4 測驗解說	64
圖 5-5 MBTI 測驗	65
圖 5-6 負面情選擇題	65
圖 5-7 主頁面	65
圖 5-8 忘記密碼	65

摘要

論文名稱：聊癒療鬱——以語言風格匹配及情感分析療鬱具有負面情緒傾向者的自然語言系統

頁數：73 頁

校別所：天主教輔仁大學 統計資訊學系

發表時間：一百一十一學年度第一學期

學位：學士班

專題生：陳琪鈞、陳雅柔、黃經圖、蔡方寧、李逸詩、李品律、吳家欣

指導教授：杜逸寧 博士

關鍵詞：聊天機器人、語言風格匹配、語言風格模型、句子結構模型、自然語言理解、情感分析、Word Embedding、word2Vec、Word2FunctionVec、N-gram、KNN

論文摘要內容：

本研究建立一套「以語言風格匹配及情感分析療癒具有負面情緒傾向者的自然語言系統」，旨在改善聊天機器人的回覆語句使之變得更多元以符合使用者的語言風格偏好以及使用語文探索與字詞計算辭典以分析使用者的情緒。本研究整合了青云客網絡商業公司之聊天資料集以及組員自創聊天資料集為此次研究的資料庫。透過 MBTI 測驗以及三階段的語言風格偏好選擇題，來瞭解使用者的語言風格偏好。使用者完成冷啟動後，以使用者偏好的語言風格作為聊天機器人的風格依據回覆使用者。為將聊天語句轉換得符合使用者的語言風格，本研究提出了創新的「語言風格模型」其結合 N-gram、句子解構模型及 Word2FunctionVector，其中句子解構模型及 Word2FunctionVector 為本研究創新建立，原語句透過語言風格模型的轉換，不僅能改善一般市面上聊天機器人回覆制式化的問題，更能貼近使用者之偏好讓使用者有如和朋友般聊天的體驗，打造其專屬的聊天機器人。為了驗證可行性，本研究設計兩個實驗驗證，其中包含驗證回覆語句以 MBTI 做為依據是否符合使用者偏好以及驗證使用者是否偏好語言風格模型轉換過後的語句，並於 2022 年 9 月 16 日至 2022 年 10 月 2 日收集 251 筆有效問卷，其中

173 筆資料為有效樣本。有效樣本中有 55.37% 的使用者較偏好與自己 MBTI 互補的回覆語句，並以 MBTI 十六種人格做分析，發現 ISTP 及 INTJ 有高達 70% 以上的使用者偏好 MBTI 互補的語句，可得知 MBTI 可能影響使用者的偏好。另外約有 53% 的使用者偏好語言風格，更以 MBTI 中的四面向做分析，發現 N-直覺型及 P-感知型的使用者皆有高達 60% 以上的使用者偏好語言風格模型轉換後的語句。可得知本研究設置的語言風格模型更改後的語句在一定程度下可使聊天機器人回覆語句變得更多元且符合使用者的語言風格偏好，成為使用者的專屬聊天機器人。

Abstract

Paper Title: Chatting your style healing your soul - building a chatbot's language style matching user and analysis user's emotion with Linguistic Inquiry Word Count Dictionary response user by using K Nearest Neighbor method

Pages: 73 pages

School: Fu Jen Catholic University

Department: Department of Statistics and Information Science

Time: December, 2022

Degree: Bachelor

Researcher: Chi-Chun Chen , Ching-Tu Huang, Fang-Ning Tsai, Yi-Shih Li, Ya-Rou Chen, Pin-Lu Li, Chia-Hsin Wu

Advisor: Ph.D. Yi-Ning Tu

Key Words: ChatBot 、 Language Style Matching 、 Language Style Model 、 Sentence structure model 、 NLU 、 Sentiment Analysis 、 Word Embedding 、 word2Vec 、 Word2FunctionVec 、 N-gram 、 KNN

Abstract:

This study establishes a "Natural Language System for Healing Negative Emotional Tendencies by Language Style Matching and Sentiment Analysis ", aiming to improve the reply sentences of chatbots, transfer them to be more diverse to match the user's language style preferences, and also the Linguistic Inquiry and Word Count Dictionary is used to analyze user's emotions. The data set of the research was establish by Qingyunke Internet Business Company and the data set which created by team members. In order to convert sentences to suit the user's language style, this research exclusively created a "language style model" which combines N-gram, sentence deconstruction model and Word2FunctionVector, among which the sentence deconstruction model and Word2FunctionVector are innovatively established for this research , the conversion of

the original sentence through the language style model can not only improve the standardization of responses from chatbots on the market, but also be closer to the user's preference so that the user can chat with chatbot like chatting with friends and create their own chatbot. In order to verify the feasibility, this study designed two experimental verifications, including verifying whether the reply sentence based on MBTI conforms to the user's preference and verifying whether the user prefers the sentence after the language style model conversion, and on September 16, 2022 By October 2, 2022, 251 valid questionnaires were collected, of which 173 were valid samples. There is 55.37% of the users in valid sample prefer the reply sentences that complement their own MBTI, and analyze the sixteen types of personality of MBTI. Founded that more than 70% of the users whose MBTI are ISTP and INTJ prefer the sentences. Indicate that MBTI may influence user preferences. In addition, about 53% of the users prefer the language style. Based on the analysis of the four aspects in the MBTI, it is found that both N-intuitive and P-perceptual users have a preference of more than 60% of the users after the language style model conversion. statement. It can be seen that the changed sentences of the language style model set in this study can make the chat robot reply sentences more diverse and in line with the user's language style preference to a certain extent, and become the user's exclusive chat robot.

謝 辭

在這一年專題研究中，首先要感謝的是我們的指導老師—杜逸寧教授，每週固定開會不斷與我們確認研究進度及方向，每當研究遇到瓶頸時，老師宛如我們的一盞明燈引領我們前進，研究過程中不斷地討論、溝通激發出許多新想法。除了研究及教學上指導我們以外，老師也相當關心每個組員，並且不吝於鼓勵與讚美，讓我們更加有自信。

也很感謝一起努力的組員們，在專題製作的這段時間，我們經歷了各種大大小小的困難，例如起初對程式的不熟悉、資料集及文獻尋找的不易等等，但因為組員們的互相扶持，並且交流討論不同意見，讓我們一一克服了許多難關。這一年，我們學習了多種演算法，創新獨有模型及功能，蒐集上萬筆資料，加入 APP 介面的美工設計，路途顛頗卻也獲益良多。很感謝組員們在課業繁忙的同時，也對專題如此重視，並在大家的堅持及同心協力下，我們才能一起完成這項艱鉅的挑戰。

最後，感謝所有親朋好友的協助，幫助我們填答如此繁瑣的問卷，並且下載系統使用後給予建設性的建議，讓我們優化系統使之更完善，並不斷給予支持及肯定，這些都是我們努力的動力，謝謝個為的陪伴與鼓勵。

陳琪鈞 黃經圖 蔡方寧 李逸詩 陳雅柔 李品律 吳家欣 謹誌於

輔仁大學統計資訊學系

中華民國一百一十一年十二月

第壹章 緒論

本章將針對本研究的研究背景、動機及目的做說明。第一節在敘述過往相關的研究及本研究發現可優化之功能，第二節將介紹本研究使用何種研究方法來解決，第三節說明將如何驗證本研究。

第一節 研究背景

新冠狀病毒席捲全球，造成許多國家經濟衰退及心理健康產生負面影響。根據世界衛生組織的數據，抑鬱症影響超過 2.64 億人口(Afshin et al., 2017)，印度甚至遭受高達 1.03 億美元的精神損失(Birla, 2019)。人工智慧的應用——聊天機器人近而倍受關注，其成為照顧人們心理健康的一種新興工具(Tewari, Chhabria, Khalsa, Chaudhary, & Kanal, 2021)。

英國電腦科學家 Alan Turing 於 1945 年至 1948 年提出近代人工智慧理論，(Turing, 1950)提道圖靈測試(Turing test)為判斷機器是否具備知能的標準。美國麻省理工學院人工智慧實驗室的德裔電腦科學家 Joseph Weizenbaum 在 1964 年至 1966 年間打造史上第一個聊天機器人「Eliza」。Eliza 被設計定位為「心理治療師」，使用者藉由與其對話來抒發心情，雖然當時 Eliza 並沒有通過圖靈測試，但不少人仍將 Eliza 視為最早的聊天機器人(許芳瑤，2020)。Eliza 會依循提問內容重覆說詞，或是針對關鍵字詞進行回答，藉此滿足提問者內心預期聽到答案，進而達成讓提問者認為對話對象是真人。儘管如此，(Weizenbaum, 1996)仍提出 Eliza 的幾個技術問題，其中一個技術問題是：若使用者輸入的內容關鍵字不在 Eliza 的回覆預設腳本中，其回覆會不合乎自然人聊天的邏輯。故本研究欲藉由情緒分析及使用語言風格匹配技術應用在機器學習，使用機器學習中的自然語言模型訓練聊天機器人使之以較人性化的方式回覆使用者，改善回覆過於制式化的問題，優化聊天機器人使之以更人性化的語句回覆使用者。

第二節 研究動機

首先，本研究將使用邁爾斯—布里格斯性格分類指標(Myers-Briggs Type Indicator, 以下簡稱 MBTI)將使用者依人格分類，希望驗證與使用者相似人格之回覆語句是否較適合使用者。再來，本研究將透過語言風格匹配模型，改善聊天機器人回覆制式化問題，欲藉此達到機器人語言風格與使用者相似。且本研究將

會利用語文探索與字詞計算(Linguistic Inquiry and Word Count, 以下簡稱 LIWC) 辭典，分析使用者情緒傾向(張瓊之，2019) (Huang et al., 2012)。最後，將計算聊天機器人與使用者之間的語言風格匹配程度，數值越高，則表兩者語言風格相似度高，達到本研究之目的；反之則表語言風格相似度低，本研究將會使機器人重新選擇語言風格模型之資料集，以尋找最適合使用者之回覆語句資料集。

第三節 研究目的

1. 語言風格模型訓練，欲產生具有語言風格之聊天答覆資料集
2. 調整 Word2Vec 向量維度至最適合本研究之維度並調整句向量計算方式
3. 計算 LIWC 判定使用者是否處於負面情緒之門檻值
4. 驗證以相似 MBTI 做為因素是否能得到更適合使用者的回覆
5. 驗證本研究為是否能透過計算語言風格相似度，正確預測使用者的語言風格偏好

首先，為了使機器人擁有自然語言處理的能力，本研究研究創新製作語言風格模型，使機器人擁有不同語言風格的回覆。再來，使用 Word2Vec 語言模型調整之向量維度使其適合本研究之語料庫，並設置適合本研究語料庫之句子向量參數。且使用 LIWC 辭典分析使用者之情緒。

最後，為了驗證本研究之假設，設計了三階段之問卷，如下(圖 1-1)。第一階段為蒐集負面情緒問答資料集(將於第參章第三節詳細說明)及計算語言風格相似度。第二階段為驗證相似 MBTI 之回覆是否更能療癒使用者的心情。第三階段為驗證本研究是否能正確預測使用者的語言風格偏好。

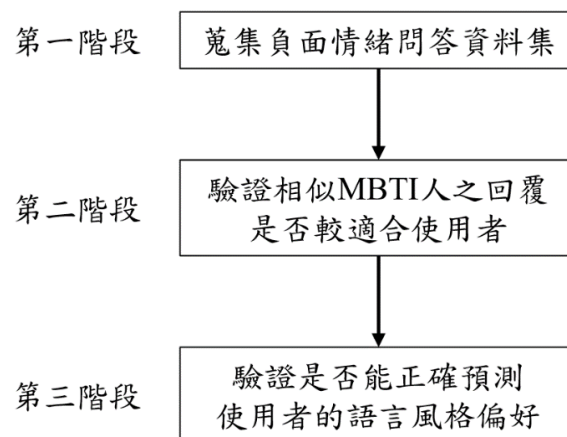


圖 1-1 三階段之問卷

第貳章 文獻探討

為建立本研究之研究架構與理論基礎，本章節將探討專家學者過去相關研究文獻。本章共有六小節，根據本研究的流程首先第一節探討語文探索與字詞計算之文獻，語文及字詞藉由文辭的分析可分成(1)語文探索與字詞計算(2)綜合性語言風格匹配度，接著第二、三節將探討介紹自然語言之演算法在聊天機器人上的應用，其中包含了 Word2Vec、及 N-Gram Language Models 兩種演算法，第四節是講解本研究使用到的中文斷詞系統，第五節為 K 鄰近值演算法之文獻探討，第六節為訓練機器人時使用到的邁爾斯—布里格斯性格分類指標。

第一節 語文探索與字詞計算的發展

(一)LIWC

在日常生活中，語言的使用，反映了每個人內在的心理狀態、思考模式以及人格特質，也因為如此，語言的使用對於心理學研究來說是一個很重要的窗口。[\(Pennebaker, Francis, & Booth, 2001\)](#)採用字詞計算的方式進行語文特性的分析，並發展出電腦程式「語文探索與字詞計算」(Linguistic Inquiry and Word Count, 以下簡稱 LIWC)。LIWC 最核心的部分在於其詞典，透過將字詞分類來進行語文特性的分析，其中 LIWC2007 包含 4 個一般描述性類別(總詞數、每句詞數等)、22 個語文特性類別(代名詞、冠詞等百分比)、23 個心理特性類別(情感詞、認知詞等)、7 個個人化類別(工作、休閒活動等)、3 個副語言學(paralinguistic)類別(應和詞、停頓詞等)以及 12 個標點符號類別(如句點、逗點等)，總計有 80 個字詞類別，具有相當好的信效度。

上述的 LIWC2007 為英文版的辭典，中文版 LIWC 便是由 LIWC2007 而來，先透過翻譯以及檢視各類別底下的字詞是否適當，然後依照中文特性進行類別的增刪。例如，增加中文特有的第二人稱單／複數、量詞、語尾助詞等類別；中文沒有動詞時態的差異，因此刪去現在式、過去式與未來式等類別，但也另建了相對應概念的各類時態標記詞。在增刪類別後，再經過斷詞系統的處理，確認分詞是否適當。建立過程經過多次的討論，最後逐詞確認後，中文版 LIWC(以下簡稱 CLIWC)才建立完成[\(Huang et al., 2012\)](#)。

LIWC 被應用在非常多層面，包括心理、商業、臨床、教育等等。分析語言的使用特性，可以知道當事人的注意力所在並探討個別差異。例如，女性使用較多的社會詞與他人相關詞；男性的語言型態則傾向較為複雜 ([Newman, Groom, Handelman, & Pennebaker, 2008](#))；自殺詩人在其作品中使用較多的第一人稱單數代名詞與死亡詞 ([Stirman & Pennebaker, 2001](#))；高外向性者是用較多社會詞、正向情緒詞，較少的負向情緒詞與複雜詞 ([Mehl, Gosling, & Pennebaker, 2006; Pennebaker & King, 1999](#))。

本研究欲使用 LIWC 辭典來分析使用者的情緒，透過斷詞系統，將使用者的聊天內容分割成字詞，並與 LIWC 辭典的各類別字詞分析比較，若是經計算後發現使用者負面情緒詞的使用率較高，聊天機器人便會主動關心使用者，幫助使用者紓解情緒，並和使用者有相似的語言風格，達到共情的效果。

(二)LSM

依照文法特性語文可以大分為內容詞(實詞)和功能詞(虛詞)。其中，功能詞(如介係詞、量詞、連接詞等)本身沒有特定的內容意涵，主要在串連句中的內容詞使語句通順完整，功能詞本身數量雖然不多，但是在語言的使用中卻涵蓋了將近 50% 的使用率。由於功能詞扮演了人們組織語言的角色，相對於內容詞，功能詞更能作為個體語言風格的指標。因此，透過 LIWC 分析各功能詞類別的使用率，可計算出兩篇對話(文本)的語文風格匹配度(language style matching, 以下簡稱 LSM)。([Ireland, & Pennebaker, 2010](#))

LSM 的運算如下：首先將兩篇對話(文本)進行 LIWC 分析，並針對功能詞使用比率，分別計算在兩篇文章中的使用率相似度。以介係詞為例，如公式(1)：

$$\text{LSMpreps} = 1 - [\text{abs}(\text{preps1} - \text{preps2}) / (\text{preps1} + \text{preps2} + 0.001)] \text{ 公式(1)}$$

- *preps1*: 第一篇文本的介係詞使用率

- *preps2*: 第二篇文本的介係詞使用率

其中 0.001 是為了避免分母為 0 而加。在公式(1)中，分子是兩篇文章使用介係詞比率的差異絕對值，也就是說，兩篇文章何者較高或較低並不重要，重要的

是兩者的相似程度，相似性越高，差異越小，反之，分子的值越大，兩者的差異性也越大。最後，將各類功能詞的 LSM 平均後得到一個整體的 LSM 指標。

(Ireland, & Pennebaker, 2010)的研究分析心理學家、文學配偶等之書信與作品的語文型態相似度。以佛洛伊德與容格的往來書信為例，結果發現他們的 LSM 隨著關係的變化而隨之變動。在關係親密無間時，彼此書信展現相當高的語文型態匹配度；關係決裂時，則得到最低的 LSM。其他相關研究也發現，在小團體溝通時，成員間彼此的 LSM 可以預測團體的凝聚力及其後續合作表現(Gonzalez, Hancock, & Pennebaker, 2010)。LSM 也可以預測在快速約會中，雙方的好感度，同時也可有效預測情侶三個月後的關係穩定度(Ireland et al., 2011)。

以上研究大多使用英語，或將文本翻譯成英文分析，但功能詞的使用並不僅限英語，而是具有語言的普同性的，中文 LSM 的計算修正為中文的各項功能詞類別(如表 2-1)，而非沿用英文功能詞。(張硯評，2011)以實驗法探討夫妻雙方的感恩行為是否有助於減緩對方的憂鬱程度，發現 LSM 對於此效果具有調節效應。亦即，在感恩表達組中，夫妻雙方電子郵件的 LSM 越高，對對方的憂鬱狀態越有緩解效果；相反的，在挫折表達組中(向對方表達日常生活遇到的挫折)，則是 LSM 越高，對方的憂鬱緩解越差。該研究顯示，LSM 的效應在中文使用者身上也可以獲得相似的效果。由上述舉例可知個體間功能詞的使用率的相似性可以作為語言風格匹配程度的指標，且預測效果皆達到正相關。

本研究欲使用 LSM 來達到聊天機器人與使用者的語言風格匹配，和使用者的聊天過程中，分析使用者各功能詞的使用率，以調整聊天機器人的語言風格，隨者使用者與聊天機器人的對話增加，能分析的文本也隨之增多，逐步將聊天機器人的語言風格調整成與使用者最相似的狀態。

表 2-1 中文語文風格匹配度(LSM)所涵蓋之功能詞

變項名稱	簡寫	總詞數	範例
特定人稱代名詞	ppron	33	他、大家、你們
非特定人稱代名詞	ipron	35	一切、這些、其他
助動詞	auxverb	32	不必、可能、應該
副詞	adverb	202	曾經、漸漸、那麼
介系詞	preps	70	從、依照、把
連接詞	conj	94	和、一旦、不僅

變項名稱	簡寫	總詞數	範例
否定詞	negate	268	不要、未必、沒有
概數詞	quant	104	一些、所有、眾多
後置詞	prepEnd	41	之中、以上、為止
特指定詞	specArt	17	本、該、每
量詞	quanUnit	123	條、頭、枝
語助詞	interjunction	34	呢、嗎、吧
多用途詞	multiFun	11	的、有、是
時態標定詞	tenseM	85	已經、之前、日後

第二節 Word2Vec 演算法

Word2Vec 是由(Mikolov, Sutskever, Chen, Corrado, & Dean, 2013)提出的語言模型，其中使用了 CBOW(continuous bag-of-word model)和 Skip-gram 訓練模型，兩者皆是神經網路的技術。以下將介紹 Word2Vec (一)的概念以及兩種方法(二)、(三)的訓練過程，並在(四)小節做總結。

(一)Word2Vec

Word2Vec 是一個能賦予詞在向量空間上意義的模型，語意上越相近的詞它們向量間的夾角將會越小。舉例來說(圖 2-1):「紅茶」和「綠茶」模型中的向量夾角會很接近，反之，「紅茶」與「電視」的夾角就會較大。

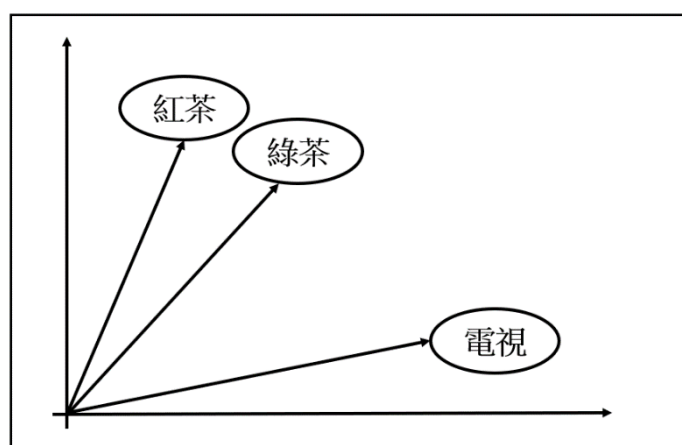


圖 2-1 Word2Vec 向量空間

Word2Vec 是根據目標詞的前後文(context)之間的關係去定義目標詞的向量，CBOW 是用前後文去預測目標詞，而 Skip-gram 則是用目標詞去預測前後文，具體訓練過程會再(二)、(三)小節說明。

(二)CBOW

此小節將以一個例子闡述，假設語料庫中只有三個句子：“the dog saw a cat”，“the dog chased the cat”，“the cat climbed a tree”，共有 8 個詞彙。將此語料庫當作模型的訓練資料集，並設定參數(windows = 2)使模型以目標詞的上下文各一個詞作為根據預測目標詞。另外設定每個詞都有 3 個隱藏特性，以上述參數為設定並以預測“cat”為目的之神經網路圖如下(圖 2-2)。

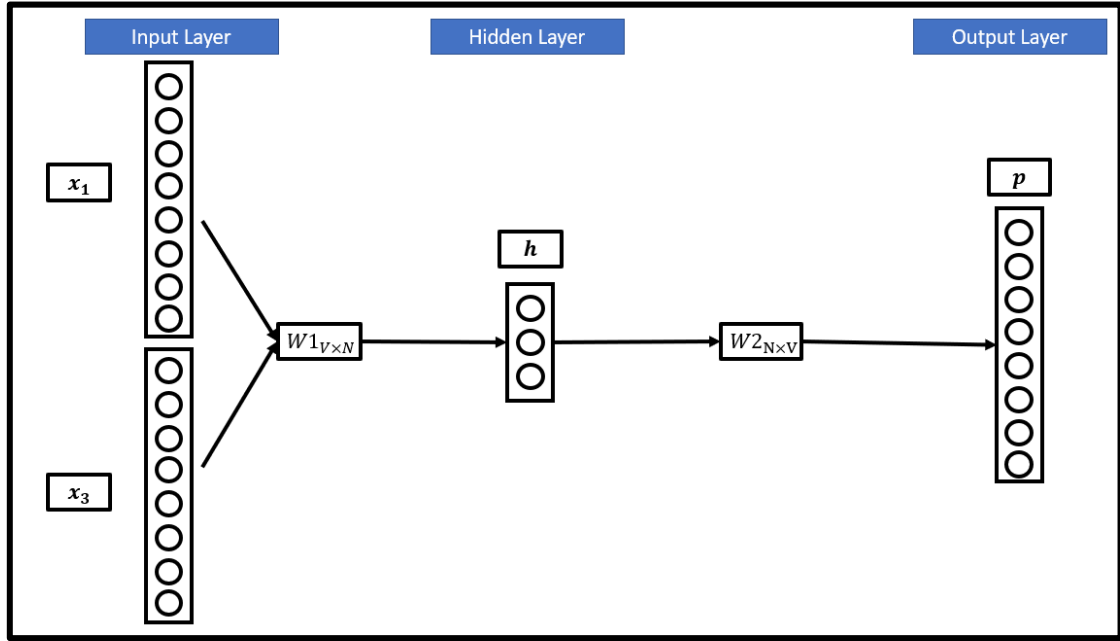


圖 2-2 CBOW 模型

第一步，再輸入層(Input Layer)，會先將“cat”的上下詞輸入，其中 x_1 是指“the”以 one-hot 方式呈現的向量($1 \times V$)， V 是語料庫中詞的數量($V=8$)。假設“cat”為 $[0,1,0,0,0,0,0,0]$ ，則 x_1 則是 $[1,0,0,0,0,0,0,0]$ 。同理， x_3 是“climbed”以 one-hot 方式呈現的向量。第二步， x_1 和 x_3 會經過第一個權重矩陣(特徵矩陣) $W1_{V \times N}$ ，形成 $h(1 \times N)$ ， N 是參數隱藏特性的數量($N=3$)。另外， $W1_{V \times N}$ 一開始是隨機亂數產生的矩陣，會根據最後的結果 p 進行更正。第二步的作用在於找到 x_1 和 x_3 的隱藏特性，並求平均取得 h 如公式(2)。第三步，將 h 與第二個權重矩陣 $W2_{N \times V}$ 進行外積形成 $p(1 \times V)$ ，如公式(3)。

$$h = \frac{1}{2} (x_1 \times W1_{V \times N} + x_3 \times W1_{V \times N}) \quad \text{公式(2)}$$

$$p = h \times W2_{N \times V} \quad \text{公式(3)}$$

其中， $W2_{N \times V}$ 不是 $W1_{V \times N}$ 的轉置， $W2_{N \times V}$ 是另一個隨機亂數產生的矩陣。 p 則代表預測結果， p_i 代表語料庫中第 i 個字的預測分數。接下來的步驟未顯示在圖 2-3 中，屬於評估權重的步驟。首先，對 p 使用 softmax function 歸一化成 y ，如公式(4)：

$$y_i = Pr(word_i | word_{context}) = \frac{\exp(p_i)}{\sum_n^v \exp(p_n)} \quad \text{公式(4)}$$

$word_i$ 代表語料庫中第 i 個字，而 $word_{context}$ 則代表了輸入的上下文。 y_i 代表第 i 個字的預測分數進行歸一化後的結果。以本次的例子來說，會希望 y_2 也就是“cat”的分數會最高，因此會針對 $W1_{V \times N}$ 和 $W2_{N \times V}$ 進行調整，使用到的調整公式在(Mikolov et al., 2013)有說明。接著重複多次第一步到第三步，不斷地調整 $W1_{V \times N}$ 和 $W2_{N \times V}$ ，直到最後的結果使“cat”的分數最高，在迭代的同時也會使用語料庫中其他組的目標詞與上下文。當模型訓練完成後， $W1_{V \times N}$ 就會是一個能區別語料庫中所有詞的詞義的一個特徵矩陣。也就是使用 Word2vec 中所有詞向量的矩陣。

(三)Skip-gram

此小節沿用上節的參數設定以及語料庫進行說明($V = 8, N = 3$)，但 Skip-gram 的做法與 CBOW 完全相反，輸入為目標詞，而輸出則為上下文的預測結果，其神經網路結構圖如下(圖 2-3)

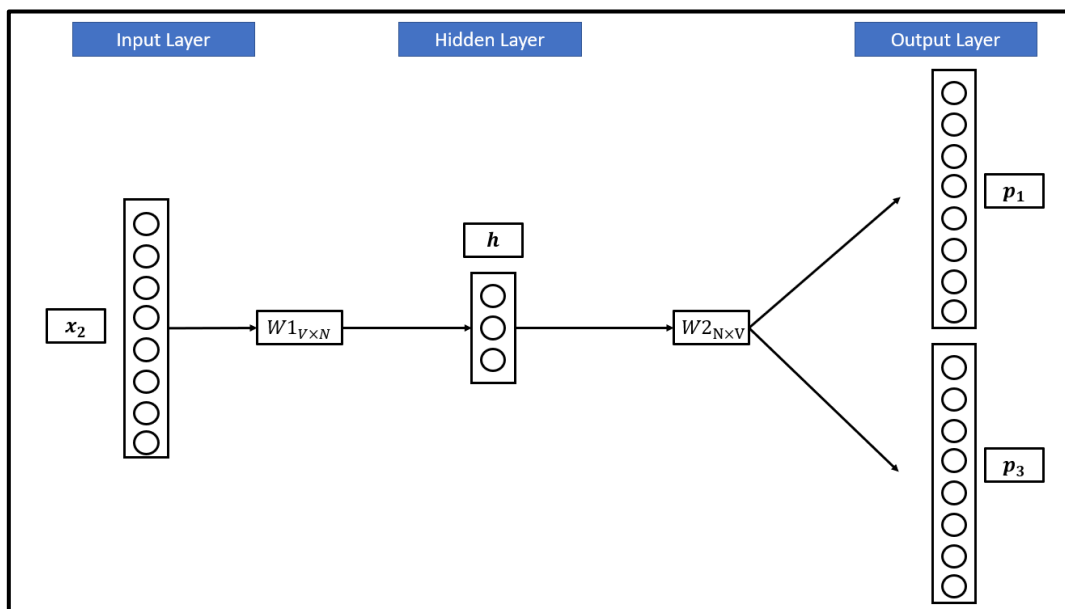


圖 2-3 Skip-gram 模型

第一步，輸入改成目標詞 x_2 。第二步， x_2 會經過第一個權重矩陣 $W1_{V \times N}$ 形成 $h(1 \times N)$ 如上述公式(2)。第三步，將 h 與第二個權重矩陣 $W2_{N \times V}$ 進行兩次外積形成 p_1 和 $p_3(1 \times V)$ ，如公式(5):

$$p_c = h \times W2_{N \times V}, c = 1 \text{ or } 3 \quad \text{公式(5)}$$

p_c 則代表針對上下文其中一個的(“the” or “climbed”)的預測結果， p_{ci} 代表針對第 c 個上下文預測語料庫中第 i 個字的分數之預測結果。接下來的步驟未顯示在圖 2-4 中，屬於評估權重的步驟。首先，對所有 p_c 使用 softmax function 歸一化成 y_c ，如公式(6):

$$y_{ci} = Pr(word_{ci} = word_c | word_{input}) = \frac{\exp(p_{ci})}{\sum_{n=1}^V \exp(p_{cn})} \quad \text{公式(6)}$$

$word_{ci}$ 代表在針對第 c 個上下文的狀況下，語料庫中第 i 個字， $word_c$ 則代表第 c 個上下文。 $word_{input}$ 則代表了輸入的目標詞。 y_{ci} 代表第 i 個字在針對第 c 個上下文的預測分數進行歸一化後的結果。以本次的例子來說，會希望 y_{11} 和 y_{33} ，也就是在針對上下文為“the”的狀況下“the”的預測分數最高以及在針對上下文為“climbed”的狀況下“climbed”的預測分數最高。為達到此目的，如同 CBOW，一樣會對 $W1_{V \times N}$ 和 $W2_{N \times V}$ 進行調整，使用到的公式同樣地在(Mikolov et al., 2013)有說明。重複多次第一步到第三步，不斷地調整 $W1_{V \times N}$ 和 $W2_{N \times V}$ ，直到最後的結果使得 y_{11} 和 y_{33} 分數最高，再迭代的同時也會使用語料庫中其他組的上下文與目標詞。相同地， $W1_{V \times N}$ 是訓練模型後的結果。

(四)總結

根據(Wang, 2014)，CBOW 和 Skip-gram 各有各的好處，Word2vec 的創始人曾說過，Skip-gram 在處理少量資料以及能給予語料庫中較稀少的詞更好的向量意義。而 CBOW 則運行的更快並且在處理常出現的詞中有更好的表現。因本研究需要快速計算大量的語料庫並給予回應，因此將採用 CBOW 作為訓練方式。

第三節 中文斷詞系統

為了分析使用者的回覆訊息，一個適合的中文斷詞系統是必要的。本研究評估斷詞系統是否適合是藉由兩個指標(1)斷詞的速度(2)斷詞的精確度。因為設計一個能對使用者的訊息進行分析再回覆的聊天機器人，分析時間為相當重要的議

題。斷詞的結果也會影響分析的結果，不好的斷詞結果可能會導致分析的誤判。基於以上兩個指標，本研究分別研究了兩個中文斷詞系統 CKIP 與 Jieba，以下將會比較兩者的差距。

(一)CKIP

CKIP 是由台灣中研院資訊所、語言所的中文語言小組所開發。[\(Ma, & Chen, 2005\)](#)中有詳細說明 CKIP 的運作原理，包括用啟發式規則(Heuristic Rules)解決一詞多意的問題，以及未知詞的偵測和斷詞。CKIP 在繁體中文的斷詞精確度上有非常好的表現，有一部分是得益於該研究強大的詞庫。CKIP 所用的詞庫是 CKIP Lab 中的詞庫小組長年的研究結果，因此在斷詞結果上 CKIP 有絕對的優勢。CKIP 目前提供多種功能包括斷詞、剖析系統、實體辨識、指代消解等，同時也具備基本的詞性標註和使用者辭典。

(二)Jieba

Jieba 為中國百度的開發者所撰寫的，初始為針對簡體中文的斷詞系統。程式檔中運用到的字典也是由中國簡體的文檔訓練得出，因此 Jieba 在繁體中文的斷詞精確度上不能說非常好。但後來有台灣的研究者對字典進行翻譯，在每個簡體詞後加上繁體版的詞，使得斷詞結果得到了改善。Jieba 斷詞的步驟如圖 2-4，用字典建構 Trie tree，根據 Trie tree 對句子進行斷詞。Trie tree 能有效地縮短查詢字典所消耗的時間。

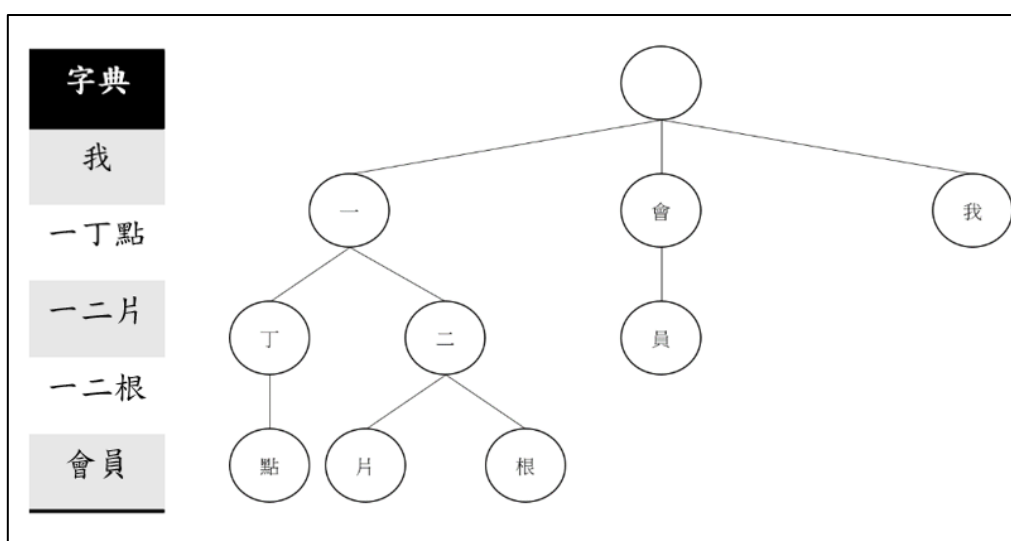


圖 2-4 舉例用字典建構 Trie tree

利用此樹狀結構，在查詢字典中最後一個詞「會員」時，就不用搜尋整部字典，只要以「會員」這兩個文字當作索引進行搜尋就可以找到對應的節點。

一、根據所有可能的斷詞結果生成有向無環圖(DAG)，再依照詞頻找出最大概率路徑。例如句子「即將來臨時」，根據所有斷詞結果會產生圖 2-5:

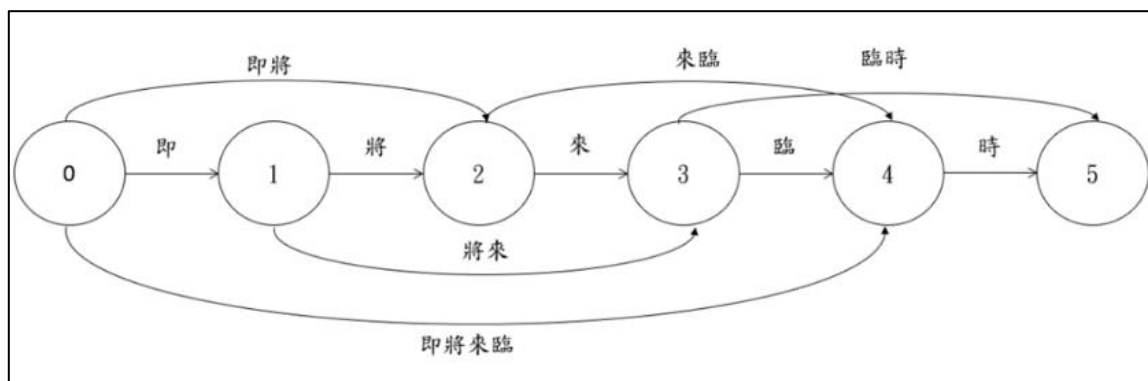


圖 2-5 有向無環圖

假設第一節點「即」的編號為 0，以上圖形可以這樣表示 {0:[1,2,4],1:[2,3],2:[3,4],3:[4,5],4:[5]}，代表「即」、「即將」和「即將來臨」都是被收錄在字典裡的詞。因此這個句子的分詞結果可能會是:

(b) 即/將/來臨/時

(c) 即將/來臨/時

(d) 即將來臨/時

透過每個詞在字典中的詞頻，判斷哪種斷詞結果可以有最大的機率。經過 Jieba，最後的斷詞結果為「即/將來/臨時」。

二、對於未登錄詞使用隱藏式馬可夫模型(Hidden Markov Model, 以下簡稱 HMM)，找出最佳斷詞結果。

在進行第二步驟時，若發現有未登錄詞，也就是不存在於字典中的詞，會先給予字典中最低的詞頻進行 DAG。隨後使用 HMM 找出最適合的結果。Jieba 將字分成四種類型，分別是開始(B)、結束(E)、中間(M)和獨立成詞(S)，再經過大量的資料訓練後得到三個機率表(1)位置轉換機率表(2)位置到單詞轉換機率表(3)詞語開頭機率表，這三張表都在程式檔中叫 finalseg 的資料夾中，作用是在找出未登錄詞屬於哪種類型，根據類型就能對未登錄詞進行斷詞。

(三)CKIP、Jieba 比較

本研究使用模擬使用者回覆的文件檔如表 2-2，對兩個中文斷詞系統進行比較。結果顯示如表 2-3。

表 2-2 測試文檔

username	內容
Alex	物價又上漲了，工資卻沒漲
Kevin	明天還要上課，好想 休假阿~
Grace	最近的都不敢去菜市場了，都是因為 Covid-19
Vincent	我讀書讀得好累喔，每天都熬夜

表 2-3 斷詞結果比較

username	CKIP 斷詞結果	Jieba 斷詞結果
Alex	物價/又/上漲/了/，/工資/卻/沒/漲	物價/又/上漲/了/，/工資/卻/沒/漲
Kevin	明天/還/要/上課/，/好/想//休假/阿~	明天/還要/上課/，/好/想//休假/阿/~
Grace	最近/的/都/不/敢/去/菜市場/了/，/都/是/ 因為/Covid-19	最近/的/都/不/敢/去/菜市場/了/，/都/是/ /因為/Covid-/19
Vincent	我/讀書/讀/得/好/累/喔/，/每/天/都/熬夜	我/讀書/讀/得/好/累/喔/，/每/天/都/熬夜

根據實驗結果，CKIP 的斷詞結果比 Jieba 好，其原因是 Jieba 使用的字典是用中國簡體直接翻譯成繁體，但實際上，兩種語言的文字結構本身就不同。在時間的表現上 CKIP 的運行時間為 12.5 秒，Jieba 則為 0.534 秒。Jieba 的效率高了將近 23.4 倍。考慮到 CKIP 的運行時間太長，本研究將使用 Jieba 作為分析使用者輸入的斷詞系統，看中其高效率的斷詞速度，可以使機器人的回覆時間較快，另外 CKIP 將作為本研究後台進行模型訓練的斷詞系統，因為模型訓練不注重運行時間而較為注重斷詞精確度。

第四節 K-Nearest Neighbors 演算法

K-Nearest Neighbors 演算法(以下簡稱 KNN)為一種高效能的演算法，適用於分類和迴歸分析之的問題，最初由 Evelyn 和 Joseph Lawson Hodges jr 提出，後來 [\(Cover & Hart, 1967\)](#)發表論文才正式介紹了 KNN，KNN 的概念是計算測試樣本與所有訓練樣本之間的歐式距離，其中 K 為最接近此測試樣本的 K 個訓練樣本，最後在 K 個訓練樣本中以多數決的方式作為分類準則。以圖 1 為例，其中方型類別(或屬性)未知，三角形及圓形分別代表一種類別，當 K=3 時，因為最接近方型

的三個樣本是兩個圓形和一個三角形，則方形將被歸類為圓形，當 $K=5$ 時，最接近方型的五個樣本為三個三角形和兩個圓形，因此方形將被歸類為三角形。如圖 2-6 所示。

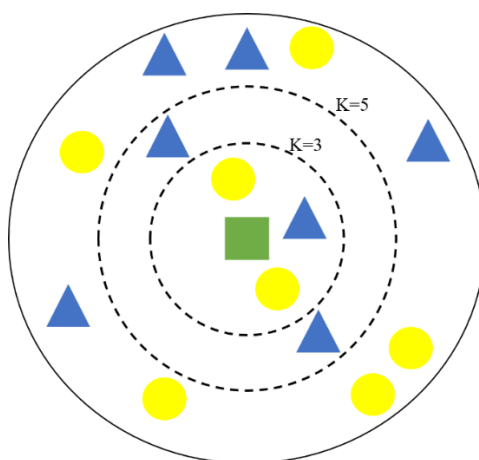


圖 2-6 KNN 圖示

在大數據時代 KNN 已被活用在各項實驗中，也演化出結合其他工具已達到降低成本、提高效率、提升正確率的演算法，例如(Zhang, Li, Zong, Zhu, & Wang, 2017)為了選擇每個測試樣本的最佳 K 值以達到降低運行成本並提高分類性能，提出兩種新的 KNN 分類算法，即 $kTree$ 與 $k*Tree$ ，證實其施行的實驗確實降低成本及提高效率；而(Deng, Zhu, Cheng, Zong, & Zhang, 2016)建議先將大數據進行 k -means(k -均值演算法)分群，再對每個群進行 KNN 分類，最後的結果證明了 KNN 的準確率及高效能。

本研究欲將問卷收集得到的資料集與使用者的性別、年齡及邁爾斯—布里格斯性格分類指標進行 KNN 演算法找出與使用者最相近的人，當使用者輸入與問卷中提及相關負面情緒的語句，即利用 KNN 的結果來回覆使用者。

第五節 邁爾斯—布里格斯性格分類指標

本研究透過 MBTI 了解使用者的人格類型。MBTI 人格測驗由四個面向構成八種不同類型，每個面向由兩種相反類型所組成，其中包含外向(Extroversion)與內向(Introversion)、知覺(Sensing)與直覺(Intuition)、理性(Thinking)與感性(Feeling)、判斷(Judging)與感知(Perceiving)。藉由不同類型的配對組合，形成 16 種不同的人格類型，分別為 ISTJ、ISFJ、INFJ、INTJ、ISTP、ISFP、INFP、INTP、ESTP、ESFP、ENFP、ENTP、ESTJ、ESFJ、ENFJ、ENTJ，每個不同人格類型都有屬於

其性格和特質。測驗試題引用《應用 MBTI 性格量表探討消費者對不同比例矩形之偏好》[\(王靜儀，2015\)](#)，共計 70 題。其中 EI 性格有 10 題，SN、TF、及 JP 各 20 題，每一題有 a 與 b 兩種選項，兩種選項分別對應同面向中兩種不同類型，以二選一的方式來計算，於最後測驗完畢後，計算兩種類型的題數。假設 EI 面向題目 10 題中，a 選項對應外向 E、b 選項對應內向 I，此面向計算結果外向 E 分數為 8、內向 I 分數為 2，則此面向為 E，其餘三個性格面向計算方式亦然，因此最後得出屬於該使用者的人格類型。本研究將使用 MBTI 做為使用者使用聊天機器人前的人格測驗，以供後續選定聊天機器人的模型進行自然語言模型的訓練。

第參章 研究方法

本章將逐步說明本研究所使用的方法。第一節將說明本研究的架構，第二節將說明本研究之研究流程，第三節將講本研究新創之語言風格模型及資料預處理，第四節講述中文斷詞，第五節講述 word2vec，第六節講述 KNN 演算法，第七節講述 LIWC 辭典分析使用者之情緒。

第一節 研究架構

本研究之核心目的對應至第壹章第三節研究目的。為使聊天機器人與使用者達到語言風格相似及改善聊天機器人回覆制式化之問題，故做了以下三個研究：

(一)Word Embedding

Word Embedding 是一種將字詞以空間向量表示的技術，目前廣泛被使用在自然語言處理上，最常被提及或使用的相關技術為 Word2Vec，在第貳章第二小節已提及。本研究在(二)訓練語言風格模型階段使用 Word Embedding 是為了能找到兩個“用法”相似的詞，Word2Vec 在某種程度上可以勝任這個角色。但 Word2Vec 是使用上下文去賦予目標字詞詞向量，而一個詞的用法主要是受上下文詞性的影響，因此在 Word2Vec 的詞向量並不能完整的代表一個詞的“用法”。所以本研究提出了新的 Word Embedding 技術稱之為 Word2FunctionVec。主要的差別在於，Word2FunctionVec 使用的是上下文的詞性去賦予目標詞詞向量，如此一來，Word2FunctionVec 的詞向量就更能代表一個詞的“用法”。關於 Word2FunctionVec 更多的細節將在第四節的(c)小節說明。

(二)句子結構模型

本研究提出了句子結構模型，使機器人可以學習、模仿某位作者的語言風格，也就是某位作者的說話方式。目前此模型的功能設計為可以更改一句話的語言風格，也就是將一句話以特定的語言風格換句話說，更多詳細的細節將會在此章第四節的第(d)小節說明。

(三)word2vec 調整句子向量詞性權重

本章第一小節(I) Word Embedding 中提及本研究使用 Word2Vec 將字詞以向量的形式表達語意，本研究發現不同詞性之詞向量對於原語句之句向量影響程度

不同，故為弱化語助詞對於原語句向量之影響，故針對不同詞性給予不同權重，如何針對詞性調整句向量權重之細節將於第三節(δ)S2VecQ 資料集說明。

第二節 研究流程

本研究研究流程將分成三個階段，(一)資料預處理、(二)訓練語言風格模型及(三)聊天語句分析及產生回覆語句，研究流程圖如下圖 3-1。第一階段資料預處理，將會在此章節的第三幾小節進行說明，其中包括(α)一般答覆資料集(以下簡稱 OrgData)及(β)負面情緒資料集(以下簡稱 NegData)，最後產生(γ)聊天答覆資料集(以下簡稱 Dataset)。第二階段訓練語言風格模型，將會在此章節的第四幾小節進行說明，其中包括(a)語料庫、(b)N-gram 模型、(c)Word Embedding 模型(d)句子結構模型及(f)多個具有不同語言風格的聊天答覆資料集(以下簡稱 LsmData)。第三階段將分析使用者輸入之聊天語句(以下簡稱 Input)及產生聊天回覆語句，此章節的第五幾小節將進行說明，其中包括首先，(1-1)計算 S2VecQ 與 Input 間的餘弦相似度、(1-2) 回覆使用者、(2-2) 使用 LIWC 辭典分析使用者的情緒、(2-3)判斷使用者有無負面情緒傾向及(2-4)聊天機器人主動關心使用者。

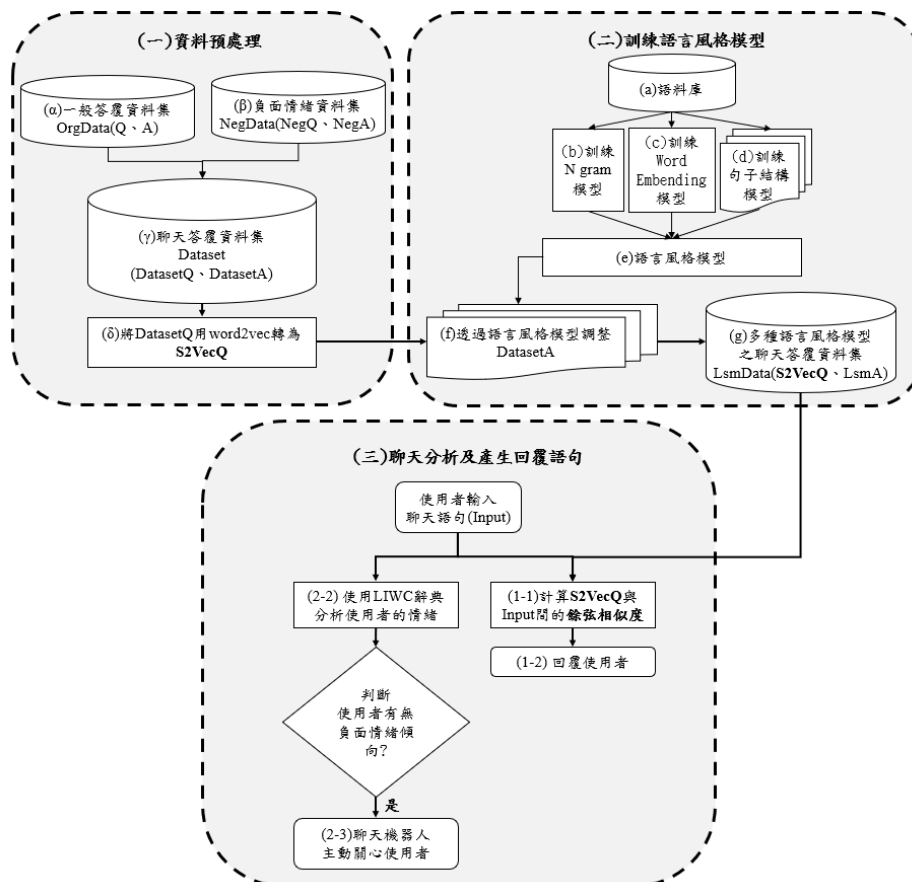


圖 3-1 研究流程圖

第三節 資料預處理

本研究所使用之聊天答覆資料集分為一般答覆與負面情緒聊天資料集。

(α)一般答覆資料集

一般答覆資料集由本研究之研究者與朋友的日常聊天對話資料集及青云客網絡商業公司之聊天資料集所組成。選用青云客網絡商業公司之聊天資料集的原因為此資料集的對話較生活化，有如朋友日常聊天一般，符合本研究欲達之成目的。兩資料集皆以一問一答的方式呈現，將資料欄位分為問題(以下簡稱 Q)及答覆(以下簡稱 A)，如表 3-1 示意：

表 3-1 一般答覆資料集之部分資料

ID	Q	A
1	喜歡你呀	謝謝你喜歡我
2	最近好常熬夜	這樣對身體不好
3	答對了	對了有獎嗎？
4	感冒一直都不好	身體不舒服就快去看醫生啊，不能拖的
5	感覺其中有蹊蹺	別想那麼多

因為原始聊天機器人交流群之資料集的 Q 皆不相同，因此資料集的主要修改依據為在 Q 不變的情況下，將 A 修改為合理的回覆，若原始回覆合理則不做更改，如表 3-2。以此方式能使 Q 在不重複的情況下持續增多，確保資料集 Q 的多元性，如表 3-3。

表 3-2 原始聊天機器人交流群資料集部分資料

ID	Q	A
1	我失望了	傷心，期望越高失望越大
2	我打噴嚏了	有人想你咯
3	真沒創意	葫蘆領孫悟空來了
4	祝您一路順風	天地間，世間萬物都有它存在的道理，比如我
5	真的好可愛	當然可愛呀

表 3-3 修改後聊天機器人交流群資料集部分資料

ID	Q	A
1	我失望了	傷心，期望越高失望越大
2	我打噴嚏了	有人想你咯
3	真沒創意	我覺得很有創意耶

ID	Q	A
4	祝您一路順風	謝謝你，下次見
5	真的好可愛	當然可愛呀

(β)負面情緒資料集

負面情緒資料集是經由本研究所設計之問卷進行蒐集，經過篩選後以回覆擁有相同 MBTI 使用者的相似問題。首先，本研究設計的問卷名稱為「針對不同壓力來源關心他人之調查」，將壓力分為六種面向，包含學業、家庭、愛情、友情、工作、健康等六種，再將不同壓力面向以情境題的方式細分為較普遍的問題，共三十五題。以學業面向為例：父母對孩子期望高、對自己沒有自信、上課不懂老師說什麼，跟不上進度、很努力念書，成績卻沒有起色，沒有天分、對自己所學沒有興趣，以上皆是壓力來源為學業面向所細分的问题，如表 3-4：

表 3-4 壓力來源為學業面向之問題

ID	Q
1	我已經很努力讀書了，但還是達不到爸媽的期望，怎麼辦...
2	我總是覺得自己比不上別人，可能我一輩子就這樣了吧...
3	每次老師點到我，我都回答不出問題，其他同學都會，我好爛...
4	我明明每天都熬夜讀書了，還是考得好差，我是不是真的沒有天分？
5	我發現我對現在學的東西真的完全沒興趣，唸得好痛苦...

其次，對於問卷的答題方式，本研究設計只要能感同身受，即可寫下相應回答，不須擁有相似經驗。本研究之初衷是希望讓使用者在聊天過後，負面情緒能得到舒緩，因此在篩選負面情緒資料集時，本研究會將較負面的回覆刪除，以確保使用者在抒發心情過後得到的回覆都是正面且有益於改善心情的。

最後，本研究計算每個 MBTI 人格類型於各問題的回覆數量是否與 MBTI 人格類型的分佈比例達一致，讓不同 MBTI 人格類型的使用者都能得到不同的回覆。

(γ)聊天答覆資料集

本研究之聊天答覆資料集由一般答覆資料集與負面情緒聊天資料集所組成。由於原來的負面情緒聊天資料集，只有 35 題，也就是 35 個 Q，因此為避免負面情緒聊天資料集，因筆數過少，在選取回覆時遭到忽略，而使有相關壓力來源之使用者無法得到更貼近之回覆，本研究將負面情緒聊天資料集的 35 個 Q 改以不

同方式詢問但仍然保留原有語意，藉此來增加負面情緒聊天資料集之 Q 的筆數。同時，也透過蒐集本研究設計之「不同壓力來源問題以不同方式詢問但保留原有語意之問卷」，以達到相同之目的，如表 3-5。最後，將一般答覆資料集與增加 Q 後之負面情緒聊天資料集合併為聊天答覆資料集。

表 3-5 將 Q 改以不同方式詢問但保留原有語意之範例

原 Q	以不同方式詢問
我已經很努力讀書了，但還是達不到爸媽的期望，怎麼辦...	每天已經很認真在唸書了，但爸媽的要求真的好難達到，好痛苦...
我總是覺得自己比不上別人，可能我一輩子就這樣了吧...	我是不是真的比不上別人，我真的覺得自己好爛喔
每次老師點到我，我都回答不出問題，其他同學都會，我好爛...	被老師點到的時候我都不會其他人都知道我真的好爛
我明明每天都熬夜讀書了，還是考得好差，我是不是真的沒有天分？	我都每天都熬夜讀書了還是考得好差到底還要我怎樣？
我發現我對現在學的東西真的完全沒興趣，唸得好痛苦...	最近唸書念得好痛苦，真的對現在學的東西沒有興趣

(δ)S2VecQ 資料集

在研究過程中發現，若使用者 Input 與 DatasetQ 之語句相差一個語助詞之語句，計算餘弦相似度後，使用者 Input 與 DatasetQ 餘弦相似度最大之值，不會為捨棄語助詞之語句。例如：使用者 Input 未加語助詞時為：「今天天氣真好」，資料集 DatasetQ 中與使用者 Input 最相似之語句為「今天天氣真好」；但若加上語助詞，如：「今天天氣真好喔」，資料集 DatasetQ 中與使用者 Input 最相似之語句為「今天好累喔」。如下表 3-6 Input 與 DatasetQ 相似度對應示意表所示。

表 3-6 Input 與 DatasetQ 相似度對應示意表

Input	DatasetQ	DatasetA
今天天氣真好	今天天氣真好	心情也跟著好起來
今天天氣真好喔	今天好累喔	帶你吃豪華大餐

原因為加語助詞語句之句向量跟未加語助詞語句之句向量餘弦相似度相差大，導致挑選出最相似之 DatasetQ 與原語句不同。其對應到之回覆語句 DatasetA 與使用者 Input 沒有關聯，所以最後產出的聊天回覆語句與使用者 Input 沒有邏輯。例如：DatasetQ「今天天氣真好」對應到的 DatasetA 為「心情也跟著好起來了」；

而 DatasetQ 「今天好累喔」對應到的 DatasetA 為「帶你吃豪華大餐」。所以當使用者 Input:「今天天氣真好喔」，將產生的聊天回覆語句為:「帶你吃豪華大餐」。

本研究認為語助詞之詞向量不可以直接設定為 0，因為「吃晚餐」跟「吃晚餐嗎」為兩種語意，前者為直述句，後者為問句。因此本研究為了弱化語助詞對句向量的影響，及強化動詞、名詞之詞向量對句向量的影響，針對不同詞性詞向量在句向量中權重的設置做超參數校調的實驗。

為了弱化語助詞及強化動詞、名詞對句向量形成的影響，本研究將 LIWC 辭典中語助詞類別的 34 組詞彙挑選出來，首先先將語助詞使用 word2vec 轉換為詞向量，再來將所有詞彙之向量長度統一為 1，而本研究為了依照 34 組語助詞向量角度做分群找尋最適合之群集數，使用了階層式分群(Hierarchical Clustering)，其中包括沃德法(Ward's method)、單一連結聚合演算法(single-linkage agglomerative algorithm)、平均連結聚合演算法(average-linkage agglomerative algorithm)及中心連結聚合演算法(centroid-linkage agglomerative algorithm)。分群結果如圖 3-2:

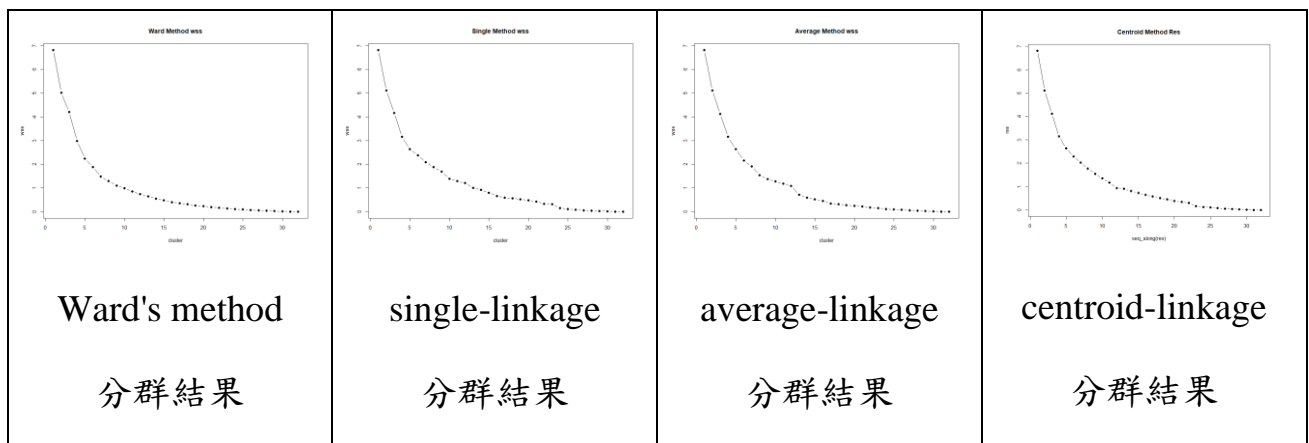


圖 3-2 階層式分群之分群結果

分群後計算群集中語助詞間向量組內距離平方和，找到最適之群集數，結果為合沃德法之群集數量為 6 群、單一連結聚合演算法之群集數量為 5 群、平均連結聚合演算法之群集數量為 6 群及中心連結聚合演算法之群集數量為 5 群。

向量長度越長表示影響句向量之因素越大，故接下來本研究取出群集中向量長度最長之詞彙作為群集之代表詞，並將群集代表詞彙兩兩計算餘弦相似度最後取平均，其數值代表其方法，最後本研究選取餘弦相似度最低之分群方法作為分

群依據，表 3-7 為實驗結果，故挑選沃德法為分群方法，將 34 組語助詞分為 6 群，其中 6 個代表詞彙分別為：「也、了、嗎、嗯哼、的」。

表 3-7 不同分群方法群集間餘弦相似度數值

分群方法	群集間餘弦相似度數值
沃德法	0.22834
單一連結聚合演算法	0.24159
平均連結聚合演算法	0.24757
中心連結聚合演算法	0.31857

語助詞詞向量之權重設定為 0.4、0.5、0.6，動詞及名詞之詞向量權重設定為 1 至 3(間隔為 0.25)，將資料集 14713 筆資料分別加上上述之 6 種語助詞，計算與資料集中 14713 筆資料未加語助詞之語句間的餘弦相似度，若加語助詞後之語句和未加語助詞之語句間的餘弦相似度值仍為 14713 筆餘弦相似度資料中最大，則標記為正確，否則標記為錯誤，再計算正確率，最後選取正確率最高的參數作為調整詞性權重設置之參數。例如當語助詞權重設定為 0.6，動詞、名詞權重設定為 1，其餘詞性維持 1，則「今天/天氣/真好/喔」這句話「今天」權重為 1，「天氣」權重為 1，「真好」權重為 1，「喔」權重為 0.6，將 word2vec 演算法中得到之詞向量乘以權重後並加總即為此句話之句向量。

超參數校調的實驗結果為設定語助詞權重為 0.6，動詞、名詞權重為 1 時，正確率相較於其他權重設置高，故詞向量在加總為句向量時將會依照詞性給予上述所定義的句向量權重，得出句向量後，將會與 DatasetQ 計算餘弦相似度，並選取餘弦相似度與 Input 最高之 DatasetQ。

第四節 訓練語言風格模型

語言風格模型的目的在於，能讓機器人回覆多樣化以及能回覆的像是使用者的朋友，其中回覆的像是使用者的朋友是根據 LSM 文獻探討中提到的，朋友間的語言風格會較為相似的特性。語言風格模型是由三個模型組合而成的包含了 N-gram、Word2Vec 及句子結構模型，其中句子結構模型為本研究創新的模型。N-gram 以及 Word2Vec 已第貳章第二、三小節詳細說明；在本研究中，會使用所有蒐集到的文本當作兩個模型的語料庫進行訓練。此小節將著重介紹句子結構模型及如何挑選訓練句子結構模型的文本。

(a)語言風格語料庫

本研究以來自不同作者之網路文章，作為提供模型訓練之語料庫。因句子結構模型需具備不同的語言風格，為了確保取樣之作者間的語言風格具差異性，故本研究利用 LSM 提取八種語言向度，包含特定人稱代名詞、非特定人稱代名詞、助動詞、副詞、介係詞、連接詞、否定詞及概數詞，計算兩兩作者間的語言風格，確認其具足夠差異性再以其語料作為模型訓練之語言風格語料庫。

LSM 的運算如下：首先將兩篇對話(文本)進行 LIWC 分析，並針對上述提及之八種語言向度，分別計算在兩作者語料間的使用率相似度。以介係詞為例，如公式(7)：

$$LSM_{preps} = 1 - [abs(preps1 - preps2)/(preps1 + preps2 + 0.001)] \text{公式(7)}$$

- LSM_{preps} 為兩作者間介係詞使用率相似度
- abs 為絕對值函式
- $preps1$:第一位作者語料的介係詞使用率
- $preps2$:第二位作者語料的介係詞使用率

依照公式(7)個別計算八種語言向度在兩作者語料間的使用率相似度後，將使用率相似度加總除以語言向度數，得兩作者間語言相似度指標。如公式(8)

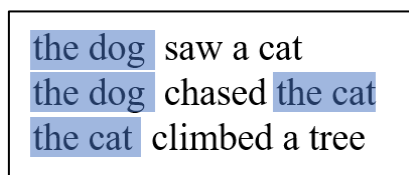
$$LSM = \frac{\sum_{i=1}^8 LSM_{preps_i}}{8} \text{公式(8)}$$

- LSM 為兩作者間語言風格相似度指標
- LSM_{preps_i} 為兩作者間各功能詞使用率的相似度

(b)N-gram 模型

N-gram 模型是一項常被用於自然語言處理的工具，用於計算文本中 N 個詞連在一起出現的機率，常見 N 的設置有 Bigram(N = 2)、Trigram(N = 3)。以下面這個語料庫為例：「the dog saw a cat」，「he dog chased the cat」，「he cat climbed a tree」。若使用 Bigram 如圖 3-3，可以得出「the」後面接「dog」的機率為 50%，

接「cat」的機率為 50%。若使用 Tri-gram 則可得出「the」後面接「dog saw」的機率為 33.33% 以此類推。



the dog saw a cat
the dog chased the cat
the cat climbed a tree

圖 3 - 3Tri-gram 舉例

本研究新增了紀錄詞性的功能在 N-gram 模型裡，舉例來說，在本研究的 N-gram 模型中，「the」後面可以接「dog」，其機率為 50% 並且記錄「dog」為名詞。此功能的目的是以及 N-gram 模型如何運用在本研究的語言風格模型將會在(e) 小節闡述。

(c) Word Embedding 模型

本研究使用到的 Word Embedding 技術有 Word2Vec 以及 Word2FunctionVec，Word2Vec 已經在文獻探討以及第三章第一節的(I)討論過，並且說明過為何 Word2Vec 的詞向量不能完整的代表一個詞的用法。此小節將著重討論 Word2FunctionVec，Word2FunctionVec 的目的在於賦予語料庫中每個詞一組空間向量，並且當兩個“用法”相近的詞其向量的夾角將會很低。以下以圖 3-3 的句子為語料庫進行說明，取「climbed」與「chased」為目標詞，其上下文分別是「dog、the」以及「cat、a」。在 Word2Vec 中，「climbed」的詞向量將被「dog、the」定義，而「chased」的詞向量則被「cat、a」定義，因此「climbed」與「chased」在 Word2Vec 的詞向量角度將會差距很大，甚至可以說兩個目標詞之間沒有相關性。但在本研究提出的 Word2FunctionVec 中，「climbed」的詞向量將被「dog、the」的詞性定義，也就是「名詞、冠詞」，而「chased」的詞向量則被「cat、a」的「名詞、冠詞」定義，因此在 Word2FunctionVec 中「climbed」與「chased」之間的夾角將為 0，也就是完全相同。因此 Word2FunctionVec 賦予的詞向量能完整地提供一個詞的“用法”，而 Word2Vec 的詞向量雖然完全以“用法”定義詞向量，但能賦予詞向量更多的含意。具體如何使用 Word2Vec 或 Word2FunctionVec 將在第(e) 小節進行說明。

(d) 句子結構模型

句子結構模型為學習同一作者使用的句子結構，訓練模型的語料庫皆來自同一位作者。句子結構的定義為(圖 3-4)，句子是由單詞組成的，再把句中的單詞進行斷詞並分析其詞性，則可以得知一句子由什麼詞性及順序組合而成。

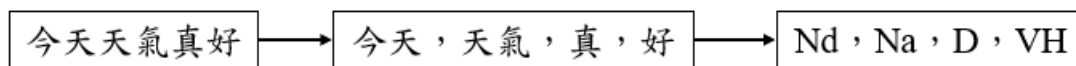


圖 3-4 句子結構示意圖

句子結構模型主要是學習一個作者如何使用句子的，與語言風格分析(LSM)一樣，將詞性當作重要因素去分析一位作者，當訓練完成後，機器人即可以模仿作者之語言風格。訓練的方法是對來源於同一作者的語料庫進行預處理，以標點符號進行斷句並且斷詞。模型將紀錄作者所使用到的所有句子結構類型，並且會計算各個句子結構的出現次數。在(e)小節將會闡述如何將句子結構模型運用於語言風格模型中。

(e) 語言風格模型

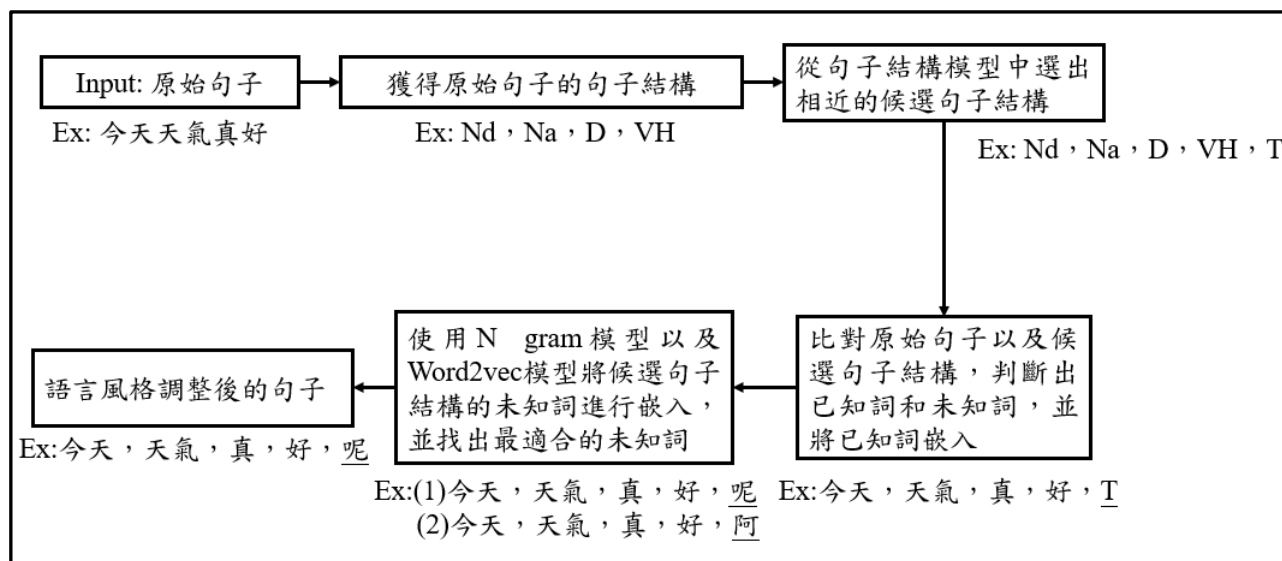


圖 3-5 語言風格模型架構圖

語言風格模型使用到(b)、(c)和(d)小節的三種模型，目的為改變句子之語言風格，即是把新字詞加入原語句中，或是刪除一些原語句之詞，以改變原語句的語言風格。可以理解為將原語句換句話說，並且使用的是句子結構模型的語言風格。為了評估改變後的語句是否具邏輯性及是否能表達原語句之語意，因此需要一個

指標進行合理性的評估，關於指標的設置將於第肆章進行說明。圖 3-5 為語言風格模型之架構圖。演算法 3-1 為語言風格模型更改語句之語言風格的虛擬碼。

演算法 3-1 語言風格更改之演算法

Input <i>S1</i>	
1	對 <i>S1</i> 進行斷詞並且標註詞性生成 <i>S1_struct</i>
2	For <i>S_struct</i> in <i>Struct_model</i> :
3	If <i>S_struct</i> 與 <i>S1_struct</i> 相似:
4	比對 <i>S_struct</i> 和 <i>S1_struct</i> 判斷未知詞並將已知詞嵌入
5	使用 <i>N_gram</i> 模型及 <i>Word embedding</i> 技術嵌入未知詞生成 <i>S2</i>
6	<i>Lsm_alter_S</i> 紀錄 <i>S2</i>
Output <i>Lsm_alter_S</i>	
●	<i>S1</i> 為欲更改語言風格之語句。
●	<i>S1_struct</i> 為 <i>S1</i> 之句子結構。
●	<i>Struct_model</i> 為句子結構模型。
●	<i>S_struct</i> 為句子結構模型中紀錄的句子結構。
●	<i>S2</i> 為 <i>S1</i> 更改語言風格之語句。
●	<i>Lsm_alter_S</i> 記錄更改後的語句。

此段落將以圖 3-4 即演算法 3-1 進行語言風格模型運作的示範，語言風格模型的輸入為一個句子稱為原始句子 *S1*，輸出則為語言風格調整後之句子，輸出可能為一個或多個句子，因此已陣列的方式輸出 *Lsm_alter_S*。以「今天天氣真好」為 *S1* 舉例，首先(演算法 3-1 的第 1 行)，獲取原語句之結構 *S1_struct* 為「Nd，Na，D，VH」，(演算法 3-1 第 3 行)再從句子結構模型中獲得相似的候選句子結構，通常候選句子結構會有多種可能。接著，(演算法 3-1 第 4 行)將比對原始句子與候選句子，並把已知詞填入對應之結構位置，如「今天，天氣，真，好」為已知詞，將對應到 *S_struct* 「Nd，Na，D，VH，T」的無底線之部份，而底線部分為新加入之未知詞，因此語句將變成「今天，天氣，真，好，T」。再來是未知詞的預測(演算法 3-1 第 5 行)，本研究使用 *N-gram* 及 *Word embedding* 兩個技術結合已找到最適合嵌入的未知詞。如欲判斷「好」後面接「T」詞性的候選未知詞有哪些可能，將使用 *N-gram* 進行判斷並計算候選未知詞的機率。以下例說明，

N-gram 模型判斷「好」後面接「T」的詞有「呢」及「阿」，兩者接在「好」後面的機率分別是 10%跟 5%，因此判斷「呢」適合嵌入，以「今天天氣真好呢」作為 *S2* 記錄到 *Lsm_alter_S*。若在語料庫中「好」後面沒有接「T」詞性的情況，將使用 Word embedding 尋找與「好」最相似之詞，並同樣使用 N-gram 進行判斷並計算候選未知詞以及對應的機率。以下例說明，N-gram 模型判斷「好」後面沒有詞性為「T」的詞，因此使用 Word embedding 尋找與「好」最相似之詞，舉例來說，「不好」是「好」最相似之詞，而「不好」在 N-gram 模型中可以接詞性為「T」的詞，因此就以「不好」進行判斷並計算候選未知詞的機率，但實際上並不會把「好」替換成「不好」，只是使用「不好」進行機率判斷而已。

另一種情況為未知詞在已知詞前面，舉例來說，「今天，天氣，真，VH，呢」，這種情況需要判斷「VH」最適合嵌入什麼未知詞。首先，如上個段落使用「好」預測「T」一樣，這裡使用「真」預測「VH」。接著將使用「呢」，回去判斷「VH」的候選未知詞誰最好。假設用「真」預測「VH」發現「好」與「好棒」都可以當作候選未知詞，接著再 N-gram 模型檢查「好」後面接「呢」的機率以及「好棒」後面接「呢」的機率，並判斷何種組合的機率最大，並將最大機率的組合進行輸出。其中，若未知詞無法接「呢」會輸出 0，若所有候選未知詞都無法接「呢」，則使用 Word embedding 尋找未知詞後面能否接與「呢」相似的詞，若有則使用對應的機率作為輸出。舉例來說「好」與「好棒」都無法接「呢」，但使用 Word embedding 發現「好」後面可以接與「呢」相近的詞，像是「吧」，並且「好吧」的機率為 30%，因此將會以「好吧」的機率代替「好呢」進行判斷。最後輸出 *Lsm_alter_S*，作為 *S1* 在特定語言風格下的說法，以上述的例子來說，「今天天氣真好」在特定語言風格下能被更改成「今天天氣真好呢」或是「今天的天氣真好」。

(f)透過不同語言風格挑整 DatasetA

在本研究中，將會對一般答覆資料集的回覆(DatasetA)進行語言風格的調整，獲得不同語言風格的一般答覆資料集。並且對於同一個回覆，通常能產出多個語言風格調整後的回覆，因為候選句子結構通常會有多個。如此一來就可以根據使用者的語言風格去選取相似的語言風格資料集，使機器人能回覆的像使用者的朋友，另外對於同一個問題能有不同句子結構的回覆。

(g)多種語言風格之聊天答覆資料集 LsmData

語言風格模型會因為句子結構模型使用的語料庫不同，得到不同的語言風格，因此本研究使用多個句子結構模型，產生各種語言風格的語言風格模型，並且每種語言風格模型都會進行(f)，因此就能獲得多種語言風格之聊天答覆資料集 LsmData。

第五節 聊天及產生回覆語句

(1-1) 計算 S2VecQ 與 Input 間的餘弦相似度

本研究為了要分析使用者輸入的聊天語句(以下簡稱為 Input)，故利用 Jieba 中文斷詞系統將使用者 Input 進行斷詞，如表 3-8 所示：

表 3-8 使用者 Input 及 Jieba 斷詞結果

Input	我把他當成最好的朋友，他卻誣賴我，我好失望好難過，心情好低落。
斷詞結果	我/把/他/當成/最好/的/朋友/，/他/卻/誣賴/我/，/我/好/失望/好/難過/，/心情/好/低落/。

經過 Jieba 斷詞後，本研究將使用 word2vec 將每一字詞轉換成詞向量，以向量表達其文字語意，為找尋適合本研究資料集之向量維度，本研究作了以下的實驗。

首先，本研究使用以新聞為例及第三小節(γ)所提及之聊天答覆資料集所組合而成之資料集，挑選出資料集中詞頻最大的前 10 個字詞，包含:的、桃園、我、市長、你、及、鄭、市府、也、在，倆兩字詞合併為一組合形成 45 個組合。再使用 word2vec 將字詞轉換成 5 至 500 維度之向量，計算每組合字詞間的餘弦相似度。同時也請本研究小組之研究員以分數 1 至 10 人工註記兩字詞間的相似度，若人工判定兩字詞相似度高則給予較高之分數，反之則給予較低之分數，並取平均值作為人工註記相似度代表值。如下表 3-9。

表 3-9 兩字詞間的相似度註記

詞 1	詞 2	研究員 1	研究員 2	研究員 3	研究員 4	研究員 5	研究員 6	研究員 7	研究員 8	average
的	桃園	1	1	1	1	1	1	1	1	1
的	我	2	1	1	1	1	1	1	1	1.125

詞 1	詞 2	研究員 1	研究員 2	研究員 3	研究員 4	研究員 5	研究員 6	研究員 7	研究員 8	average
的	市長	1	1	1	1	1	1	1	1	1
的	你	1	1	1	1	1	1	1	1	1
的	及	5	3	8	1	1	3	1	3	3.125
的	鄭	2	1	1	1	1	1	1	1	1.125
的	市府	1	1	1	1	1	1	1	1	1
的	也	4	3	8	1	1	7	1	3	3.5
的	在	4	3	8	1	1	7	1	3	3.5
桃園	我	2	1	6	1	1	4	2	1	2.25
桃園	市長	5	3	6	2	5	4	2	1	3.5
桃園	你	5	1	6	1	1	4	2	1	2.625
桃園	及	1	1	1	1	1	1	1	1	1
桃園	鄭	1	1	1	1	1	3	1	1	1.25
桃園	市府	5	2	8	2	5	6	5	3	4.5
桃園	也	1	1	1	1	1	1	1	1	1
桃園	在	1	1	1	2	1	1	1	1	1.125
我	市長	5	1	10	1	3	6	1	3	3.75
我	你	10	10	10	5	8	8	10	7	8.5
我	及	4	1	1	2	1	1	1	1	1.5
我	鄭	1	1	1	2	1	5	2	8	2.625
我	市府	3	1	9	2	1	4	2	1	2.875
我	也	6	1	1	2	1	1	1	1	1.75
我	在	1	1	1	2	1	1	1	1	1.125
市長	你	3	1	10	2	3	7	2	9	4.625
市長	及	1	1	1	1	1	1	1	1	1
市長	鄭	1	2	1	1	1	3	2	8	2.375
市長	市府	8	5	10	1	5	4	5	1	4.875
市長	也	1	1	1	1	1	1	1	1	1
市長	在	1	1	1	1	1	1	1	1	1
你	及	1	1	1	1	1	1	1	1	1
你	鄭	1	1	1	1	1	4	1	6	2
你	市府	3	1	9	1	3	6	2	1	3.25
你	也	1	1	1	1	1	1	1	1	1
你	在	1	1	1	1	1	1	1	1	1
及	鄭	1	1	1	1	1	1	1	1	1

詞 1	詞 2	研究員 1	研究員 2	研究員 3	研究員 4	研究員 5	研究員 6	研究員 7	研究員 8	average
及	市府	1	1	1	1	1	1	1	1	1
及	也	1	3	10	1	3	7	1	3	3.625
及	在	5	2	8	1	1	6	1	2	3.25
鄭	市府	1	1	1	1	1	2	2	1	1.25
鄭	也	1	1	1	1	1	1	1	1	1
鄭	在	1	1	1	1	1	1	1	1	1
市府	也	1	1	1	1	1	1	1	1	1
市府	在	1	1	1	2	1	1	1	1	1.125
也	在	6	2	8	3	1	6	1	2	3.625

再來計算 5 至 500 維度字詞間的 similarity 和人工註記字詞相似度間的 Pearson 相關係數，如下表 3-10，選取 Pearson 相關係數最高之維度，110 維度作為本研究資料集之向量維度。

表 3-10 不同維度下的 Pearson 相關係數

維度	Pearson	維度	Pearson	維度	Pearson	維度	Pearson	維度	Pearson
1		101	0.511122	201	0.486164	301	0.481037	401	0.480385
2		102	0.494111	202	0.522899	302	0.498177	402	0.47673
3		103	0.497715	203	0.480242	303	0.478013	403	0.469711
4		104	0.484934	204	0.453887	304	0.510152	404	0.504028
5	0.332381	105	0.500632	205	0.506112	305	0.459882	405	0.491115
6	0.296312	106	0.49767	206	0.513917	306	0.469711	406	0.479175
7	0.291171	107	0.538336	207	0.495436	307	0.511386	407	0.459217
8	0.235204	108	0.490316	208	0.499125	308	0.472934	408	0.482996
9	0.258465	109	0.498208	209	0.458314	309	0.45448	409	0.453345
10	0.319293	110	0.54899	210	0.508554	310	0.490132	410	0.508728
11	0.291562	111	0.504186	211	0.498815	311	0.458476	411	0.468432
12	0.343097	112	0.492543	212	0.479453	312	0.507705	412	0.463772
13	0.433928	113	0.532702	213	0.503395	313	0.479756	413	0.461846
14	0.440138	114	0.526949	214	0.461229	314	0.500068	414	0.506788
15	0.386334	115	0.513079	215	0.476972	315	0.439725	415	0.487908
16	0.529585	116	0.507923	216	0.460199	316	0.44916	416	0.462036
17	0.425595	117	0.480793	217	0.485428	317	0.443236	417	0.478434
18	0.390075	118	0.501898	218	0.508447	318	0.488202	418	0.47807
19	0.395298	119	0.524312	219	0.496983	319	0.470713	419	0.475006

維度	Pearson	維度	Pearson	維度	Pearson	維度	Pearson	維度	Pearson
20	0.458336	120	0.515893	220	0.50092	320	0.459044	420	0.481016
21	0.45228	121	0.496049	221	0.494824	321	0.497971	421	0.489564
22	0.468539	122	0.483086	222	0.492594	322	0.450685	422	0.475775
23	0.538852	123	0.52166	223	0.484746	323	0.503319	423	0.453864
24	0.4729	124	0.523078	224	0.49352	324	0.508256	424	0.466274
25	0.463063	125	0.476874	225	0.489598	325	0.501563	425	0.482556
26	0.496835	126	0.474785	226	0.507287	326	0.45434	426	0.473791
27	0.462813	127	0.482959	227	0.478119	327	0.468659	427	0.486788
28	0.449877	128	0.492582	228	0.465278	328	0.506925	428	0.471145
29	0.4923	129	0.478411	229	0.457325	329	0.488818	429	0.481944
30	0.545699	130	0.522132	230	0.47094	330	0.490332	430	0.476126
31	0.507781	131	0.511404	231	0.471709	331	0.490157	431	0.506167
32	0.489304	132	0.527967	232	0.498484	332	0.500713	432	0.480048
33	0.48468	133	0.468056	233	0.496348	333	0.455844	433	0.467065
34	0.46334	134	0.498573	234	0.473145	334	0.488917	434	0.479146
35	0.48421	135	0.481938	235	0.460416	335	0.507334	435	0.491723
36	0.485243	136	0.499984	236	0.480244	336	0.484108	436	0.497532
37	0.529267	137	0.504495	237	0.509143	337	0.486234	437	0.492976
38	0.49836	138	0.516244	238	0.495254	338	0.46967	438	0.476948
39	0.450025	139	0.5105	239	0.496746	339	0.477224	439	0.481631
40	0.51819	140	0.516008	240	0.489027	340	0.47895	440	0.490035
41	0.492671	141	0.50163	241	0.491519	341	0.473317	441	0.493189
42	0.506364	142	0.475144	242	0.475581	342	0.490541	442	0.471794
43	0.46093	143	0.493035	243	0.480573	343	0.462498	443	0.45538
44	0.478217	144	0.484136	244	0.508733	344	0.489383	444	0.493746
45	0.521407	145	0.478286	245	0.501871	345	0.462441	445	0.453228
46	0.534724	146	0.520212	246	0.476605	346	0.517011	446	0.465632
47	0.496756	147	0.481011	247	0.490358	347	0.469338	447	0.461096
48	0.516279	148	0.493846	248	0.455929	348	0.480753	448	0.491747
49	0.503113	149	0.491779	249	0.478721	349	0.489849	449	0.463299
50	0.483393	150	0.510742	250	0.474489	350	0.484885	450	0.466817
51	0.532659	151	0.489784	251	0.48266	351	0.467681	451	0.463697
52	0.525298	152	0.487726	252	0.506237	352	0.47818	452	0.498092
53	0.474574	153	0.475757	253	0.531281	353	0.477591	453	0.481666
54	0.463365	154	0.521553	254	0.486514	354	0.468689	454	0.48693

維度	Pearson	維度	Pearson	維度	Pearson	維度	Pearson	維度	Pearson
55	0.504883	155	0.500246	255	0.485791	355	0.48573	455	0.47798
56	0.500234	156	0.472037	256	0.486118	356	0.476008	456	0.46489
57	0.513388	157	0.501854	257	0.507291	357	0.521036	457	0.485462
58	0.472074	158	0.490678	258	0.473372	358	0.479623	458	0.475045
59	0.526671	159	0.519706	259	0.499488	359	0.479719	459	0.480037
60	0.481755	160	0.496506	260	0.475174	360	0.486014	460	0.485268
61	0.509105	161	0.466121	261	0.494319	361	0.483115	461	0.462247
62	0.487749	162	0.516904	262	0.504612	362	0.48987	462	0.479248
63	0.5076	163	0.490713	263	0.499616	363	0.441438	463	0.449211
64	0.503943	164	0.541423	264	0.492521	364	0.498453	464	0.499016
65	0.510171	165	0.503171	265	0.511868	365	0.467914	465	0.47354
66	0.512923	166	0.469023	266	0.47092	366	0.494006	466	0.485366
67	0.522253	167	0.476926	267	0.479593	367	0.481228	467	0.459666
68	0.510541	168	0.460943	268	0.488321	368	0.468554	468	0.455663
69	0.472723	169	0.526775	269	0.478858	369	0.478882	469	0.490042
70	0.493105	170	0.484359	270	0.490222	370	0.469763	470	0.483163
71	0.454474	171	0.493808	271	0.489682	371	0.471913	471	0.46152
72	0.48508	172	0.495853	272	0.498346	372	0.472946	472	0.440598
73	0.523986	173	0.481225	273	0.492277	373	0.495987	473	0.463722
74	0.501553	174	0.503827	274	0.511433	374	0.449175	474	0.492029
75	0.480834	175	0.491123	275	0.476508	375	0.494704	475	0.482707
76	0.535086	176	0.476842	276	0.488528	376	0.503292	476	0.470507
77	0.504921	177	0.488568	277	0.49378	377	0.485127	477	0.457507
78	0.482625	178	0.495185	278	0.493907	378	0.482901	478	0.489298
79	0.501977	179	0.484283	279	0.476386	379	0.45627	479	0.428022
80	0.493926	180	0.473078	280	0.489946	380	0.461223	480	0.478979
81	0.543276	181	0.482539	281	0.47518	381	0.493126	481	0.475086
82	0.486649	182	0.495689	282	0.490608	382	0.492961	482	0.489928
83	0.483041	183	0.479974	283	0.452781	383	0.497295	483	0.483669
84	0.491452	184	0.509442	284	0.492187	384	0.469528	484	0.446812
85	0.509102	185	0.500068	285	0.48695	385	0.477134	485	0.484293
86	0.497392	186	0.521953	286	0.505763	386	0.501532	486	0.440202
87	0.501856	187	0.489165	287	0.48223	387	0.453029	487	0.470197
88	0.52451	188	0.492919	288	0.474542	388	0.468954	488	0.498842
89	0.483712	189	0.498681	289	0.472668	389	0.446884	489	0.452616

維度	Pearson	維度	Pearson	維度	Pearson	維度	Pearson	維度	Pearson
90	0.485704	190	0.505779	290	0.506859	390	0.459257	490	0.487579
91	0.471182	191	0.45756	291	0.472008	391	0.508792	491	0.463935
92	0.518035	192	0.489019	292	0.472888	392	0.52919	492	0.488106
93	0.516066	193	0.502837	293	0.475519	393	0.470083	493	0.484103
94	0.541739	194	0.480693	294	0.530628	394	0.477183	494	0.467023
95	0.521341	195	0.485659	295	0.484324	395	0.466008	495	0.462141
96	0.470405	196	0.498889	296	0.497958	396	0.462311	496	0.4749
97	0.485234	197	0.500977	297	0.469223	397	0.473034	497	0.484735
98	0.513443	198	0.514375	298	0.462689	398	0.473769	498	0.459078
99	0.492085	199	0.503999	299	0.486932	399	0.459175	499	0.44258
100	0.498731	200	0.502245	300	0.530651	400	0.45976	500	0.473939

經由上面實驗結果，本研究將以 Input 使用 word2vec 將每一字詞轉換成一 110 維度之向量，並以詞向量之方式表達字詞的語義資訊。再將語句中每一字詞的詞向量依照詞性調整權重並加總，形成句向量，依詞性調整權重設置之方式將於第肆章說明。計算完使用者 Input 之句向量後，會再計算使用者 Input 之向量與 S2VecQ 向量間的餘弦相似度，餘弦相似度計算公式，如公式(9)，並以二維向量為例，如公式(10)。

$$\text{cosin} - \text{Similarity} = \cos \theta = \frac{A \cdot B}{|A| \cdot |B|} = \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}} \quad \text{公式(9)}$$

- A_i 、 B_i 分別代表 A 和 B 的各分量

$$\cos \theta = \frac{a \cdot b}{|a| \cdot |b|} = \frac{(x_1, y_1) \cdot (x_2, y_2)}{\sqrt{(x_1)^2 + (y_1)^2} \times \sqrt{(x_2)^2 + (y_2)^2}} = \frac{x_1 x_2 + y_1 y_2}{\sqrt{(x_1)^2 + (y_1)^2} \times \sqrt{(x_2)^2 + (y_2)^2}} \quad \text{公式(10)}$$

- $a = (x_1, y_1)$
- $b = (x_2, y_2)$

兩向量點距離若越近，則向量間夾角會越小，餘弦相似度值會越大；故本研究計算完使用者 Input 之向量與 S2VecQ 向量間的餘弦相似度後，會選取與使用者 Input 餘弦相似度最高之 S2VecQ 語句對應的 DatasetA 回覆使用者。

(1-2) 回覆使用者

本研究將利用 DatasetA 答覆使用者，在選取 DatasetA 前會判斷 S2VecQ 是否為本研究透過問卷收集到的 NegQ，若是，則會使用 K-Nearest Neighbors 演算法（以下簡稱 KNN），在負面情緒聊天資料集中挑選與使用者 MBTI 最為相似之問卷填寫者的 NegA 做回覆。本研究之 KNN 演算法利用歐幾里得距離公式計算使用者與問卷填寫者之距離，如公式(11)。

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} \quad \text{公式(11)}$$

- $d(x, y)$ = x 與 y 兩點間之距離
- x_n = x 之第 n 個屬性值
- y_n = y 之第 n 個屬性值

若使用者(x)MBTI 為 ENTP(x_1)，表 3-11 為使用者輸入之聊天語句，表 3-12 為聊天答覆資料集之部分資料，由上一小節提及計算使用者輸入之聊天語句與聊天答覆資料集的餘弦相似度之方法，得出使用者 Input 與表 3-13 第 4 條「每天都聽到爸媽在房間吵架，壓力真的好大...」有最高的餘弦相似度，則取得 NegQ1，NegQ1 為負面情緒資料集之語句。

表 3-11 使用者輸入之聊天語句

Input	每天都聽到爸媽在房間吵架，壓力真的好大...
-------	------------------------

表 3-12 聊天答覆資料集之部分資料

ID	答 Q	覆 A	餘弦相似度
1	你好	你好哇	0.87173
2	今天要吃甚麼呀	吃漢堡	0.88684
3	還要多久才會下課呀，好無聊	再撐一下	0.94258
4	每天都聽到爸媽在房間吵架，壓力真的好大...	NegQ23	1
5	晚安	晚安	0.900209

表 3-13 NegQ23 資料集之部分資料

ID	MBTI	NegA
1	ISTJ	聽音樂！
2	INFP	集合大家開一次家庭會議可能是個好方法
3	ISTJ	我一直都在
4	ENTP	可以試著跟他們談談,不然就是把事情的來龍去脈跟可以分享的人好好說出來
5	INFJ	嗯… 你可以多出門走走！

因為 KNN 演算法是利用歐幾里得距離公式計算，所以需將使用者資訊轉換成虛擬變數(以下稱為 dummy variable)的型態，MBTI 分為 EI、SN、TF 及 JP 四個面向，EI 面向中 E 以 0、I 以 1 表示；SN 面向中 S 以 0、N 以 1 表示；TF 面向中 T 以 0、F 以 1 表示；JP 面向中 J 以 0、P 以 1 表示，轉換結果如下表 3-14，ENTP 將會轉換成[0,1,0,1]。

表 3-14 使用者資訊轉換為虛擬變數 dummy variable 型態

E, I	S, N	T, F	J, P
0	1	0	1

並將 NegQ23 資料集中資料根據上述規則轉換成 dummy variable，結果如下表 3-15，最後再利用 KNN 計算使用者與哪一位問卷填寫者最接近。

表 3-15 使用者資訊轉換為 dummy variable

ID	E, I	S, N	T, F	J, P	NegA
1	1	0	0	0	聽音樂！
2	1	1	1	1	集合大家開一次家庭會議可能是個好方法
3	1	0	0	0	我一直都在
4	0	1	0	1	可以試著跟他們談談,不然就是把事情的來龍去脈跟可以分享的人好好說出來
5	1	1	1	0	嗯… 你可以多出門走走！

得出結果如表 3-16，使用者與 4 號問卷填寫者距離最小最為相似，則挑選表 3-15 中 4 號問卷填寫者的 NegA：「可以試著跟他們談談,不然就是把事情的來龍去脈跟可以分享的人好好說出來」回覆使用者。

表 3-16 使用者與各問卷填寫者之歐幾里德距離

ID	E, I	S, N	T, F	J, P	d(使用者, 問卷填寫者)
1	1	0	0	0	1.732050808
2	1	1	1	1	1.414213562
3	1	0	0	0	1.732050808
4	0	1	0	1	0
5	1	1	1	0	1.732050808

(2-2)使用 LIWC 辭典分析使用者的情緒

本研究採用 LIWC 字典分析使用者於聊天過程中是否有負面傾向。研究方法為蒐集 30 篇憂鬱症患者發布於網路之文章，並將此 30 篇文章分別透過 Jieba 斷詞，如表 3-17 中斷詞結果。排除不在 LIWC 字典之詞及詞性類別為人稱代名詞、非特指人稱代名、助動詞、常用副詞、介係詞、連接詞、否定詞、量詞等九類字詞，為分析之總詞數，如表 3-17 中總詞數。接著分析總詞數中屬於負面情緒之詞，最終計算出負面情緒詞個數佔總詞數個數之比例，如公式(12)。統計 30 篇文章(詳見附錄一)負面情緒詞比例之平均值為 0.0560，為本研究判斷使用者是否有負面傾向之準則，如表 3-17，負面情緒詞比例為 0.2857，將視為有負面傾向。

$$\text{負面情緒詞比例} = \text{負面情緒詞個數} / \text{總詞數個數} \quad \text{公式(12)}$$

表 3-17 計算負面情緒詞比例之範例

	分析過程	詞數
例句	我今天心情有點差壓力好大	
斷詞結果	我/今天/心情/有點/差/壓力/好/大	8
總詞數	今天/心情/有點/差/壓力/好/大	7
負面情緒詞	差/壓力	2
負面情緒詞比例	0.2857=2/7	

(2-3)聊天機器人主動關心使用者

於分析完使用者聊天紀錄後，一旦發現使用者有負面情緒，聊天機器人便會傳送關心訊息給使用者。

第肆章 實驗結果

本章節將敘述本研究所使用之回覆語句資料集、各演算法的結合及實驗的結果。第一節為資料收集，第二節為語言風格模型，第三節為本研究如何挑選適當回覆語句回覆使用者之流程，第四節將透過本研究設計之實驗驗證具有語言風格的聊天機器人回覆語句是否讓使用者覺得跟聊天機器人聊天和跟朋友聊天的感受是相同的。

第一節 資料收集

(一)一般答覆資料集

本研究之一般答覆資料集經過資料預處理後，共有 14713 筆，詳細資料來源已於第三章第三節說明。資料集中包括研究者與朋友的日常聊天對話資料集 708 筆及修改後的聊天機器人交流群資料集 14005 筆，如表 4-1：

表 4-1 一般答覆資料集分類

資料集名稱	筆數
研究者與朋友的日常聊天對話資料集	708
聊天機器人交流群資料集	14005
一般答覆資料集	14713

(二)負面情緒資料集

本研究經由問卷所收集並選用的負面情緒資料集共有 3687 筆，其中不同壓力面向之問題的收集筆數，如表 4-2：

表 4-2 負面情緒資料集各問題收集筆數

題號	Q	筆數
Q1	不管我怎麼修改，客戶都一直不滿意我的提案，到底想要我怎麼樣？	99
Q2	公司的同事都是為了利益在交朋友，讓我感覺壓力好大，跟他們的關係也不和睦...	101
Q3	好想我男/女朋友，下次見面要再三個月後了，他能每天在我身邊該有多好	116
Q4	好羨慕別人都有健康的身體，為甚麼人生這麼不公平，就我天生有殘缺...	101
Q5	我已經很努力讀書了，但還是達不到爸媽的期望，怎麼辦...	157
Q6	我只把他當朋友，他卻跟我告白，好尷尬...我們的友誼還可以繼續嗎？	104
Q7	我好喜歡他，可是他已經有男/女朋友了，我該怎麼處理這段感情...	100
Q8	我好想離職，但又找不到更滿意的工作...可能是我能力不足吧	106

題號	Q	筆數
Q9	我弟今天偷拿我的錢，被我發現居然還反過來罵我!一氣之下，我就動手打了他...	85
Q10	我把他當成最好的朋友，他卻在別人面前誣賴我，我真的好失望好難過...	109
Q11	我明明每天都熬夜讀書了，還是考得好差，我是不是真的沒有天分?	113
Q12	我爸媽很不喜歡我男/女朋友，叫我跟他快點分開，我該怎麼辦?	102
Q13	我的業績總是達不到公司標準，就算再怎麼努力都沒用...	113
Q14	我突然發現我最好的朋友居然和我喜歡同一個人，我該如何是好?	107
Q15	我們都在一起那麼久了，他居然在情人節跟我提分手，他怎麼能這麼狠心!	102
Q16	我真的好累，每天都要半工半讀，好想辭職，但家裡沒有我這份薪水又過不下去...	101
Q17	我真的好喜歡他，好想跟他告白，但又怕被拒絕，怎麼辦?	107
Q18	我發現我男/女朋友背叛我，我親眼看到他在街上跟別人牽手擁抱!	107
Q19	我發現我對現在學的東西真的完全沒興趣，唸得好痛苦...	116
Q20	我總是覺得自己比不上別人，可能我一輩子就這樣了吧...	106
Q21	我還這麼年輕，為什麼會罹患癌症，老天爺怎麼可以這樣對我...	103
Q22	每天去醫院看我奶奶的時候都覺得好難受，但又沒辦法做甚麼，真的好無力...	121
Q23	每天都聽到爸媽在房間吵架，壓力真的好大...	98
Q24	每次老師點到我，我都回答不出問題，其他同學都會，我好爛...	108
Q25	每次我報告時，主管總是故意挑我毛病，對我百般挑剔，我真的快受不了了!	101
Q26	每次都因為一點小事就跟我男/女朋友吵起來，他也不願意跟我溝通，好煩...	102
Q27	每次跟朋友吵架，他都不願意聽我意見，我真的不知道怎麼跟他說了	94
Q28	受到疫情的影響，我投資的餐廳都倒閉了，損失了好幾百萬，唉...	97
Q29	爸媽又開始在親戚面前誇獎姊姊了，但說到我就開始搖頭，唉...	98
Q30	前幾天不幸被裁員了，但我真的很需要這份工作，怎麼辦...?	108
Q31	為什麼明明是弟弟搶我東西，我還要被罵，是不是因為我是女生?	89
Q32	從小我父母就離婚了，每次看到同學一家人出遊，都好希望爸媽同時在我身邊...	109
Q33	最近不知道為什麼朋友突然不理我了，我到底做錯什麼...?	101
Q34	最近我常常莫名其妙就想哭，做什麼事都提不起勁，好痛苦...	104
Q35	最近事情好多、壓力好大，連想休息的時候都睡不著，只能靠安眠藥才勉強入睡...	102

(三)使用者 MBTI

針對負面情緒資料集本研究計算出各 MBTI 人格類型的回覆數，如圖 4-1，並將各 MBTI 人格類型的回覆數除以總回覆數得到各人格類型之比例再將其與實際 MBTI 人格類型之分佈比例(總比例)做比較，如圖 4-2。

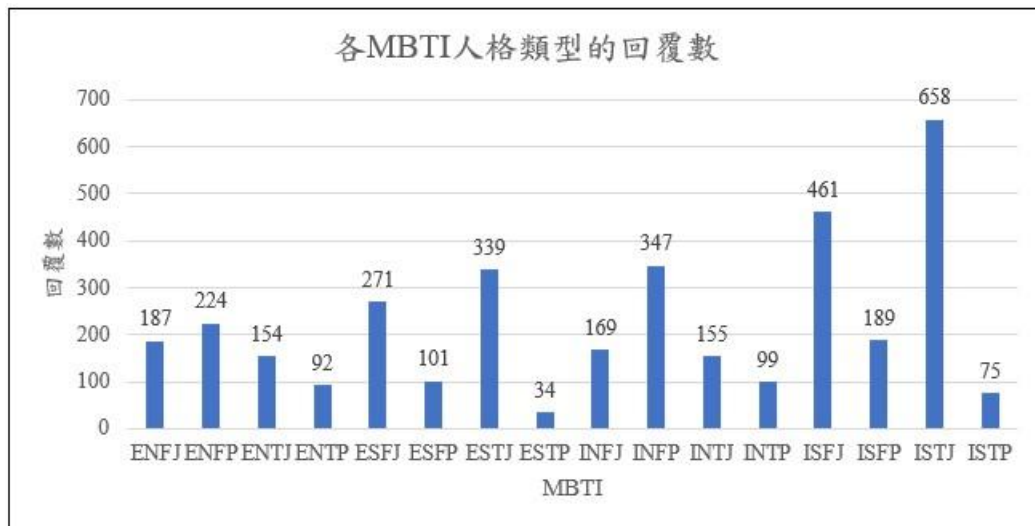


圖 4-2 各 MBTI 人格類型的回覆數

由圖 4-2 可知 ENTP、ESTP、INTP 及 ISTP 四種人格的個別回覆數皆低於 100 則，為本研究問卷較缺乏之人格類型，因此進一步比較其與實際 MBTI 人格類型的分佈比例是否一致，比較後得到此四種人格類型之比例皆確實較低。

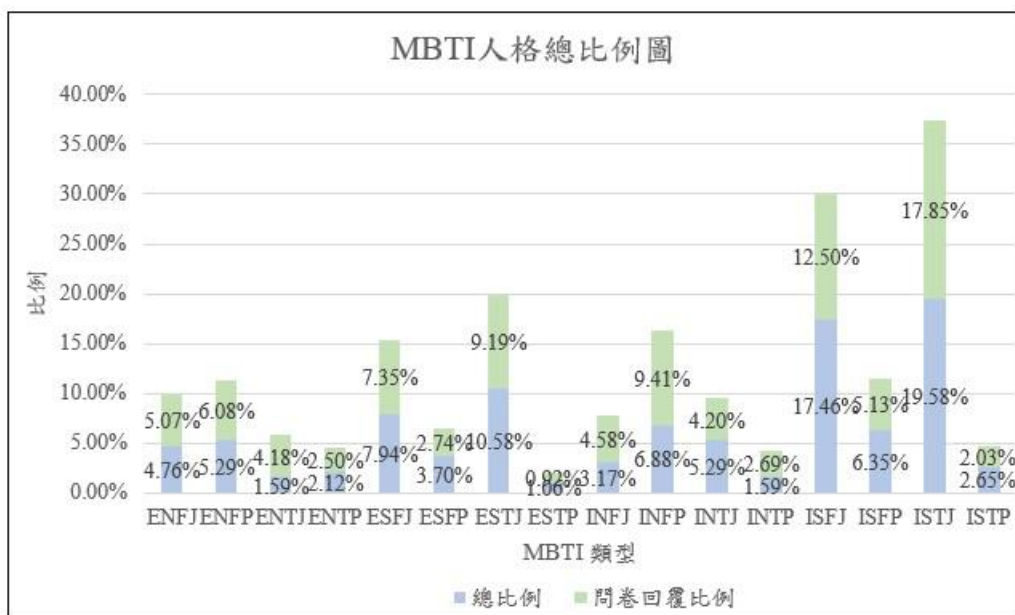


圖 4-1 問卷回覆比例與實際 MBTI 人格類型之分佈比例比較

第二節 語言風格模型

(一) 句子結構模型的多種語言風格語料庫

由於句子結構模型會因語料庫的語言風格不同，而使語言風格模型有不同的語言風格，所以本研究欲找尋不同語言風格的語料庫，而進行了以下實驗。

由於 LSM 介於 0 至 1，接近 0 為較不相似，接近 1 為較相似，並無一數值可界定兩文本間語言風格之相似程度。為得一判定兩文本間語言風格相似程度之門檻值，本研究利用負面情緒資料集，挑選完整填寫問卷 35 題之 52 位受測者資料，將每筆資料第 1 至 18 題與第 19 至 35 題之回覆內容各拆分成一組，共 5356 組。再將每組回覆內容進行斷詞，計算八個功能詞使用比例，最後計算每兩組之間之 LSM 指標。

本研究透過建立不同門檻值(門檻值設立由 0.4 至 0.96，以 0.02 為間隔)之混淆矩陣，以找尋分類正確率最高之門檻值，以兩組間是否來源於同一問卷填寫者為目標變數(來源於同一填寫者為 YES，反之為 NO)。但因目標變數 YES 只有 52 筆，NO 有 5304 筆，資料呈現長尾分配，故本研究透過 SMOTE(Synthetic Minority Oversampling Technique)演算法，將目標變數 YES 的比例從放大 10 至 100 倍(以 10 作為間隔)計算混淆矩陣，如下圖 4-3。

由於本研究在使用 LSM 指標之門檻值時，主要考慮此門檻值所預測出語言風格相似之情況，實際上也為相似，所以以 Precision 作為主要判斷指標，Recall 作為次要考量指標。經過計算發現將目標變數 YES 放大 100 倍時，得出之 Precision 值較其他放大倍數高，因此選擇放大 100 倍時所計算之混淆矩陣作為 LSM 指數門檻值判斷依據，如下表 4-3。

表 4-3 將目標變數 YES 放大 100 倍之評測指標

門檻值	Precision	Recall	F-measure
0.40	0.4977	1.0000	0.6646
0.42	0.4979	1.0000	0.6648
0.44	0.4981	1.0000	0.6649
0.46	0.4985	1.0000	0.6654
0.48	0.4994	1.0000	0.6662
0.50	0.5004	1.0000	0.6670
0.52	0.5017	1.0000	0.6682
0.54	0.5027	1.0000	0.6690
0.56	0.5038	1.0000	0.6700
0.58	0.5046	1.0000	0.6707
0.60	0.5064	1.0000	0.6723
0.62	0.5095	1.0000	0.6751

門檻值	Precision	Recall	F-measure
0.64	0.5109	0.9912	0.6743
0.66	0.5135	0.9796	0.6738
0.68	0.5167	0.9648	0.6730
0.70	0.5192	0.9356	0.6678
0.72	0.5263	0.9111	0.6672
0.74	0.5341	0.8795	0.6646
0.76	0.5412	0.8313	0.6556
0.78	0.5490	0.7548	0.6356
0.80	0.5574	0.6605	0.6046
0.82	0.5752	0.5550	0.5649
0.84	0.6036	0.4410	0.5096
0.86	0.6514	0.3416	0.4482
0.88	0.6890	0.2113	0.3235
0.90	0.7168	0.1026	0.1795
0.92	0.7520	0.0352	0.0673
0.94	0.7755	0.0072	0.0143
0.96	0.0000	0.0000	0.0000

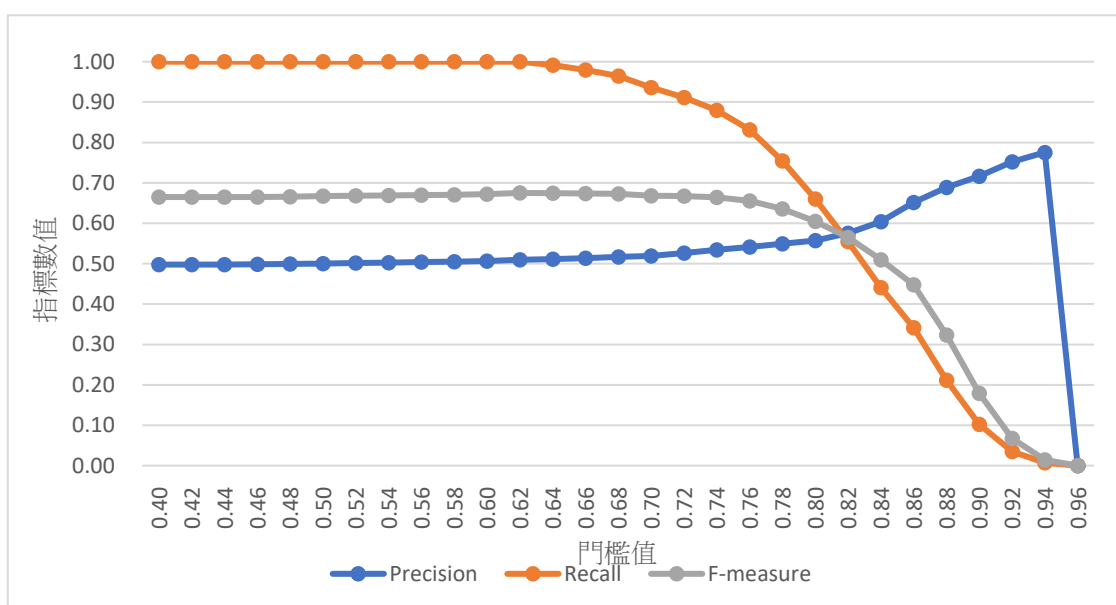


圖 4-3 將目標變數 YES 放大 100 倍之評測指標作圖

雖然本研究以 Precision 做為主要參考指標，但仍將 Recall 納入次要考量指標，故選擇適合的門檻值時，基於 Precision，會以 0.94 為第一優先，但 0.94 的 Recall 過低，0.92、0.9 同理；而 0.88 與 0.86 的 Precision 皆高於 0.65，雖然 0.86 的

Precision 相較於 0.88 低約 0.03，但 Recall 高了約 0.13，因此，本研究最後挑選 0.86 作為判定兩文本間語言風格是否相似之門檻值。本研究於網路上¹蒐集了 30 位不同作者之網路文章(詳見附錄二)，作為提供模型訓練之語料庫，作者名單如表 4-4。

表 4-4 蒐集之網路作者名單

ID	作者	ID	作者
1	A&F	16	肆一
2	不朽	17	Nir 說
3	男孩與貓	18	middle
4	恆思榆	19	riceandshine
5	溫如生	20	HOOK
6	蘇乙笙	21	好日曆
7	張西	22	Sandy
8	豆苗先生	23	Andy
9	Aliceeecccc	24	大姚
10	林達陽	25	illy
11	知寒	26	冒牌生
12	凌性傑	27	海島居民
13	否思	28	大坦誠
14	taylorleemw	29	lun
15	怪奇事務所	30	peter

¹2022 年 5 月 30 日至 2022 年 7 月 17 日於論壇 Dcard 與社交媒體 Instagram，蒐集作者當時論壇上所有文章

欲確保作者間的語言風格具差異性，30 位作者間兩兩計算 LSM，並利用門檻值 0.86 判斷兩兩作者間語言風格是否不相似，得出兩兩作者間語言風格不相似之結果如下表 4-5。

表 4-5 語言風格不相似之兩兩作者

ID	作者 1	作者 2	LSM
1	A&F	林達陽	0.8523
2	A&F	知寒	0.8328
3	A&F	凌性傑	0.8249
4	A&F	怪奇事務所	0.8048
5	A&F	riceandshine	0.8541
6	A&F	Andy	0.8219
7	A&F	海島居民	0.8222
8	A&F	lun	0.8444
9	不朽	Andy	0.8589
10	男孩與貓	知寒	0.8523
11	男孩與貓	怪奇事務所	0.8593
12	溫如生	知寒	0.8585
13	溫如生	怪奇事務所	0.8600
14	溫如生	Andy	0.8514
15	溫如生	illy	0.8580
16	蘇乙笙	知寒	0.8453
17	蘇乙笙	凌性傑	0.8556
18	Aliceeecccc	怪奇事務所	0.8463
19	Aliceeecccc	Andy	0.8541
20	Aliceeecccc	海島居民	0.8463
21	林達陽	知寒	0.8339
22	林達陽	否思	0.8460
23	林達陽	middle	0.8472
24	林達陽	Andy	0.8499
25	林達陽	illy	0.8338
26	知寒	凌性傑	0.8559
27	知寒	怪奇事務所	0.8137
28	知寒	riceandshine	0.8393
29	知寒	Andy	0.8212
30	知寒	海島居民	0.7840

ID	作者 1	作者 2	LSM
31	知寒	大坦誠	0.8172
32	知寒	lun	0.8147
33	凌性傑	Andy	0.8140
34	凌性傑	illy	0.8576
35	凌性傑	大坦誠	0.8498
36	否思	怪奇事務所	0.8386
37	否思	Andy	0.8577
38	否思	海島居民	0.8267
39	否思	lun	0.8549
40	怪奇事務所	肆一	0.8477
41	怪奇事務所	Nir 說	0.8498
42	怪奇事務所	middle	0.8465
43	怪奇事務所	riceandshine	0.8448
44	怪奇事務所	Andy	0.8513
45	怪奇事務所	illy	0.8263
46	怪奇事務所	大坦誠	0.8410
47	怪奇事務所	lun	0.8418
48	肆一	Andy	0.8515
49	肆一	海島居民	0.8398
50	肆一	lun	0.8484
51	Nir 說	Andy	0.8427
52	Nir 說	海島居民	0.8384
53	Nir 說	lun	0.8408
54	middle	Andy	0.8477
55	middle	海島居民	0.8111
56	middle	lun	0.8195
57	Andy	illy	0.8381
58	Andy	海島居民	0.8473
59	illy	海島居民	0.8431
60	illy	lun	0.8548

挑選出兩兩作者間語言風格均不相似且 LSM 平均最低之集合當作最終欲作為模型訓練之語言風格語料庫，如下表 4-6。本研究最終挑選了四位作者之語言風格語料庫，此語言風格語料庫將會是本研究語言風格模型可轉換之風格。

表 4-6 最終挑選出之作者語料庫

ID	作者	ID	作者
1	A&F	3	riceandshine
2	知寒	4	怪奇事務所

(二)語言風格模型參數調整

本專題的語言風格模型共有三個參數如表 4-7，此小節的實驗目的為，找到最適參數組合，使模型可以產生出最多合理的句子。合理的定義為模型再更改原語句語言風格的情況下，不影響更改後的句子表達原語句的語意。

表 4-7 語言風格模型參數定義

參數	定義
w_e	使用到何種 word embedding 技術(Word2Vec 或是 Word2FunctionVec)
f_s	篩選候選結構: 0 為會省略重複詞性的版本。1 為 0 的優化版，若需省略，只能省略語助詞。2 為 0 再新增長度限制。3 為 1 再新增長度限制。
c_s	是否使用 TF-IDF 以及餘弦相似度方法篩選句子。0 為不使用，1 為使用。

為了判斷哪種參數組合最好，需要先判斷模型更改後的句子是否合理，為此本研究設立一指標幫助判斷。合理性指標的實驗是由本研究小組中 7 名成員，以人工註記的方式判斷由模型生成出的 500 組原語句與更改語言風格後之語句的組合，並判斷更改語言風格後之語句是否能傳達原語句的語意，並由少數服從多數的方式決定，若大於或等於 4 人認為是合理的就會將此更改語言風格後之語句視為合理。本研究將判斷合理性的問題看作分類問題，因此將人工註記的結果視為答案(目標類別)，並分析 500 組組合的多屬性 (其他變數)如表 4-8。最後使用 Weka 跑決策樹，尋找 3 個指標表現最好以及過擬和最不明顯的樹當作指標，詳細數據如表 4-9。

表 4-8 500 組組合的屬性值範例

原語句	更改語言風格後之語句	餘弦相似度	相差長度	相差詞性	人工註記結果
晚安 晚安	晚安 喔	0.7918	0	VH:-1,T:1	1
確實 呢	鼓勵 確實 呢	0.7702	1	VF:1	0

表 4-9 決策樹表現

詞性因素型態	修剪	正確率	Precision	Recall	F-measure	交叉驗證 K	正確率	Precision	Recall	F-measure	建模正確率 -交叉驗證 正確率
nominal(0,1)	pruned	90.16%	0.777	0.723	0.749	16	83.13%	0.6	0.505	0.548	7.03%
nominal(0,1)	unpruned	92.57%	0.89	0.723	0.798	12	81.92%	0.563	0.485	0.521	10.64%
numeric	pruned	89.76%	0.821	0.634	0.715	12	83.93%	0.627	0.515	0.565	5.82%
numeric	unpruned	92.37%	0.839	0.772	0.804	14	83.73%	0.625	0.495	0.552	8.63%
nominal(-1,0,1)	pruned	88.35%	0.753	0.634	0.688	16	81.72%	0.481	0.376	0.422	6.63%
nominal(-1,0,1)	unpruned	93.37%	0.915	0.743	0.82	16	81.92%	0.568	0.455	0.505	11.45%

表 4-9 中，詞性因素型態指的是如何表達更改後之語句與原語句相差的詞性。假設更改後之語句比起原語句少了 2 個“VH”並且多了 1 個“T”，在 nominal(0,1) 的情況下會，“VH”的屬性值會是 1 同時在“T”的屬性值也會是 1。在 nominal(-1,0,1) 的情況下，“VH”的屬性值會是 -1 同時在“T”的屬性值會是 1。如果是在 numeric 的情況下，“VH”的屬性值會是 -2 同時在“T”的屬性值會是 1。如此分別是因為資料的型態會影響到決策樹的表現。另外表 4-9 中的最後一欄位建模正確率-交叉驗證則是過擬合的參考值。最終本研究決定使用詞性因素型態為 numeric 且有修剪的決策樹當作分類合理性的指標，因為這棵樹的過擬合狀況最少，並且 3 個指標在交叉驗證下也表現得最好。

擁有合理性指標後，就開始語言風格模型的參數調整。參數調整的實驗是使用 riceandshine 作者的語言風格做為測試，更改聊天資料集中所有回覆語句的語言風格，表 4-10 為語言風格模型在 riceandshine 語言風格下各種參數的表現。

表 4-10 語言風格模型在 riceandshine 語言風格的表現

w_e	f_s	c_s	合理數量	原始筆數	正確率
Word2Vec	2	0	1333	7733	17.24%
Word2Vec	3	0	1766	12589	14.03%
Word2Vec	0	1	2389	22384	10.67%
Word2Vec	1	1	1779	14916	11.93%
Word2Vec	2	1	1333	7733	17.24%
Word2Vec	3	1	1012	4382	23.09%

w_e	f_s	c_s	合理數量	原始筆數	正確率
Word2FunctionVec	3	1	1020	4377	23.30%

由於參數 f_s 與 c_s 會互相影響，而 w_e 獨立於另外兩個參數，因此實驗的步驟為找出最好的 f_s 與 c_s 組合，在調整 w_e 。其中 f_s 與 c_s 的組合中缺少了 (0,0) 和 (1,0)，是因為這兩組參數的運算量過大，本研究的電腦設備無法運算。表 4-10 中的原始筆數代表模型所生成出的全部更改語言風格後的語句，原始筆數的多寡會被 f_s 和 c_s 直接影響。越嚴格的參數組合會導致產生出的語句越少，但同時也會越精準。合理數量則模型更改後的所有語句，在經過前一小節得出的合理性指標(決策樹)分類後，被判斷為合理的數量。正確率的算法則是合理數量除以原始數量。當模型是在 w_e 為 Word2Vec 的情況下，另外兩個參數為 (3,1) 所產生出合理句子的正確率最高。若使用 w_e 為 Word2FunctionVec 在 (3,1) 的情況下，產生出的合理數量些微的提高了，並且正確率也提高了一點，因此最後本研究採用 (Word2FunctionVec, 3, 1) 作為語言風格模型的參數設定。

第三節 挑選適當語句回覆使用者

延續上一小節生成之語言風格模型，當使用者輸入語句時，本研究將挑選適當語句回覆使用者，如下圖 4-4。首先將提取使用者輸入的聊天語句(以下簡稱 Input)，並結合多種語言風格之聊天答覆資料集(簡稱 LsmData)，接下來(1)將會將 Input 使用 word2Vec 轉換為句向量與 S2VecQ 計算餘弦相似度，取出與 Input 餘弦相似度最高的 S2VecQ，(2)判斷 S2VecQ 是否與 NegQ 相同，若相同則表 S2VecQ 為本研究設計之負面情緒資料集，並會(3-1)使用 KNN 選取與使用者 MBTI 相似人之回覆語句回覆使用者；若與 NegQ 不同則使用 S2VecQ 對應到之 DatasetA 回覆使用者。在計算 Input 與 S2VecQ 餘弦相似度的同時，本研究將會分析使用者 Input，首先(1-2)會判斷使用者與聊天機器人聊天對話是否閒置 30 分鐘以上，若閒置 30 分鐘以上則如流程圖中(1-3)使用 LIWC 辭典分析使用者的情緒，LIWC 辭典之詳細分析流程已於第三章第五節說明，若判斷使用者有負面情緒傾向，則(1-4)聊天機器人將會主動關心使用者。

挑選適當語句回覆使用者流程圖

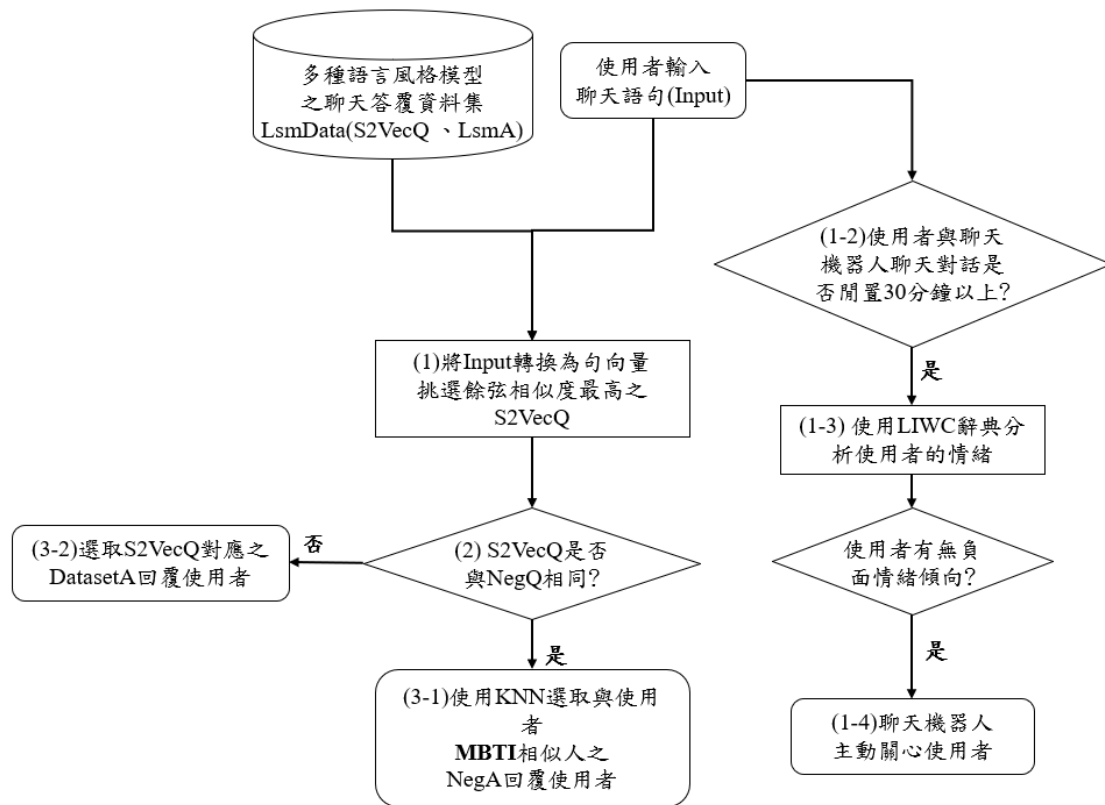


圖 4-4 挑選適當語句回覆使用者流程圖

為將使用者輸入的語句(以下簡稱 Input)以向量方式表達其語意，首先，本研究將提取 Input 並使用 Jieba 中文斷詞系統進行斷詞，如下表 4-11。

表 4-11 使用者 Input 及 Jieba 斷詞結果示意表

Input 1	今天天氣真好
斷詞結果 1	今天/天氣/真好
Input 2	我把他當成最好的朋友，他卻誣賴我，我好失望好難過，心情好低落。
斷詞結果 2	我/把/他/當成/最好/的/朋友/，/他/卻/誣賴/我/，/我/好/失望/好/難過/，/心情/好/低落/。

再使用 word2vec 演算法，將每一字詞以 110 維度之詞向量表示其語意，向量維度設置方式已於第三章第五節說明。並將每一字詞之詞向量加總形成 Input 語句之句子向量。研究過程中發現語助詞會影響句向量之計算，故在計算句子向量時本研究將針對詞性給予字詞不同權重做調整，最後再加總形成 Input 語句句向量，並挑選出 LsmData 中的 S2VecQ 與 Input 餘弦相似度最高之語句，如圖 4-4 (1)。

選出 S2VecQ 後將判斷 S2VecQ 是否為本研究設計之負面情緒問卷之問題，如圖 4-4 (2)。若 S2VecQ 為負面情緒問卷之問題如圖 4-4 (3-1)，則將使用 KNN 演算法挑選與使用者 MBTI 相似人之語句 NegA 回覆使用者；若 S2VecQ 不為負面情緒問卷之問題如圖 4-4 (3-2)，則將使用一般聊天語句回覆使用者。

本研究將於使用者聊天過程中記錄聊天內容，匯集一段時間內的聊天紀錄如圖 4-4 (1-2)，若使用者於聊天過程中閒置 30 分鐘以上未回覆，則視為下一次的分析紀錄。透過 LIWC 分析使用者之聊天紀錄如圖 4-4 (1-3)，計算負面情緒詞之比例，若該比例大於 0.07，則視使用者於此段時間的聊天過程有負面傾向，本 app 將會在下次與使用者聊天過程中跳出關心之訊息如圖 4-4 (1-4)。

第四節 實驗設計與驗證

(一)驗證 MBTI 做為因素是否更適合使用者

研究將從 NegQ(已於第參章第三節詳細說明)隨機挑選出 10 題，利用歐式距離計算出與使用者 MBTI 距離最近及最遠之 NegA 當作選項。測驗結束後記錄使用者選擇幾次最近的選項，依填答情況分為 0~4 題偏好互補性格、5 題無偏好、6~10 題偏好相似性格，因無偏好與欲驗證之假設無關故刪除，最後計算各 MBTI 喜好之比例，如公式(13)、公式(14)。

$$\text{MBTI 偏好相似比例} = \frac{\text{此MBTI 偏好相似性格人數}}{\text{此MBTI 總人數}} \quad \text{公式(13)}$$

$$\text{MBTI 偏好互補比例} = \frac{\text{此MBTI 互補相似性格人數}}{\text{此MBTI 總人數}} \quad \text{公式(14)}$$

2022 年 10 月 2 日研究問卷 251 筆資料中 173 筆有效資料計算的實驗結果如表 4-12，大多數人格偏好性格互補的回覆，尤其 ISTP 及 INTJ 有高達七成以上的使用者傾向互補性格之回覆，而較明顯喜歡性格相似回覆的有 ESTP 及 INFJ 兩種人格，但交叉比對 MBTI 的四面相並沒有明顯的趨勢。

表 4-12 各 MBTI 人格偏好相似或互補之比例

MBTI 人格	偏好相似比例	偏好互補比例
ENFJ	0.4167	0.5833
ENFP	0.4667	0.5333
ENTJ	0.5455	0.4545

MBTI 人格	偏好相似比例	偏好互補比例
ENTP	0.5714	0.4286
ESFJ	0.4667	0.5333
ESFP	0.5	0.5
ESTJ	0.3929	0.6071
ESTP	0.6667	0.3333
INFJ	0.625	0.375
INFP	0.4412	0.5588
INTJ	0.3	0.7
INTP	0.5	0.5
ISFJ	0.4483	0.5517
ISFP	0.5	0.5
ISTJ	0.3182	0.6818
ISTP	0.1667	0.8333

(二)使用者是否偏好語言風格模型轉換過後的語句

於第肆章第二節提及本研究將產出四個不同作者之語言風格模型，以下分別稱為 LsmData 1(S2VecQ、LsmA1)、LsmData 2(S2VecQ、LsmA2)、LsmData 3(S2VecQ、LsmA3)、LsmData 4(S2VecQ、LsmA4)。本研究將從 LsmData 中挑選 15 題 S2VecQ 作為驗證之題目，S2VecQ 對應的 LsmA 做為選項。

挑選 15 題驗證題目之依據為，相同 S2VecQ 之下，四個 LSM 模型都有將 DatasetA 生成出 LsmA。最後，本研究將挑選同一個 S2VecQ 產出 LsmA 之作者的各種組合當作問卷第三階段之選項，如下表 4-13。

表 4-13 相同 S2VecQ 產出 LsmA 之作者的各種組合

題號	S2VecQ	DatasetA	LsmA	產出 LsmA 之作者
1	我不要	不勉強	我不太勉強	作者 2、作者 3
			其實不勉強了	作者 1、作者 4
2	我是你的好老婆	可能是	可能是嗎	作者 1、作者 3
			這可能是	作者 2、作者 4
3	游泳課根本沒在幹嘛	浪費時間	浪費這段時間	作者 2、作者 3
			浪費一段時間	作者 1、作者 4
4	你再囂張阿	我冷靜	所以我要冷靜了	作者 2
			我很冷靜	作者 1、作者 3、作者 4
5	你運勢怎麼樣	應該不錯	目前應該是很不錯	作者 2

題號	S2VecQ	DatasetA	LsmA	產出 LsmA 之作者
			應該很不錯呀	作者 1、作者 3、作者 4
6	我更無語	我也無語	我也真是無語了	作者 1、作者 2、作者 3
			我也覺得超無語	作者 4
7	沒事，休息一會就好了	快點休息	現在快點快點休息	作者 3
			快點去休息	作者 1、作者 2、作者 4
8	屬蛇適合戀愛嗎	不適合	不是很適合	作者 3
			不適合嗎	作者 1、作者 2、作者 4
9	你什麼時候變這麼乖了	一直都乖	我一直都不乖	作者 1
			一直都不乖	作者 2、作者 3、作者 4
10	我挺好	我也不錯	我也還不錯呀	作者 1、作者 2、作者 3
			但我也很不錯呀	作者 4
11	這樣會不會太沒誠意了	心意最重要	有個心意最重要	作者 1
			心意總是最重要	作者 2、作者 3、作者 4
12	好緊張	超緊張	我也超緊張	作者 1、作者 2、作者 3、作者 4
13	我不說對嗎	不對	不對呀	作者 1、作者 2、作者 3、作者 4
14	我很好，你呢	還不錯	還不錯呀	作者 1、作者 2、作者 3、作者 4
15	我惱怒	不要生氣	不要生氣了	作者 1、作者 2、作者 3、作者 4

此驗證為本研究之問卷第三階段之驗證，共有 15 道選擇題，使用者若選擇語言風格模型轉換後的語句達 8 道題以上，將視為偏好語言風格。本研究以偏好語言之人數除以總人數，從 2022 年 10 月 2 日研究問卷 251 筆資料中 173 筆有效資料計算的實驗結果顯示，約有 53% 的使用者偏好語言風格。本研究更以 MBTI 四面向來作分別，發現 N-直覺型及 P-感知型的使用者分別有多達 60.3% 及 62.8% 偏好語言風格模型轉換

表 4-14 MBTI 各面向之語言風格偏好程度

MBTI 面向	偏好語言風格比例
E-外向	0.5854
I-內向	0.5
N-直覺	0.6038
S-實際	0.4677
F-情感	0.5645
T-思考	0.4906
J-判斷	0.4722
P-感知	0.6279

第伍章 系統介面

在開啟本 APP 後，首先會看到的是登入頁面(圖 5-1)，若是尚未註冊帳號及密碼，點擊左下角的「註冊帳號」即可進入註冊頁面(圖 5-2)，在此頁面輸入其相對應之基本資料以註冊個人帳號。而登入頁面右下角的「忘記密碼」會進入到忘記密碼的頁面(圖 5-3)，僅須回答密碼提示問題之回答，就可以找回密碼。



圖 5-1 登入頁面 圖 5-2 註冊頁面 圖 5-3 忘記密碼 圖 5-4 測驗解說

登入後，會先看到測驗解說頁面(圖 5-4)，此頁面為解釋本 MBTI 心理測驗的用意及教學使用者如何完成，使用者須依題目往左或往右滑動以選取較接近自身行為、個性的選項。MBTI 心理測驗題目畫面如(圖 5-5)，完成 70 題後測驗，會顯示使用者的人格為何。接著，本 APP 將會進行三階段測驗。首先第一階段的測驗為 15 題負面情緒選擇題(圖 5-6)，本研究將生活中煩惱分成六面向並建立不同情境，貓咪將向使用者傾訴她的煩惱，使用者將選取他會安慰貓咪的回覆語句，此測驗不需擁有個人經驗，只要能感同身受安慰貓咪即可填答，本階段是為了解使用者講話的語言風格為何。第二、三階段測驗皆是依據題目選出使用者較偏好之回覆，本 APP 根據以上測驗結果為使用者客製化聊天機器人，讓使用者能有和朋友聊天的感覺。

此外，本 APP 的主頁面(圖 5-7)也有隨機的鼓勵小語能開啟使用者一天的好心情。修改密碼(圖 5-8)則是可更改註冊帳號時所設定的密碼。

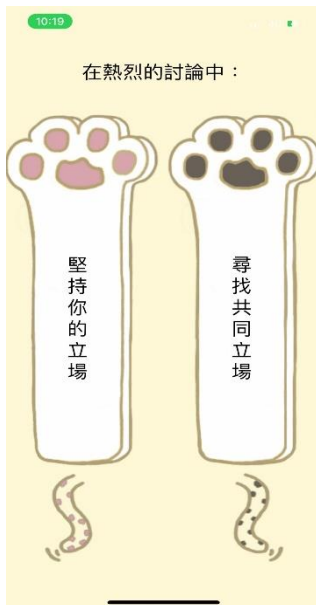


圖 5-5 MBTI 測驗

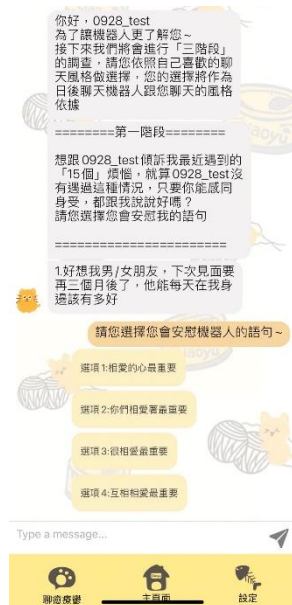


圖 5-6 負面情選擇題



圖 5-7 主頁面



圖 5-8 忘記密碼

第陸章 結論與建議

本章第一節介紹本研究的結果與發現，第二節說明研究限制及未來發展建議。

第一節 研究結果

目前市面上的聊天機器人 App 功能都相當完善，但使用者的黏著度都及長期使用的意願都不高，原因為聊天機器人回覆語句不多元導致聊天過於制式化，使用上不如和真人聊天般可以達到排解負面情緒的效果，因此本研究創新了一套能夠客製化語言風格的聊天機器人。

由於每個使用者語言風格偏好盡不相同，本研究提出了創新的「語言風格模型」其中包括 N-gram、句子解構模型以及 Word2FunctionVector，用以轉換原語句的語言風格，在不改變原語意的情況下使回覆語句變得更多元，且使用語文探索與字詞計算辭典分析使用者的聊天內容，若使用者輸入的負面情緒用詞達一定比例時，機器人將會主動關心使用者，另外本研究使用 MBTI 人格測驗分析使用者的人格，並結合 KNN 演算法了解使用者偏好與自己人格相似或互補的人聊天。最後，為驗證整體概念以及語言風格模型的可行性設計兩個實驗。

首先，本研究為了驗證回覆語句以 MBTI 做為依據是否符合使用者偏好，依比例統計出 MBTI 十六種人格中有九種人格偏好與自己人格互補的回覆語句，其中 ISTP 及 INTJ 兩種人格有高達 70% 以上的使用者傾向互補性格之回覆，而 ESTP 及 INFJ 兩種人格有高達 63% 以上的使用者傾向相似性格之回覆。由此僅能觀察出 ISTP、INTJ、ESTP 及 INFJ 四種人格的偏好，其他人格則無明顯趨勢。

另外，為了驗證使用者是否偏好語言風格模型轉換過後的語句，依比例統計出約有 53% 的使用者偏好語言風格，更以 MBTI 四面向做分析，觀察出 N-直覺型及 P-感知型皆有 60% 以上的使用者偏好語言風格模型轉換後的語句。

透過以上的實驗結果可得知本研究設置的語言風格模型更改後的語句在一定程度下可使聊天機器人回覆語句變得更多元且符合使用者的語言風格偏好，成為使用者的專屬聊天機器人。

第二節 研究限制與未來發展

(一)研究限制

本研究旨在改善機器人回覆語句制式化的問題，希望透過語言風格模型的轉換，將原語句在保有原語意的情況下變化得更多元，因為下列限制導致無法反映出語言風格模型的可行性。

1. 訓練資料集的不足

為了訓練聊天機器人自動化回覆，需要大量聊天對話資料集，但因聊天對話內容較隱私，導致目前無一開放平台有相關的資料集以訓練模型。

2. 問卷有效樣本數不足

由於本研究為了分析使用者的語言風格偏好，以 MBTI 人格測驗及三階段問卷作為依據，但因 MBTI 人格測驗題數高達 70 題，且三階段問卷共 35 題，造成問卷題數過多，難以收集到有效樣本。

3. 硬體設備的限制

本研究的語言風格模型是使用重組句子的方式改變語言風格。在模型重組句子時，根據句子的長度要判斷的可能性可以達到上千萬種，因此受限於硬體設備，只能將可能性限制在一定的數量下，才能夠執行。但語言風格模型如果可以在毫無設限的條件下執行，重組句子的效果會更好。

(二)未來發展

受限於專題時限以及比賽時間，本專題的機器人以及語言風格模型一些地方可以更加完善，目前雖已經可以進行聊天，但如果可以加上以下幾點的改善可以使模型以及機器人更加完美。

1. 語言風格模型完全自動化

如(一)研究限制第三點所述，語言風格模型適用重組句子的方式改變語句的語言風格，在改變風格的同時，為了避免影響原本的語意，本專題使用了機器學習(決策樹)的方式進行初步的判斷改變後的語句是否改變原始的語意。實際上，決策樹無法有效的進行語意是否改變的評估，因此本研究在決策樹篩選過後再透過組員人工的檢查，最後才能決定改變風格後的語句是否能傳達原始的語意。在

未來，可以結合中央研究院辭庫小組對於中文句結構樹的研究，創作出能自動判斷語意是否改變的系統，達到節省人力的同時，也可以使語言風格模型更好的重組句子。

2. 聊天的限制

目前本研究的機器人以及語言風格模型，只能讀懂一句話(不含逗號連接)語意或語言風格。在未來的研究中，可以加入對於長篇文章中每句話的語意分析，使聊天的過程更加流暢。

3. 自然語言理解以及機器人的回覆系統

本研究機器人是使用 Word2Vec 了解語意，並且回覆是建構在聊天資料集上。這樣的系統若很適合功能型聊天機器人(例如客服機器人)，因為使用者輸入的內容會被限縮在一定的範圍內。但本專題的機器人是為了聊天而創造的，因此若能結合網路搜尋以及深度學習的方式建構聊天系統，就能使機器人的回覆更多元並且更符合時事。但受限於專題的時間，因此將上述方案放在未來發展。

參考文獻

- [Afshin, A., Sur, P. J., Fay, K. A., Cornaby, L., Ferrara, G., Salama, J. S., ... Murray, C. J. \(2019\). Health effects of dietary risks in 195 countries, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet*, 393\(10184\), 1958-1972.](#)
- [Birla, N. \(2019, September 10\). Mental health may hurt India to tune of \\$1.03 trillion; here's a dose for cos. *The Economic Times*.](#)
- [Cover, T., & Hart, P. \(1967\). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13\(1\), 21-27.](#)
- [Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. \(2016\). Efficient kNN classification algorithm for big data. *Neurocomputing*, 195, 143-148.](#)
- [Huang, C. L., Chung, C. K., Hui, N., Lin, Y. C., Seih, Y. T., Ben C. P., ... Pennebaker, J. W. \(2012\). The development of the Chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology*, 54\(2\), 185-201.](#)
- [Ireland, M. E., & Pennebaker, J. W. \(2010\). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology*, 99\(3\), 549-571.](#)
- [Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. \(2011\). Language style matching predicts relationship initiation and stability. *Psychological Science*, 22, 39-44.](#)
- [Ma, W. Y., & Chen, K. J. \(2005\). Design of CKIP Chinese word segmentation system. *Chinese and Oriental Languages Information Processing Society*, 14\(3\), 235-249.](#)

Mehl, M. R., Gosling, S. D., & Pennebaker, J. W. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology*, 90, 862- 877.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, 26.

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, 45(3), 211-236.

Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, 77(6), 1296-1312.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count: LIWC*. Mahwah, NJ: Lawrence Erlbaum Associates.

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, 63(4), 517-522.

Tewari, A., Chhabria, A., Khalsa, A. S., Chaudhary, S., & Kanal, H. (2021, April). *A survey of mental health chatbots using NLP*. In proceedings of the International Conference on Innovative Computing & Communication (ICICC), New Delhi.

Turing, A. M. (1950). Computing machine and intelligence. *Mind*, 59, 443-460.

Wang, H. (2014). *Introduction to Word2vec and its application to find predominant word senses*. Retrieved from Nanyang Technological University, Institute for Division of Linguistics and Multilingual Studies Web site: <http://compling.hss.ntu.edu.sg/>

Weizenbaum, J. (1996). ELIZA-A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1), 36-45

Zhang, S., Li, X., Zong, M., Zhu, X., & Wang, R. (2017). Efficient kNN classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems*, 29(5), 1774-1785.

王靜儀(2015)。應用 MBTI 性格量表探討消費者對不同比例矩形之偏好。感
性學報,3(1)。

張硯評(2011)。感恩表達與配偶之生活適應(未出版博碩士論文)。國立台灣大
學，台北市。

張瓊之(2019)。可轉變對話風格的聊天機器人。國立臺灣大學電機工程學研究
所學術論文，已出版，臺北市。

許芳瑀(2020)。虛擬參考資訊服務革新-對話服務平台之比較研究。國立臺灣
師範大學圖書資訊學研究所學術論文，已出版，臺北市。

附錄一 蒐集之網路文章來源

個版	篇數	
Whyserious	300	https://www.dcard.tw/f/whysoserious
Talk	150	https://www.dcard.tw/f/talk
Mood	150	https://www.dcard.tw/f/mood
Relation	150	https://www.dcard.tw/f/relationship
Job	150	https://www.dcard.tw/f/job
freshman	142	https://www.dcard.tw/f/freshman
food	127	https://www.dcard.tw/f/food
Girl	115	https://www.dcard.tw/f/girl
makeup	113	https://www.dcard.tw/f/makeup
exam	111	https://www.dcard.tw/f/exam
meme	101	https://www.dcard.tw/f/meme
youtuber	88	https://www.dcard.tw/f/youtuber
Pet	80	https://www.dcard.tw/f/pet
Trending	68	https://www.dcard.tw/f/trending
Funny	60	https://www.dcard.tw/f/funny
Dressup	60	https://www.dcard.tw/f/dressup
Tv	33	https://www.dcard.tw/f/tvepisode
2019_ncov	32	https://www.dcard.tw/f/2019_ncov
marriage	18	https://www.dcard.tw/f/marriage
boy	10	https://www.dcard.tw/f/boy

附錄二 蒐集之網路作者文章來源

作者名	社群平台	網址
A&F	DCARD	https://www.dcard.tw/@a1234887
不朽	IG	https://www.instagram.com/taylorlmw/
男孩與貓	DCARD	https://www.dcard.tw/@dkt.0727
恆思榆	IG	https://www.instagram.com/struggledog/
溫如生	IG	https://www.instagram.com/issmisally/
蘇乙笙	IG	https://www.instagram.com/siahuei/
張西	IG	https://www.instagram.com/ayrichang/
豆苗先生	IG	https://www.instagram.com/mr.doumiao/
Aliceeecccc	IG	https://www.instagram.com/aliceeecccc/
林達陽	IG	https://www.instagram.com/poemlin0511/
知寒	IG	https://www.instagram.com/infernowords/
凌性傑	IG	https://www.instagram.com/lshjet/
否思	IG	https://www.instagram.com/fourzpoem/
taylorleemw	IG	https://www.instagram.com/taylorleemw/
怪奇事務所	IG	https://www.instagram.com/incrediville_tw/?igshid=YmMyMTA2M2Y=
肆一	IG	https://www.instagram.com/fourone4141/
Nir 說	IG	https://www.instagram.com/nir_talk/
Middle	IG	https://www.instagram.com/mid810/
riceandshine	IG	https://www.instagram.com/riceandshine.co/
HOOK	IG	https://www.instagram.com/helloiamhook/
好日曆	IG	https://www.instagram.com/calendar.tw/
吳姍儒	IG	https://www.instagram.com/sandywis/
Andy	IG	https://www.instagram.com/real.nd_official/
大姚	IG	https://www.instagram.com/day.dayao/
illy	IG	https://www.instagram.com/illyqueen/
冒牌生	IG	https://www.instagram.com/inmywordz/
海島居民	IG	https://www.instagram.com/thecoastlander.lc/
大坦誠	IG	https://www.instagram.com/bigtan___bibi/
lun	IG	https://www.instagram.com/lun.tw/
peter	IG	https://www.instagram.com/peter825/