機率與統計 期末報告

使用BERT進行疾病推斷

國立金門大學

資訊工程系

組長：魏仲彥

指導教授：張珀銀 老師

目錄

(一)摘要

疾病引響人類的生活，許多人患了重大疾病往往因為缺乏知識而選擇忽視，導致延誤就醫，造成嚴重的後果，甚至死亡。根據世界衛生組織 2020 年提供的全因死亡報告[1]，可以得知因疾病死亡的人數之多，我們對於其重視程度也隨之提高。因此，開發一種有效預測和診斷疾病的方法非常重要。在過去的幾十年裡，人工智能和機器學習的進步使疾病推斷的技術成為可行的方法。最近，一些研究使用機器學習來診斷疾病。然而，有些方法只能有限或是固定的方面進行疾病分析[2, 3, 4]。一旦牽扯到廣泛的疾病分析，這些方法就無法有效預測和診斷疾病。

為了解決這個問題，在本專題中，我們提出了一種機器學習方法，可以使用動態數據有效地預測和診斷疾病。我們的方法基於由 Transforms[12] decoder 部分延伸出來的 BERT[10]。我們首先將患者的動態數據轉換。然後，我們使用疾病的類別和該類別會有的問診相關對話來訓練我們的模型。實驗結果的準確率表明，我們的方法可以有效地預測和診斷疾病，可以讓患者可以提前預防，進行更進一步的身體保養。

透過使用 BERT 模型，我們希望能夠藉由與用戶的對話來推斷用戶的疾病。為了訓練和評估模型，我們收集了大量的疾病對話數據，並使用準確率作為模型的評估指標。在實驗中，我們的模型在預測疾病方面表現出色，達到了 90% 的準確率。通過使用 BERT 模型，我們的研究為疾病對話推斷提供了一種有效的方法，並且對於解決在醫療領域中常見的問題有所貢獻。然而，由於數據集的局限性，我們的模型尚未在現實中得到實際的測試。因此，未來的工作將致力於在更大和更真實的數據集上測試和擴展我們的模型。

(二)緒論

研究背景

目前人工智慧在自然語言處理、影像辨識、視覺檢索、醫療影像分析等領域已有了很大的進步[5, 6]，尤其是最近幾年來，隨著電腦規格的快速提升，並且各領域人才的成長，人工智慧的實力也在逐年增強。針對這一點，我們想要透過人工智慧技術，來做疾病分析和推斷。

在醫療領域中，快速且準確地推斷病人的疾病是至關重要的。因此，我們希望通過使用自然語言處理技術來解決這個問題，並提出一種使用 BERT 模型進行疾病對話推斷的方法。

研究動機

現今疫情肆虐全球，大多數人生了小病也不會出門，而是選擇待在家裡靜養，這樣可能會導致身體出現嚴重不適時，無法及時就醫而造成無法挽回的結果。本

專題開發的疾病預測方式，可以提早偵測到身體出現的症狀可能的疾病，從而達到預防的效果。

研究問題

本計畫所要探討【研究問題】包含：

(1) 如何運用神經網路進行疾病分析？

(2) 如何使用動態資料進行疾病分析？

(3) 是否可以使用 BERT 進行疾病推斷?

(4) 使用 BERT 模型進行疾病對話推斷的效果如何？

(5) 對於疾病分析技術的改善與發展?

研究目的

(1) 提出一種基於 BERT 的疾病分析技術。

(2) 利用互動的方式，讓使用者可以透過 APP 了解目前自身疾病。

(3) 透過 BERT 的預訓練模型，結合資料集進行疾病分類

(4) 討論本專題對於疾病推斷的準確度，和應用成果。

(5) 說明本專題研究成果對於疾病推斷的貢獻及未來發展。

(三)文獻探討與回顧

我們蒐集了台灣衛生福利部的國人死亡數據，並把這些資料整理成圖表，以顯示 102 年到 110 年，人口因疾病死亡和其他因素的分布比較圖(見圖 1、表 1)。

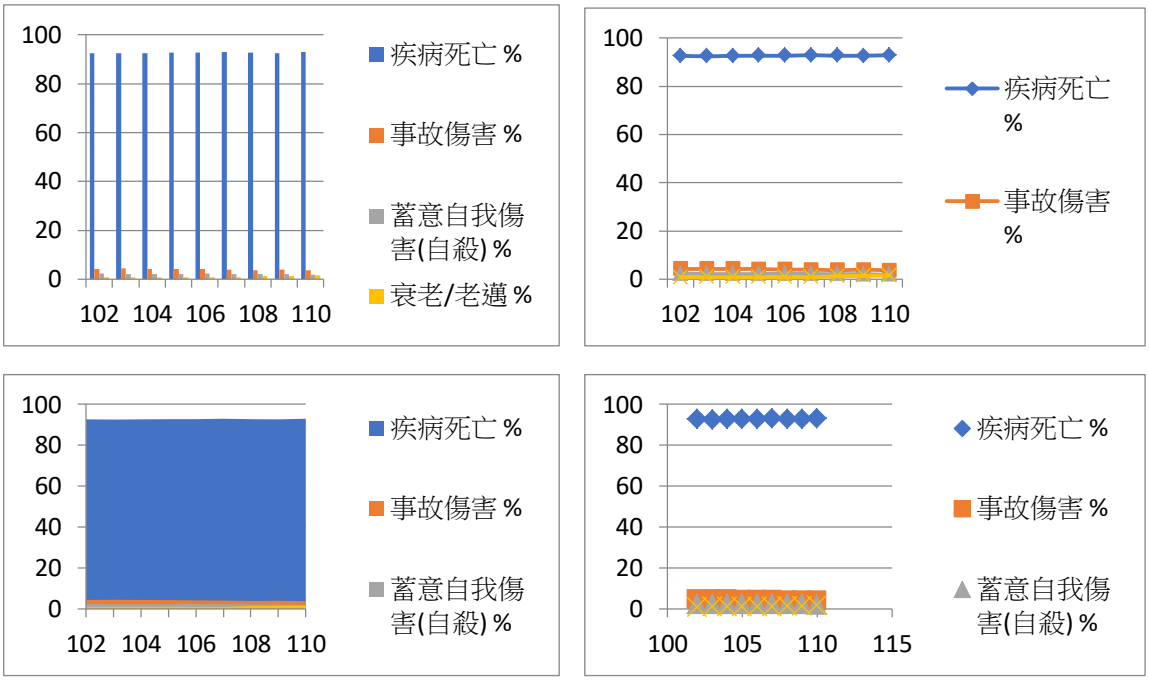

圖 1 直方圖、折線圖、區域圖、散佈圖

死亡類型	Obs	Mean	Max	Min	SD	CV
所有死亡原因	9	100	100	100	0	0
疾病死亡	9	92.68888889	92.9	92.5	0.136422546	0.018611111
事故傷害	9	4.077777778	4.4	3.7	0.243812314	0.059444444
蓄意自我傷害(自殺)	9	2.177777778	2.3	1.9	0.120185043	0.014444444
衰老/老邁	9	1.055555556	1.5	0.8	0.265099562	0.070277778

表 1 敘述統計量表

可以由衛生福利部整理的數據發現，大多數的死亡原因都是因為疾病所導致，可見提早發現疾病的重要性。

近年來，有許多研究都在使用人工智慧來判斷疾病，但大多都是針對一種類型的疾病，像是使用 RF 預測肝病[2]、使用 SVM 預測帕金森氏症[3]或是 LSVM 預測心血管疾病[4]，只有少數方法像是 CART，或是 Cosine Similarity[11]是針對廣義的疾病做預測。

CART

用決策樹裡面的 CART 做為模型，可以針對許多不相關的特徵進行數 據處理，對於本專題的多種疾病和症狀的推測，CART 模型可以高效率處理。此 外，CART 是一個白盒模型，易於推導和觀察，而且只需要一次建構， 就可以反覆使用，而且就算輸入的症狀很多，CART 的樹狀分類也可以有效的識別出疾病的類別，像是圖 2，就是經由給定症狀標籤，經由整數轉換後，依次判別輸入症狀的內容，以達到多種疾病的分析。但是在輸出上，無法顯示機率，只會有唯一解，導致在特殊情況下，輸出的疾病類別會誤差非常大

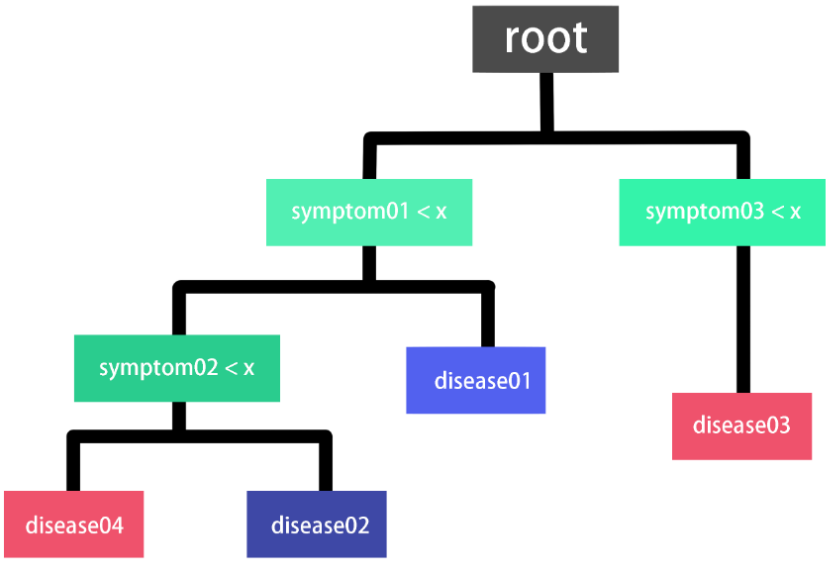

圖 2　CART

Cosine Similarity

使用余弦相似度判斷兩個向量的相似度，公式如(1)，可以使用陣列表示疾病，做訓練資料，利用多種特徵，來推斷疾病。

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}} \quad (1)$$

我們把症狀使用紀錄下來(見表 2)，並寫出對應的疾病對應的 vector(見表 3)，假設今天有一位使用者輸入"我常常覺得噁心、胃痛，且想吐"，就可以由使用者輸入的向量與數據及預設向量做內積，去推估疾病，可以根據不同狀況輸出不同的疾病機率，增加模型靈活性。

症狀	Flagged
頭痛	0
想吐	1
胃痛	1
想吐	1
消化不良	0
流鼻水	0

表 2　症狀編碼

疾病	vector	similiarity
流感	[1, 0, 0, 1, 1, 1]	21%
食物中毒	[0, 1, 1, 1, 0, 0]	100%

表 3　疾病編碼

CART 和 Cosine Similarity 能夠有效處理症狀推測疾病的問題，但是也各有缺點(見表 4)，這兩種方法對於疾病的預測還有的最大問題就是，當輸入是資料集以外的資料，那就會輸出很嚴重的錯誤資訊，導致疾病的誤判。因此，我們的研究旨在通過使用 BERT 模型來解決這個問題，並有良好的疾病推斷的效果。

模型	優點	缺點
CART	便於推導和解釋結果	輸出的答案不會有機率
COS	可以快速得出答案	資料集的欄位超級多
BERT	使用自然語言理解，準確率高	訓練時間比較久

表 4　模型方法比較

(四)研究方法與步驟

本專題使用維基百科的資料進行資料蒐集，再根據疾病和症狀延伸出各種資料和對話，最後轉換成為資料集。我們利用 Cross validation 將資料集分為兩成的測試集和八成的訓練集，再分割成十份，分別訓練，我們會把得到的 10 組 Accuracy 指標做加總後平均，得到我們最後的準確率，建立完模型後，等待使用者輸入，模型將會回傳症狀分析的疾病推斷結果。圖 3 介紹了專案研究流程。

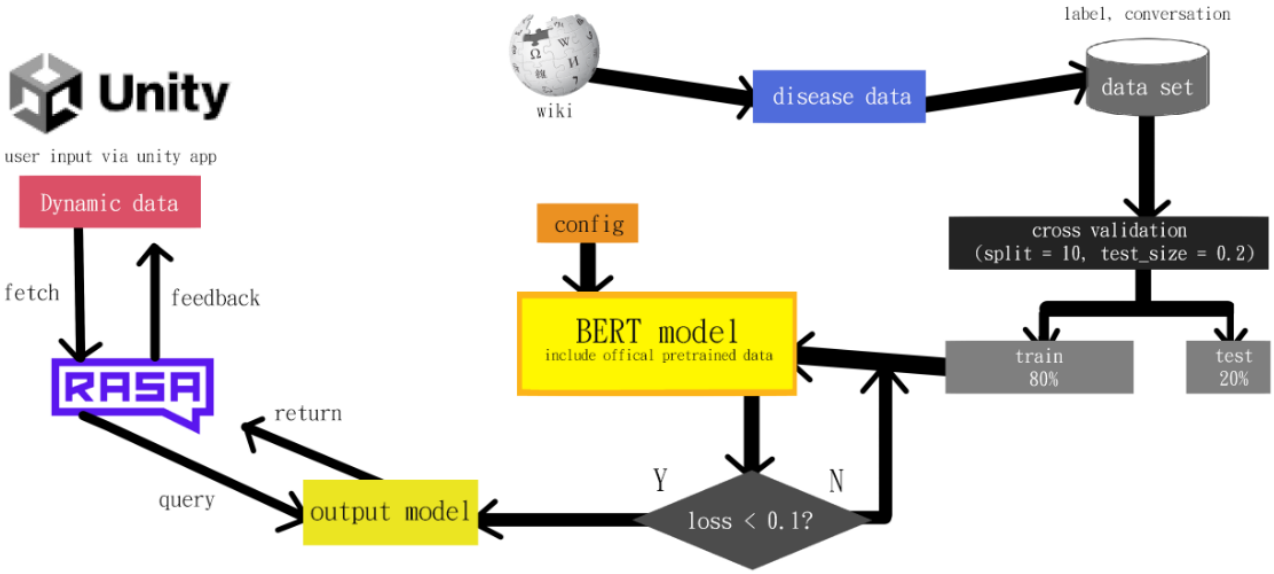

圖 3　專案架構圖

BERT

BERT (Bidirectional Encoder Representations from Transformers) 是一種用於自然語言處理 (NLP) 的預訓練深度學習模型，由 Google 研發。 BERT 能夠對句子進行雙向編碼，也就是同時考慮句子中所有詞的前後文脈信息。這使得 BERT 在許多 NLP 任務中都有出色的表現，包括情感分析、機器翻譯、關鍵字提取和問答系統等。

本專題使用的 BERT 的架構包括了 12 個 transformer 層(見圖 4)，這是一種能夠有效地學習長距離依賴關係的神經網絡架構。 BERT 在訓練過程中使用了一種稱為自適應學習的技巧，使得模型能夠根據輸入的句子動態地適應不同的輸入長度。此外，BERT 也使用了殘差連接和自注意力機制(self-attention)[12]，這使得模型能夠更好地捕捉句子中的相關性。

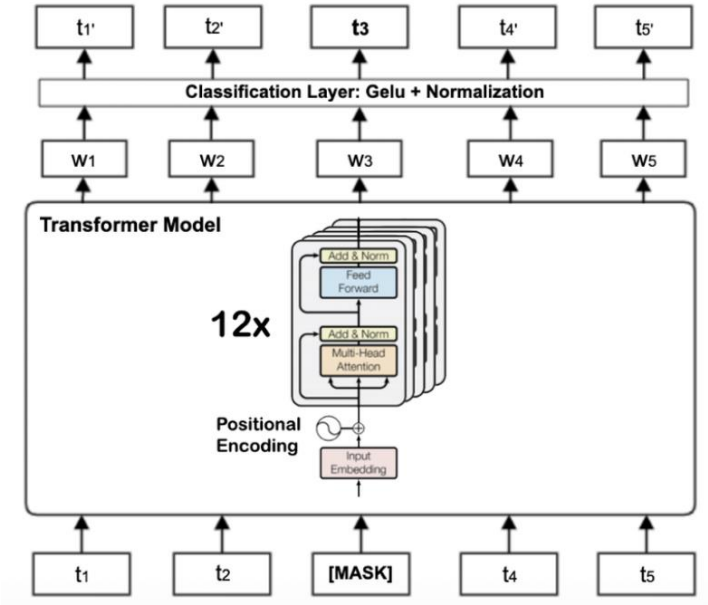

圖 4　BERT 架構圖

數據收集

本專題使用了 BERT 的預訓練模型 bert-base-chinese，進行 BERT 模型 token 的輸入初始化，並加入我們從維基百科蒐集的資料，目前我們的資料集裡面擁有 109 種疾病的資料，疾病和其可能發生的症狀所衍生的對話資訊(見表 5)，資料數量一共有 623 筆。

id	text
流行性感冒	發燒、咳嗽、喉嚨痛、流鼻水、肌肉酸痛、疲倦和頭痛
流行性感冒	我最近常嘔吐和腹瀉
急性咽喉炎	咽部疼痛，吞嚥疼痛及吞嚥困難
....	....
中暑	出現身體發熱、多汗、頭暈、頭痛、口渴、四肢倦怠等症狀
中暑	頭暈、頭痛、口渴、呼吸急促、意識不清等症狀
消化不良	持續性食慾不振、無故消瘦

表 5　BERT 訓練資料

資料訓練

本專題使用 BERT 作為預訓練模型再進行微調 (fine-tuning)， 我們輸入到

BERT 模型中的微調參數一共有四個 input_ids、token_type_ids、attention_mask、labels，其中 input_ids 每個 token 的索引值，用於對應 Token Embeddings。可以用 BERT 模型提供的 vocab.txt 查到；而 token_type_ids 是負責識別句子界限，第一句中的每個詞給 0，第二句中每個詞給 1；界定自注意力機制範圍， 1 是有意義的文本序列內容，進入後續運算，0 則不進入運算；labels 主要是由訓練集所定義，我們先把字串文本轉為整數，再把整數代號視為類別。本專題使用 pytorch 的 AdamW 進行梯度運算，最後在 forword 時，透過 pytorch 的 Linear 分類器辨別輸出的詞向量，判斷輸出的疾病類別，因為 BERT 已經學會了許多有用的語言表徵，所以能夠取得更好的模型結果。

我們輸入資料一次 10 筆(batch_size)，完整資料餵的次數會隨著損失函數而變化，圖 5 我們其中一次的損失函數結果。

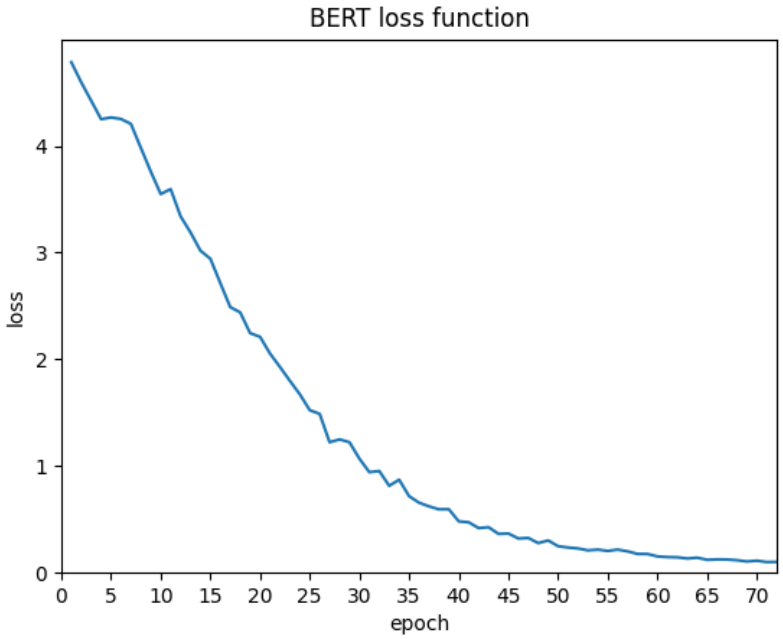

圖 5　BERT loss function

資料集相關性分析

在本研究中，我們使用 BERT 模型來訓練。我們的訓練資料包含 623 筆文本樣本，分佈於 110 個不同的類別。我們將資料拆分為訓練集和測試集，並使用訓練集來訓練。 我們使用這個模型訓練出來的 word embedding 向量來驗證資料集中文字資料在不同類別之間的相關性(圖 6)。我們使用余弦相似度(Cosine Similarity)來計算各個詞彙向量之間的距離，並使用 t 檢定檢驗不同類別之間的距離是否有顯著差異。我們使用左尾檢定法(圖 7)，虛無假設為 “余弦相似度>=0.21”， 對立假設則是: “余弦相似度<0.21”，在對每一個資料集的文字資料都做假設檢定的情況下，我們發現，在不同類別之間的距離有顯著差異 ($p < 0.05$)，代表拒絕虛無假設，並表明這些詞彙在不同類別之間的相關性較低(余弦相似度

<0.21)。此外，我們使用散佈圖來視覺化不同類別之間的距離分佈(圖 8)。我們發現，不同類別之間的距離呈現出明顯的分離情形，進一步證實了這些詞彙在不同類別之間的相關性較低。

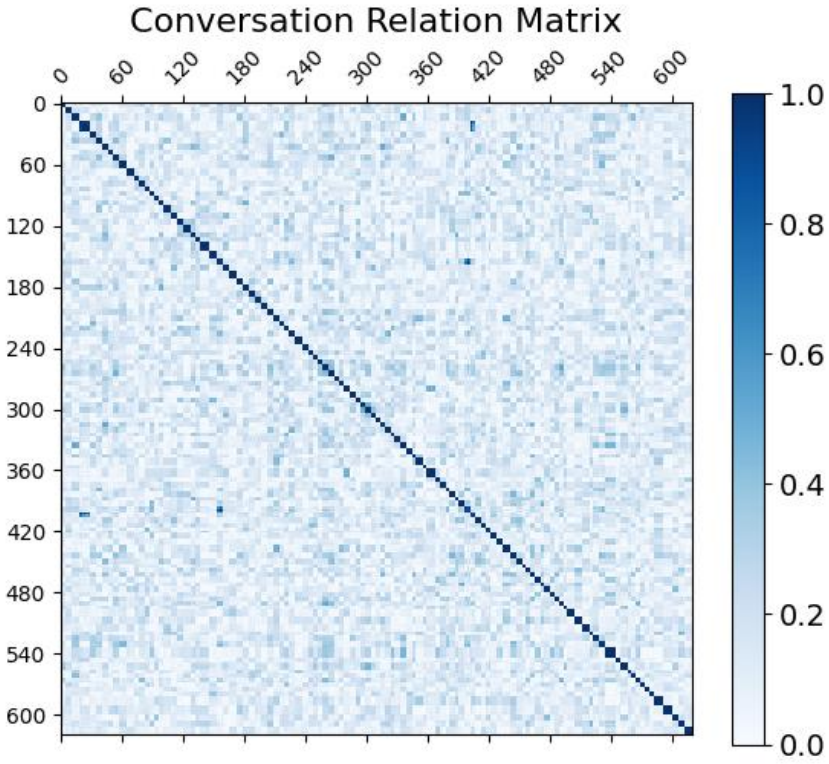

圖 6　word embedding relationship

在圖 7 與圖 8 的部分，因資料較多無法一一顯示，故從資料集中抽取第 1、201、258、301、401、501 筆資料做圖形分析。

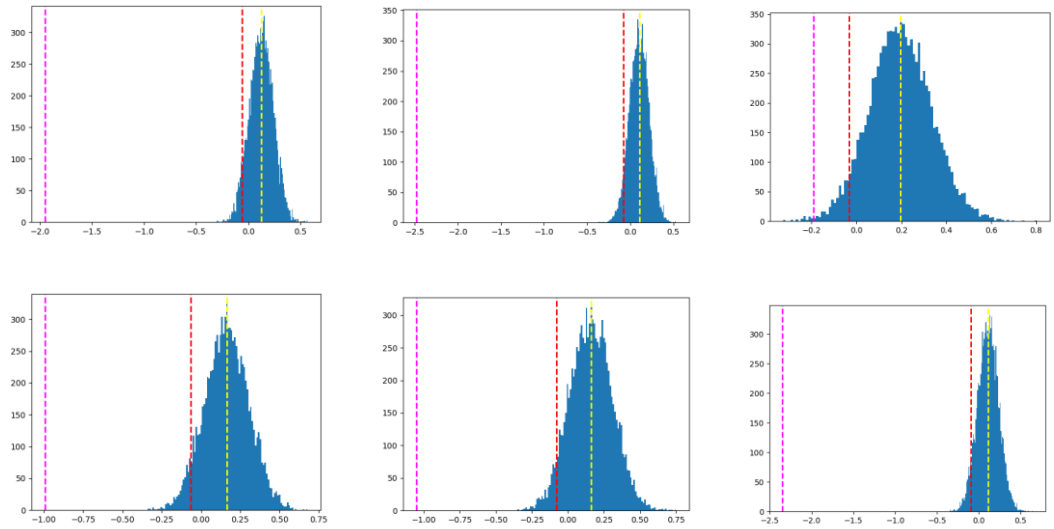

圖 7　hypothesis　testing

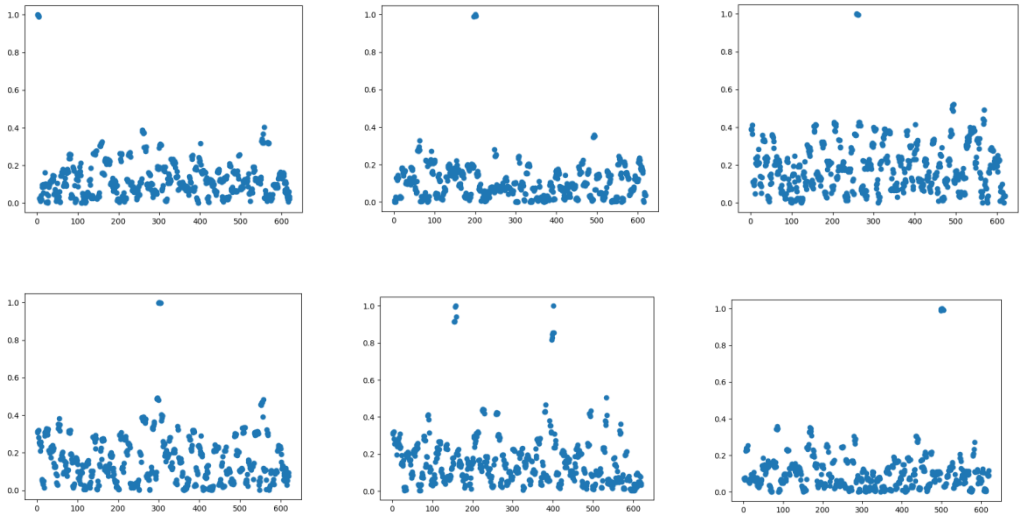

圖 8　word embedding scatter

模型性能評估

本專案使用的訓練方式是使用 Cross validation 進行十次訓練，我們把十次的指標結果相加後平均，最終得到 4 個模型指標: Accuracy、Weight Precision、Weight Recall、F1-Score

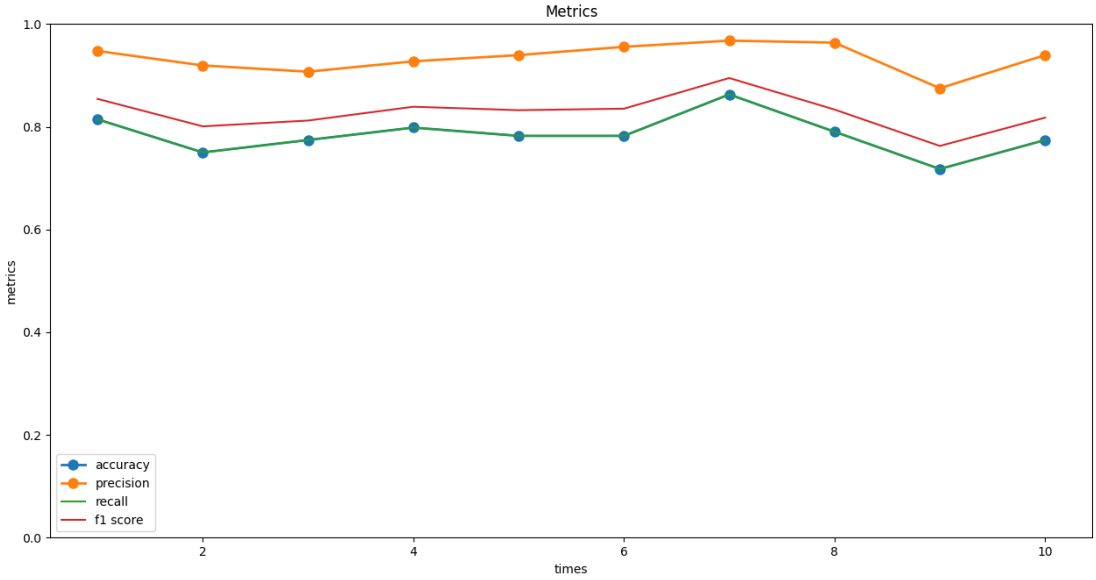

圖 9　BERT model Metrics

Accuracy

所有比較方法將使用準確性評估，公式如(1)。準確性範圍為 0 到 1 之間，越 接近 1 表示準確性越高。除了計算二進位精度外，同時考慮閾值，默認為 0.5。每個預測值都會與閾值進行比較。大於閾值的值設置為 1，小於或等於閾值的值設置為 0。本專題模型的 Accuracy 約為 0.7846，

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \quad (1)$$

Weighted average

因為模型是採用多分類，所以我們使用 weighted average 方法來計算模型指標，該方法給不同類別不同權重，某類別在數據集中出現的比例越大，那他的權重就會越大，每個類別乘權重後再進行相加。該方法考慮了類別不平衡情況，它的值更容易受到常見類（majority class）的影響，但我們的數據分布均勻，所以很適合使用這個指標類別。

Weighted Precision

Precision 是對預測結果而言的，所有被預測為正的樣本中實際為正的樣本的概率。就是在預測為正樣本的結果中，我們有多少把握可以預測正確，公式如(2)。本專題模型的 Precision 為 0.9343。

$$\text{Weighted Precision} = \sum_{i=1 \text{ to } n} w_i \times Precision_i \quad (2)$$

Weighted Recall

Recall 對原樣本而言的，實際為正的樣本中被預測為正樣本的概率。含意類似於: 寧可錯殺一千，絕不放過一個，公式如(3)。本專題模型的 Recall 為 0.7846。

$$\text{Weighted Recall} = \sum_{i=1 \text{ to } n} w_i \times Recall_i \quad (3)$$

F1- score

F1-score 是利用 Precision 和 Recall 的調和值 (2 * Precision * Recall / (Precision + Recall))來預測 Precision 和 Recall 之間的平衡點，Precision 和 Recall 最接近時，F1 會最大，F1、Precision、Recall 這三者皆越大越好模型的。本專題的 F1-score 為 0.8282，可見模型可以對相關的症狀做出準確的疾病推斷。

實際應用

本專案使用 unity 做為開發介面，建立一個小型手機問診 APP，使用者可以根據自身的身體狀況進行提問，而我們透過 server 進行 BERT 模型的回應，讓使用者可以得知自身有可能出現的疾病。

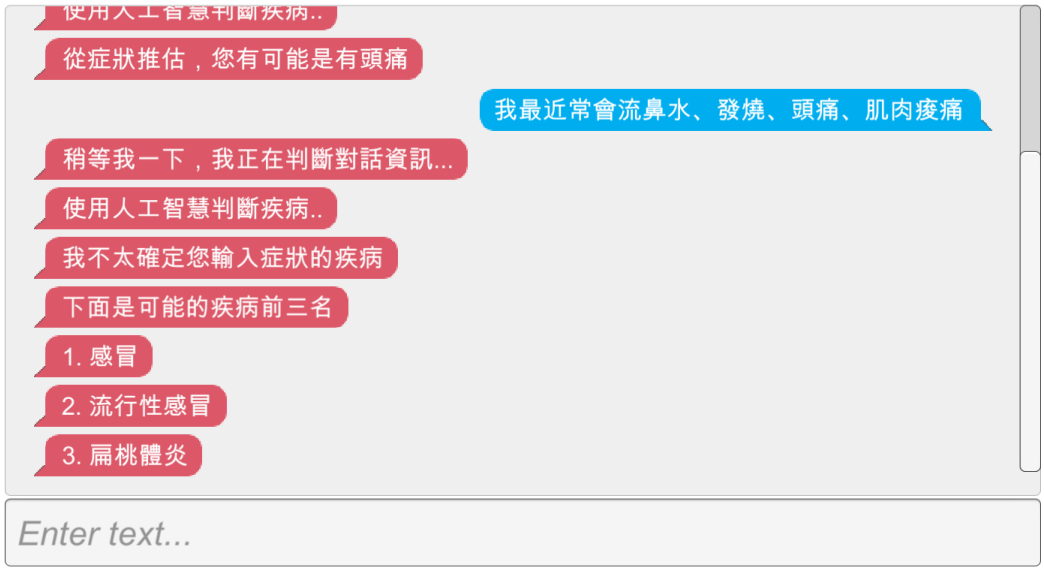

圖 10　使用 BERT 模型進行 AI 問診應用 APP

(五)預期結果

本專案使用一種機器學習方法，可以使用動態數據有效地預測，我們的方法基於 BERT 模型。我們預期，使用 BERT 模型進行疾病對話推斷可以實現較高的準確率。通過對大量的疾病對話數據進行訓練，我們期望 BERT 模型能夠學習到有用的特徵，並能夠有效地解析病人提出的問題，從而有效地診斷疾病。具體而言，本計畫之預期結果如下：

1. 利用深度神經網路進行自然語言理解，能夠分析使用者輸入，從而有效地進行疾病推斷。
2. 使用 Unity 當作 client UI 介面，輸入相關的資訊，輸入會從網路上送到專案指定公共 IP 位置，進行模型預測
3. 根據 BERT 模型訓練出來的模型參數，輸入會導入對應的位置，BERT 模型會輸出對應的疾病。
4. 整理模型結果，得到 BERT 模型的準確度和其他多種指標，並使用假設檢定佐證資料集的類別相關程度低。
5. 本專題未來會加入更多模型訓練，並反饋到資料及，期望本專題的資料集可以延續使用，並讓系統可以造福更多使用者。

(六)參考文獻

1. John Aponte, Kelly Bienhoff, Bob Black, Freddie Bray, Zoe Brillantes, Steph an i e Bur ro ws, Diana Estevez, Juliana Daher,Jacques Ferlay, Marta GacicDobo, Patrick Gerland, , Philippe Glaziou, Lucia Hug, Kacem Iaych, Robert Jakob, Li Liu, Rafael Lozano, Mary Mahy, Colin Mathers, Ann-Beth Moller, William Msemburi, Mohsen Naghavi, Abdisalan Noor, Minal K. Patel and Danzhen You,

WHO methods and data sources for country-level causes of death 2000-2019. Global Health Estimates Technical Paper WHO/DDI/DNA/GHE/2020.2, pp.3-5, 9-15, 26 WHO data

2. Achraf Benba, Abdeliah Jibab, Ahmed Hammouch and Sara Sandabad Voiceprints, analysis using MFCC and SVM for detecting patients with Parkinson's disease, 2015, pp.301-303 SVM Parkinson's disease
3. Sateesh Ambesange, Vijayalaxmi A, Rashmi Uppin, Shruthi Patil, Vilaskumar Patil, Optimizing Liver disease prediction with Random Forest by various Data balancing Techniques, 2020, pp.98-102 RF Liver disease
4. Hwin Dol Park, Youngwoong Han, Jae Hun Choi, Frequency-Aware Attention based LSTM Networks for Cardiovascular Disease, 2018, pp.1503-1505 LSTM Cardiovascular Disease
5. Baohua Sun, Lin Yang, Wenhan Zhang, Patrick Dong, Charles Young, Jason Dong, Michael Lin, Demonstration of Applications in Computer Vision and NLP on Ultra Power-Efficient CNN Domain Specific Accelerator with 9.3TOPS/Watt, 2019, pp.1213-1223
6. Andreas S. Panayides, Amir Amini, Nenad D. Filipovic, Ashish Sharma, Sotirios A. Tsaftaris, Alistair Young, David Foran, Nhan Do, Spyretta Golemati, Tahsin Kurc, Kun Huang, Konstantina S. Nikita, Ben P. Veasey, Michalis Zervakis, Joel H. Saltz, Constantinos S. Pattichis, AI in Medical Imaging Informatics: Current Challenges and Future Directions, 2020, pp.1-15
7. WHO(n.d.), World Health Statistics 2022, May 20, 2022, from https://www.who.int/news/item/20-05-2022-world-health-statistics-2022
8. NaNNN, Super invincible in-depth exploration of multi-class model Accuracy, Precision, Recall and F1-score, May 3, 2022,from https://zhuanlan.zhihu.com/p/147663370
9. Rohit Kundu, Precision vs. Recall: Differences, Use Cases & Evaluation, October 21, 2022, from https://www.v7labs.com/blog/precision-vs-recall-guide
10. Jacob Devlin Ming-Wei Chang Kenton Lee Kristina Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding Oct 11, 2018 , pp.1~16
11. minoguep, rasa_medical_diagnosis_bot, Jun 4, 2020 , from https://github.com/minoguep/rasa_medical_diagnosis_bot
12. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, Attention Is All You Need, Jun 12 2017, pp.1-15