新北市政府 113 年度自行研究報告

應用 AI 機器學習技術於統計調查編碼檢核之研究

研究機關：新北市政府主計處
研究人員：趙唯毓、張弼超
研究期程：113 年 1 月至 12 月

新北市政府113年度自行研究成果摘要表

計畫名稱	應用 AI 機器學習技術於統計調查編碼檢核之研究
期程	113年1月至12月
經費	無
緣起與目的	本處辦理各項政府統計調查，需對相關統計項目進行編碼檢核，而其項目內容屬自然語言，故寫法因人而異。現行主要以人工判斷搭配參考excel詞庫方式進行編碼檢核，其過程不僅耗費人力，亦時有錯誤發生，爰為減輕編碼檢核之人力作業，本研究應用機器學習技術於統計調查編碼作業，使編碼檢核流程自動化，可節省人力並提高處理編碼檢核之效能及確度，以精進政府統計調查工作。
方法與過程	彙整113年1至10月家庭收支記帳調查之消費品項編碼資料，並分為訓練、驗證、調參、測試及實際應用資料集，前2項資料集提供模型訓練及驗證之用，調參集用於調整超參數以選擇最佳模型，接著測試集用於測試最佳模型之有效性，最終將最佳模型作實際應用；研究方法採監督式機器學習模型，進行分類及編碼，首先針對調查項目如消費品項內容(資料X，例如：雞腿便當)，將其中文字轉換為模型可接受之格式，再提供標準答案之編碼(資料Y，例如：90301)訓練模型，另用交叉驗證法避免過擬合等問題，透過上述訓練及驗證資料集之方式，逐步獲得模型之適當參數，驗證模型之有效性；最後將建置完成之模型運用於每月家庭收支記帳調查之消費品項編碼檢核作業，以節省人力並提高編碼檢核效能及確度。

研究結論及未來展望	本研究建立基於深度學習 LSTM 的自動化編碼檢核模型，該模型充分利用深度學習在文本處理與特徵提取方面的優勢，使其能自動從文字特徵中學習分類邏輯。最佳模型在測試資料集準確率達 97.17%，而在實際應用資料集，模型從 3 萬 3,038 筆資料中抓出 1,162 筆可能存在錯誤的編碼，達到減少人工檢核負擔、提升編碼效率與準確度之目的，顯示 AI 機器學習具解決政府統計調查編碼作業困難之潛力。 為追求模型準確率可達99%以上(更準確地抓出錯誤編碼)，以下提出未來可精進之部分。可使用網格搜索等方法更詳細地挑選超參數(如：學習率、批次大小等)，以獲得最適模型。亦可打破 CPU 運算的限制，採用 GPU 運算資源來訓練更強大的模型(例如 BERT)，以應對更複雜的資料處理需求。此外，為確保模型效果穩定，建議定期檢查資料品質並及時更新訓練詞庫，使模型能適應最新的資料特徵變化。
備　　註	

目次

圖次

表次

壹、研究摘要

本研究針對政府統計調查中家庭收支記帳調查之消費品項的編碼檢核流程，提出基於人工智慧(AI)機器學習技術的自動化解決方案。現行作業主要依賴人工判斷與 Excel 詞庫進行編碼檢核，不僅耗費大量人力，且因判斷依據缺乏一致性，常出現錯誤。本研究以減輕編碼檢核的人力負擔為目標，應用深度學習技術結合自然語言處理方法，開發統計編碼檢核模型，期望提升編碼準確性與效率，促進政府統計工作的智能化與自動化。

研究數據來源為 113 年 1 至 10 月新北市家庭記帳調查資料。為建構有效模型，首先對數據進行文本分詞與標記，並透過詞嵌入技術將文字轉換為稠密向量表示，使其適配深度學習網絡。模型採用長短期記憶網絡 (Long Short-Term Memory, LSTM)，其具有處理長序列數據的優勢，能捕捉編碼任務中隱含的上下文語義關聯性。研究過程中使用 K-fold 交叉驗證方法，確保模型的穩定性與泛化能力，同時針對超參數進行調整，包括學習率、批次大小和隱藏層單元數量，進一步提升模型效能。

研究結果顯示，最佳模型在測試集上的準確率達到 97.19%，Top-3 準確率為 98.38%。在應用於實際數據時，從 3 萬 3,038 筆資料中自動檢核出 1,162 筆可能存在編碼錯誤的紀錄，成功減少 96.48% 的人工檢核工作量，這結果顯示，AI 技術在統計編碼應用中能有效降低人力成本並提高編碼準確性。

儘管本研究之模型結果在準確性和效率方面有良好表現，但本研究也發現，模型表現會因數據分布不平衡、標記樣本數量不足等因素而有所限制。因此，固定規則與模型學習結果的結合仍需進一步探索，以滿足實務需求的變化。為此，未來研究可考慮採用網格搜索 (Grid Search) 和自適應學習方法進一步優化模型超參數。此外，引入更高效的 GPU 訓練框架如 BERT 等預訓練語言模型及大規模數據增強技術，也有助於提升模型的適應能力。

綜上所述，本研究展現了 AI 深度學習技術在政府統計調查領域的廣泛應用潛力，透過構建基於 LSTM 的自動化統計編碼檢核模型，不僅完成高準確度的文本分類，還有效減輕人工工作負擔，提供未來推動智能化政府統計工作之參考。

貳、研究背景及主旨

一、研究背景

政府統計調查涵蓋多項專業數據處理工作，涉及行業別、職業別及消費品項的分類與編碼檢核。這些編碼檢核作業對於確保調查數據的一致性與準確性具有關鍵作用。然而，現行的作業模式主要依賴人工判斷，輔以 Excel 詞庫完成，導致效率不足、人力成本高昂，且容易發生錯誤，對統計工作的品質與時效性造成不利影響。

隨著人工智慧(AI)領域中的機器學習(Machine Learning, ML)技術快速發展，這些技術在自然語言處理(NLP)和自動分類領域的應用日益廣泛，為編碼檢核作業的自動化提供了可行的解決方案。通過結合統計詞庫與機器學習模型，系統能夠高效處理大量文本數據，實現精準的分類與匹配，不僅顯著提升檢核效率，還能有效降低錯誤率。這一技術進步為政府統計編碼流程的優化提供了全新的可能性，將有助於大幅提升統計工作的整體效能與可靠性。

二、研究主旨

本研究旨在運用機器學習領域中的子集合，深度學習技術開發一套自動化的編碼檢核系統，以解決現行統計調查編碼流程中存在的人力耗費高、錯誤率高及效率不足等問題。研究將結合自然語言處理技術與深度學習模型，透過對歷史編碼數據及詞庫的深度學習，實現對記帳消費品項的自動分類與檢核。預期此系統能顯著提升編碼檢核的準確性與處理效率，同時減少人工作業負擔，最終助力政府統計調查工作邁向數位化與智能化，滿足現代統計需求。

參、 相關研究及文獻探討

一、自然語言處理（Natural Language Processing, NLP）

自然語言處理是人工智慧與計算語言學的交叉領域，其核心目的是使機器能夠理解、解釋和生成人類語言。隨著深度學習技術的發展，NLP 在文本分類、情感分析、機器翻譯、自動摘要和信息檢索等領域取得了顯著突破。傳統 NLP 技術依賴規則式方法與統計模型，例如基於 n-gram 的語言模型和隱馬爾可夫模型(HMM)，但這些方法在處理語意複雜或長距離依賴的文本時效果有限。

近年來，深度學習模型(如 RNN、LSTM 和 Transformer)的引入為 NLP 提供了新的解決方案。特別是基於 Transformer 架構的模型（如 BERT、GPT 系列），通過預訓練與微調的方式，在各類 NLP 任務中取得了卓越的表現。BERT（Bidirectional Encoder Representations from Transformers）通過雙向上下文的編碼方式，顯著提升了模型對語義的理解能力，被廣泛應用於命名實體識別（NER）、文本分類及關係抽取等應用場景。而 GPT 系列模型則在生成式任務（如文章生成和對話系統）中展現極大的潛力。

在分類與檢核相關研究中，機器學習和 NLP 的結合尤為重要。例如，使用 Word2Vec 或 GloVe 等詞嵌入技術進行特徵提取，搭配 SVM 或隨機森林等傳統分類模型，已被成功應用於行業分類和產品分類。然而，深度學習方法逐漸取代傳統技術，特別是使用 Transformer 模型進行多標籤分類，展現了更高的準確率與效率。此外，結合統計詞庫與語意分析的模型能夠更準確地匹配類別，提高自動化分類的精度。

總體而言，NLP 技術的持續進步正推動自動化數據處理的革新，尤其在文本分類與檢核應用中，結合深度學習與專業詞庫

的研究方向成為當前的重點。這些進展為本研究開發統計編碼檢核系統奠定了堅實的技術基礎，但基於 Transformer 架構的模型(如 BERT、GPT 系列)運算量龐大，通常需使用 GPU 之運算能力，故本研究僅考慮可用於 CPU 計算之模型，如 RNN 及 LSTM。

二、機器學習於政府統計之應用

機器學習在政府統計領域的應用正逐漸成為一個重要的研究方向。隨著數據規模的快速增長與數據來源的多樣化，傳統統計分析方法在面對結構化與非結構化數據時面臨挑戰，而機器學習以其強大的模式識別與預測能力，為政府統計數據的處理與分析提供了新工具。聯合國歐洲經濟委員會（以下簡稱 UNECE）於 2019 年成立 ML 專案小組，嘗試將 ML 技術應用於精進政府統計作業，其主要針對「資料分類與編碼」、「資料檢核與插補」及「影像資料探勘」三類主題規劃研究。

與本研究主旨相關之資料分類與編碼目的在於有效整合多態樣的輔助資源與政府統計資料，建置自動分類及編碼系統。美國勞工統計局 (BLS) 按年辦理職業傷害及疾病調查，需對職業別、傷害或疾病性質等調查項目進行編碼，傳統手工分類方法需耗費大量人力且容易出錯。為應對此挑戰，BLS 開發了基於自然語言處理（NLP）的機器學習模型，使用了詞嵌入技術（如 Word2Vec）和長短期記憶（LSTM）網絡，從文本描述中學習職業相關的語義和分類特徵。該模型通過自動化分析職業描述、職責和技能要求等文本數據，準確預測出對應的職業分類編碼。隨著模型的應用，BLS 將職業分類的準確性提高了 85% 以上，且大幅減少了人工審核的時間，使分類流程更具效率。

另一方面，歐洲統計局（Eurostat）在產業分類的自動化過程中面臨跨語言、跨國界的挑戰，特別是在需要將描述不一的產業資料統一編碼時，傳統方法難以適應。為此，Eurostat

應用了基於深度學習的多語言變壓器模型（如 XLM-R）進行產業文本的語義分類。該模型能夠從數據集中學習不同語言之間的語意關聯，進行準確的產業分類，同時滿足跨國標準的要求。研究表明，Eurostat 使用這一方法後，跨國產業分類的準確率提升了 20%，並且能夠適應數據持續更新的需求，顯著減少了人員的編碼負擔。

這些成功實例顯示了機器學習技術在政府統計編碼中的潛在效能，尤其在需要大量文本資料分類的應用場景中，機器學習能有效地解決多語言處理、非結構化文本分析和編碼標準化等難題，為未來的政府數據分類工作提供了可靠的技術支持。

肆、 研究方法

一、研究流程

以下為研究流程，本研究使用 Microsoft Excel 試算表進行資料收集作業，並撰寫 Python 程式語言執行資料預處理至研究成果等作業。

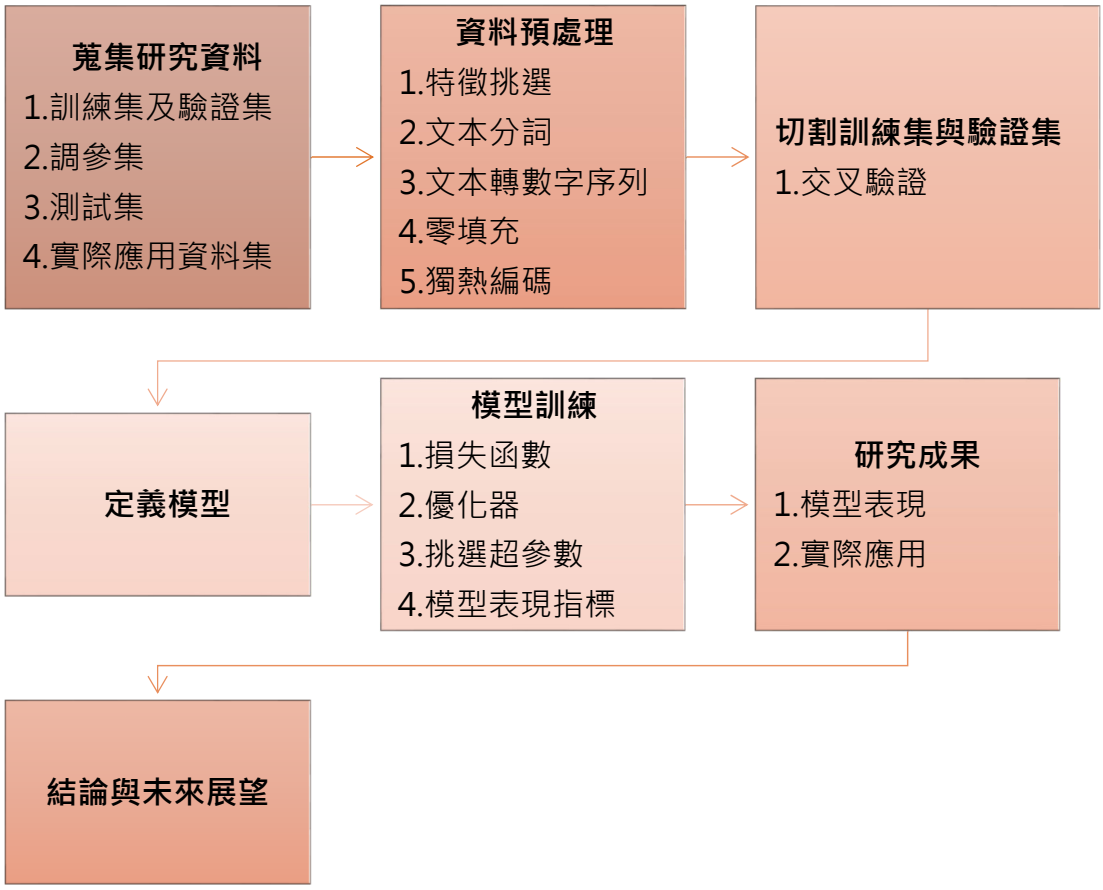

圖 1 研究流程圖

二、研究資料

本研究蒐集 113 年 1 至 10 月新北市家庭收支記帳調查之消費品項編碼原始資料計 27 萬 2,684 筆，其中編碼(依變數，又稱標籤)共 808 個類別，而 1 至 9 月調查資料已完成人工編碼檢核，即標籤正確無誤，10 月調查資料則尚未完成檢核，即標籤可能

有誤。

（一）訓練集與驗證集：113 年 1 月至 7 月新北市家庭收支記帳調查原始資料(計 19 萬 278 筆)，用於模型訓練及驗證與調整模型。

（二）調參集：113 年 8 月新北市家庭收支記帳調查原始資料(計 4 萬 2,478 筆)，用於調整超參數，以提升模型性能。

（三）測試集：113 年 9 月新北市家庭收支記帳調查原始資料(計 3 萬 9,929 筆)，用於測試模型之成效。

（四）實際應用資料集：113 年 10 月新北市家庭收支記帳調查原始資料(計 3 萬 3,038 筆)，用於觀察模型是否實際抓出有誤編碼。

三、資料預處理

（一）特徵挑選

為提升模型性能並降低電腦計算成本，選擇與分類消費品項規則關聯性較高的特徵。原始資料中的變數包含消費品項、單位、數量、是否購自餐飲服務、購買地點代碼、金額及分類消費品項之科目代碼(表 1)。依過往經驗法則，選擇消費品項、是否購自餐飲服務及購買地點代碼，這些類別型特徵與分類規則有直接關係，故作解釋變數(表 1 淺澄色)訓練模型。科目代碼為資料標籤，其為類別（categorical）變數(表 1 澄色)，因此本研究為一個分類問題。

表 1 新北市家庭收支記帳調查原始資料內容範例

消費品項	單位	數量	是否購自餐飲服務*	購買地點代碼	金額(元)	科目代碼
咖啡	杯	1	否	4	40	41110
咖啡	杯	1	是	8	40	90401

*註：餐飲服務係指販售者有現場調理動作，可即時依消費者需求調整口味，且供立即食用者。

（二）文本分詞

文本分詞是自然語言處理中的一個基本步驟，用於將連續的文本分割成更小的單位，如單詞、子詞或短語。這一過程為下游任務（如文本分類、情感分析、機器翻譯）提供基礎數據單位。分詞的目的是將語言的連續性和複雜性轉化為可以被機器理解和處理的形式。

一般中文文本分詞是將一連串文字切割成多個詞彙，如雞腿便當即會切割為「雞腿」、「便當」，而下一步驟會將這些詞彙轉成數字序列方便電腦處理，這時需建立字典，用字典中所含之詞彙轉換成對應數字。但字典裡的詞彙有限，尤其是面對稀有詞、新創詞或拼寫錯誤時，傳統方法如基於詞的標記化(Tokenization)往往無法處理。故本研究使用 WordPiece 分詞技術(表 2)，將單詞拆分為更小的子詞單位（subwords），從而有效應對字典裡詞彙不足的問題。本研究使用 Hugging Face 團隊預訓練繁體中文的 transformers 模型進行文本分詞。

表 2 消費品項中文字文本分詞範例

消費品項	基於詞的標記化	WordPiece Tokenization（基於子詞）
雞腿飯	‘雞腿’，‘飯’	‘雞’，‘腿’，‘飯’
雞蛋布丁	‘雞蛋’，‘布丁’	‘雞’，‘蛋’，‘布’，‘丁’

（三）文本轉成數字序列

文本是一種非結構化數據，為了讓機器理解，需要轉化為結構化數據。通過將句子中的單詞（或其他語言單位，如子詞或字符）映射到一個有限集合（即詞彙表）的數字索引(表 3)，實現模型可操作的數值形式。詞彙表（Vocabulary）是一個包含所有文本語料中可能出現的語言單位（Tokens）的集合。雖然數字序列本身是符號化的，但通過本研究後續定義模型段落

之嵌入層（Embedding Layer）可以將這些數字映射為密集向量（dense vectors），以保留語義信息，為建模提供基礎。

表 3 消費品項中文字文本轉成數字序列範例

消費品項	WordPiece Tokenization	數字序列
雞腿飯	‘雞’, ‘腿’, ‘飯’	[7430, 5597, 912]
雞蛋布丁	‘雞’, ‘蛋’, ‘布’, ‘丁’	[7430, 2791, 5158, 2353]

（四）零填充（Zero Padding）

在自然語言處理或時間序列建模中，不同輸入序列的長度可能不一致，而深度學習模型通常要求輸入序列的形狀固定，方便批次計算（Batch Processing）。故常用於解決此問題的方法為零填充，零填充即對短於目標長度的序列填充零值，長於目標長度的序列則刪除序列長度至目標長度，使所有序列的長度相等。使用零值填充而非其他數值的原因在於零值能降低數據範圍擴大的風險，保持模型訓練的穩定性。本研究取原始數字序列中最長之序列 26 為基準，將長度低於 26 之其他序列在詞彙前面補零，使所有序列的長度一致。

表 4 消費品項數字序列零填充範例

消費品項	WordPiece Tokenization	數字序列補零
雞腿飯	‘雞’, ‘腿’, ‘飯’	[0,...,0, 7430, 5597, 912]
雞蛋布丁	‘雞’, ‘蛋’, ‘布’, ‘丁’	[0,...,0,7430, 2791, 5158, 2353]

（五）獨熱編碼（One-hot Encoding）

獨熱編碼是將類別型數據轉換為數值表示的一種常用技術，其核心概念是將每個類別表示為一個高維的稀疏向量，保留類別之間的獨立性，以及使類別之間沒有順序性，避免直接數值編碼帶來的隱含偏見。

編碼方法為假設有 A 個類別，每個類別對應到一個 A 維向量。向量中的第 i 個位置對應第 i 個類別，該位置值為 1，其他位置值為 0。本研究所含類別變數有是否購自餐飲服務、購買地點及科目代碼，皆對其做獨熱編碼轉化為模型可理解之數值形式。

表 5 以類別變數購買地點進行獨熱編碼範例

購買地點	獨熱編碼
1.百貨公司	[1,0,0,0,0,0,0,0,0,0]
2.超市	[0,1,0,0,0,0,0,0,0,0]
3.量販店	[0,0,1,0,0,0,0,0,0,0]
4.連鎖便利商店	[0,0,0,1,0,0,0,0,0,0]
5.市場	[0,0,0,0,1,0,0,0,0,0]
6.行動商店及攤販	[0,0,0,0,0,1,0,0,0,0]
7.特定商店	[0,0,0,0,0,0,1,0,0,0]
8.其他商店	[0,0,0,0,0,0,0,1,0,0]
9.網路商店	[0,0,0,0,0,0,0,0,1,0]
10.無實體店鋪	[0,0,0,0,0,0,0,0,0,1]

四、 切割訓練集與驗證集

在資料分析研究中，測試集是不參與模型訓練的資料，僅用於評估模型最終的性能，但為了避免模型過擬合(Overfitting)並提升泛化能力(Generalization ability)，除了一筆資料用於訓練模型外，我們仍需一筆資料用於驗證模型訓練成效是否良好，並根據成效調整模型參數。故通常會將訓

練集資料切割成訓練集及驗證集，皆是用於調整模型參數之資料集。

本研究使用 K 折交叉驗證法來訓練及驗證模型，交叉驗證是一種提高模型泛化能力的數據評估方法，通過重複切分和訓練模型來獲得更穩定的性能估計。以下是 K 折交叉驗證法步驟：將數據隨機分為 k 個不重疊的子集，每次選取其中 k－1 個子集作為訓練集，剩下的一個作為驗證集，依次輪轉，進行 k 次，本研究取用 k 值為 5（表 6）。交叉驗證確保所有數據都參與模型的訓練與驗證，並平滑了因單一切分帶來的隨機性影響，解決模型過擬合與欠擬合(Underfitting)問題，提高評估結果的穩定性和可靠性。

表 6 交叉驗證範例

第 k 次	訓練及驗證資料集				
1st	D1	D2	D3	D4	D5（驗證集）
2nd	D1	D2	D3	D4（驗證集）	D5
3rd	D1	D2	D3（驗證集）	D4	D5
4th	D1	D2（驗證集）	D3	D4	D5
5th	D1（驗證集）	D2	D3	D4	D5

五、定義模型

人工智慧領域之子集合深度學習，其模型能夠自動擷取資料特徵並進行有效處理，特別是在自然語言處理等應用中展現了顯著的效果。遞歸神經網絡(RNN)被廣泛運用於自然語言處理，因其能夠處理序列資料並捕捉時間步長之間的關聯性。然而，傳統的 RNN 在長序列資料處理上存在梯度消失和爆炸的問題，因此長短期記憶網絡(LSTM)作為一種改良的 RNN 模型，有效解決這些問題，在各類自然語言處理任務中展現了更強的表現，故本研究選擇 LSTM 作深度學習統計編碼模型。本研究使用 Python 編

寫而成的開源神經網路庫(Keras)定義模型，模型共有三層，每一層各有其功能，如下所列。

（一）詞嵌入層(Embedding Layer)

詞嵌入層是深度學習中設計以處理文本數據的一層，其將離散的詞彙(如單詞或子詞)轉換成一個實數詞向量(Vectors of real numbers)表示，更精準地說法是張量(Tensor)。這些向量捕捉詞彙間的語義和語法關係，使語義上相似的詞在嵌入空間(連續向量空間)中彼此接近，為後續深度學習模型提供更加結構化的輸入。詞嵌入層的核心是嵌入矩陣E，大小為$V \times D$，其中V是詞彙表的詞彙量，D是嵌入向量的維度，本研究設定V=25,000，代表我們限制詞彙表只能包含 25,000 個詞彙，一旦詞彙量達到這個大小以後，剩餘的新詞彙都會被視為未知，以避免詞彙表過於龐大，而 D 固定為 256 維。詞嵌入層的計算方式是將輸入文本(數字序列形式)傳遞到詞嵌入層，層內根據數字索引查找嵌入矩陣中的對應向量，並輸出一個多維矩陣，其每一行代表一個詞的嵌入向量。

本段落使用簡單示例說明詞嵌入層運作方式，假設詞彙表包含 ["貓", "狗", "狼"]，對應的索引為 [0, 1, 2]，嵌入矩陣大小為 3x2（詞彙數量為 3，嵌入維度為 2），嵌入矩陣：

$$E = \begin{bmatrix} 0.6 & 0.2 \\ 0.5 & 0.3 \\ 0.3 & 0.8 \end{bmatrix}$$

代表狗(索引 1)的詞嵌入層輸出向量即為[0.5,0.3]，這一嵌入向量將參與後續模型的學習和運算。另可觀看貓與狗在語義關係上較相近，故兩個在嵌入空間中的向量距離較近，而與狼的詞向量較遠(圖 2)。

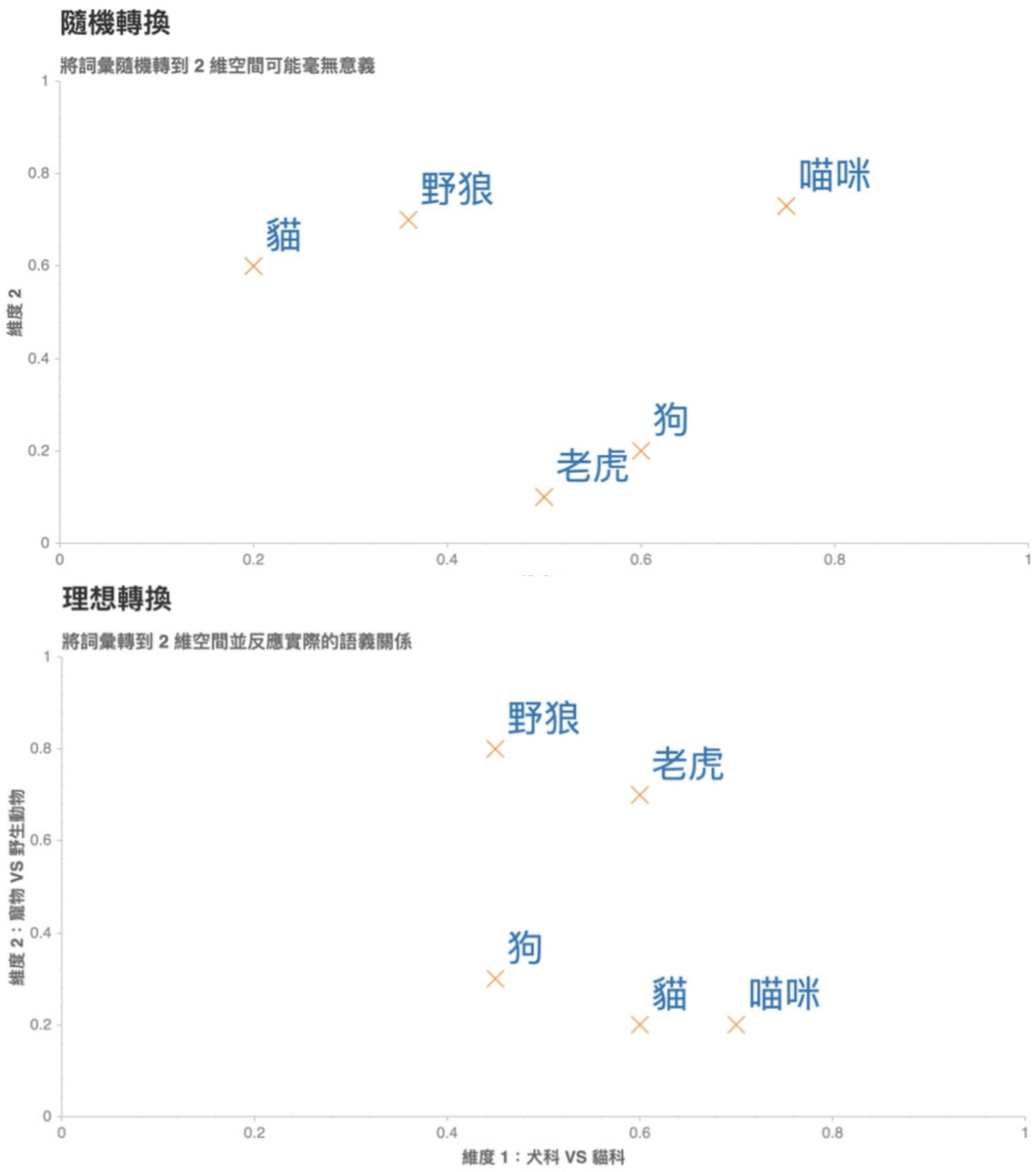

圖 2 詞嵌入層示例圖

資料來源：https://reurl.cc/862Y6d

（二）長短期記憶模型（Long Short-Term Memory, LSTM）層

選擇 LSTM 原因為其具有「記憶單元」與「門機制」(如輸入門、遺忘門、輸出門）等特性。記憶單元可以保存長期的上下文信息，其梯度在反向傳播中得以保留，因而避免了梯度的消失或爆炸，且記憶單元內的狀態通過加法(累加更新記憶)來替代傳統 RNN 中的逐層非線性變換(乘法)進行信息傳遞，能防止

梯度指數級變化。門機制則使用三個閘門來動態調節信息的流入、更新和輸出，以下分別說明：

1. 遺忘閘(Forget Gate)：確定在特定時間步長中應從記憶體中省略哪些資訊，即查看前一狀態(h_{t-1})和當前輸入 x_t，並計算函數。

$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] + b_f\right)$$

W 是權重矩陣，負責將輸入和隱藏狀態(Hidden State)轉換為每個門的輸入信號，形狀依據輸入和隱藏層大小而定。b 是偏誤項(Bias)，用於調整每個門的輸出，賦予模型更多的靈活性。σ 是 sigmoid 函數，輸出值範圍為[0,1]，表示記憶的保留比例。

2. 輸入閘(Input Gate)：決定新的信息是否加入記憶單元。

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

與 i_t 配合的候選記憶 $\tilde{C}_t$，決定信息的重要程度：

$$\tilde{C}_t = tanh(W_C \cdot [h_{t-1}, x_t] + b_C)$$

3. 輸出閘 (Output Gate)：決定記憶單元的哪些部分作為當前時間步的輸出。

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

4. 記憶單元更新(Cell State Update)：綜合遺忘閘和輸入閘，更新當前的記憶狀態。

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$$

5. 隱藏狀態：根據輸出閘和記憶單元的當前狀態計算出當前時間步的隱藏狀態。

$$h_t = o_t \cdot tanh(C_t)$$

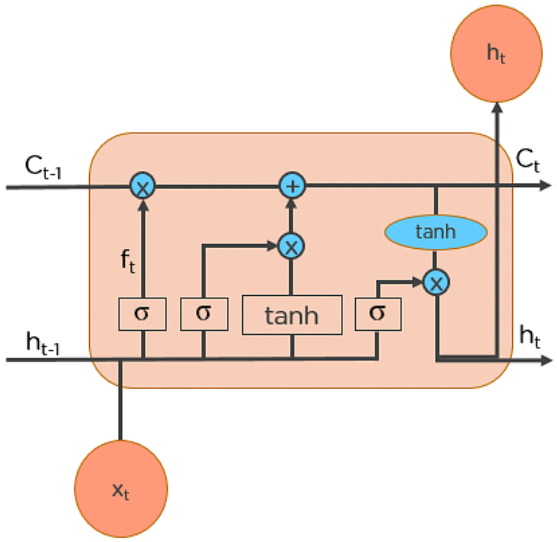

圖 3 LSTM 模型運作示例圖
資料來源：https://reurl.cc/xpEbrL

（三）全連接層(Fully Connected Layer)

全連接層是人工神經網絡中的基本組成部分，其主要功能是將前一層的輸出與下一層的每個神經元完全連接，實現不同特徵的線性組合，從而進行更高層次的特徵提取或分類。在全連接層中，每個神經元與上一層的所有神經元相連，並將輸入數據映射到新的特徵空間。連接的權重和偏誤是可訓練參數，會在模型訓練過程中進行更新。給定上一層的輸出 x（大小為樣本數 n），全連接層的操作可以表示為：

$$y = f(Wx + b)$$

其中 W 為權重矩陣，大小為$m \times n$，m是當前層的神經元數量。b 是偏誤向量，大小為m。f 為激活函數，本研究使用 Softmax 函式，Softmax 函式能將某層中的所有神經元中的數字作正規化，即壓縮到 0 至 1 之間，並讓所有數字的和等於 1，以解釋為對應分類的發生機率。本研究設置全連接層作為神經網絡的最後一層，將高層特徵映射到目標空間，進行分類(輸出機率分布)。

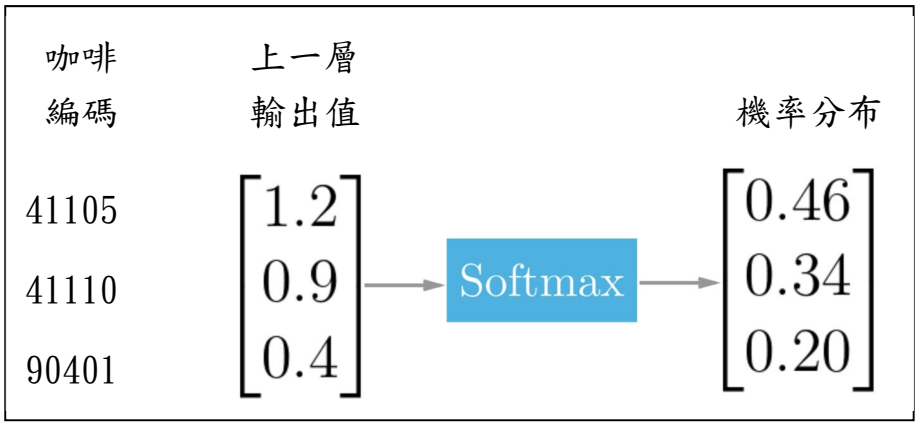

圖 4 全連接層運作範例圖

（四）神經網路架構

使用 model.summary()來看每一層的參數量以及輸出的張量（圖 5）。Param#為每一層所包含的模型參數（Parameters），在深度學習的過程中，這些參數都會不斷地被調整，直到能讓模型做出好的預測。本研究模型共計 670 萬 1,352 個參數，其中詞嵌入層包含最多參數，原因是為了將詞彙表裡的每個詞彙都建立一個 256 維度的詞向量，因此參數量計 640 萬(25,000 × 256)個最多。

本研究每個樣本輸入模型的張量形狀是(None, 38)，即張量的特徵維度是 38 維，其中由 26 維消費品項中文字轉數字序列、10 維購買地點之獨熱編碼及 2 維是否購自餐飲服務之獨熱編碼所構成；LSTM 層設定 128 維輸出向量，全連接層(Dense)輸出 808 維向量，表示本研究 808 個類別之機率分布，另所有層的張量形狀的第一個維度都是 None，其代表可以是任意的數字。

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 38)	0
embedding (Embedding)	(None, 38, 256)	6,400,000
lstm (LSTM)	(None, 128)	197,120
dense (Dense)	(None, 808)	104,232

Total params: 6,701,352 (25.56 MB)
Trainable params: 6,701,352 (25.56 MB)
Non-trainable params: 0 (0.00 B)

圖 5 神經網路架構圖

六、模型訓練

（一）損失函數

損失函數是模型訓練用來衡量預測結果與實際結果之間差距的函數。在訓練過程中，目標是透過調整模型參數以最小化損失函數的值，從而提高預測的準確性。損失函數根據具體的學習任務和數據類型的不同，可以分為多種類型。在本研究多類別分類問題中，類別交叉熵(Categorical Cross-Entropy)是最常用的損失函數之一。該損失函數對每個樣本計算預測機率分佈與真實標籤之間的交叉熵，對於每個樣本 i，其交叉熵損失函數 L_i可以表示為：

$$L_i = -\sum_{a=1}^{A} y_{i,c} \log(p_{i,a}), \qquad i = 1, \ldots, \text{訓練集樣本數 n},$$

目標： $\min \sum_{i=1}^{n} L_i$,

其中 A 是類別總數，本研究分類數量共 808 個類別。$y_{i,a}$ 是樣本 i 在類別 a 上的值。如果 $y_{i,a} = 1$，則表示樣本 i 屬於第 a 類別；如果 $y_{i,a} = 0$，則表示不屬於該類別。$p_{i,a}$ 是模型對樣本 i 屬於第 a 類別的預測機率。

預測機率 p 越接近真實標籤 y，交叉熵的值越小，表示模型預測越準確。本研究模型的輸出是每個類別的預測機率，類別交叉熵能夠處理這些多維的機率分佈並計算預測誤差，其亦對預測錯誤的情況給予較大的懲罰，幫助模型加速學習，高效地衡量和指導模型參數的優化過程。

（二）優化器(Optimizer)

優化器是模型訓練過程中的核心組件，用於調整模型的參數(如權重和偏差)以最小化損失函數，從而提高模型對訓練數據和測試數據的預測能力。常用的最佳化算法是梯度下降法（Gradient Descent)，其原理是給定損失函數 $L(\theta)$ 和參數初始值 θ_0，每次迭代計算梯度 $\nabla L(\theta_t)$，並以負梯度方向更新參數，第 t 次迭代：

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla L(\theta_t)$$

其中 $\eta > 0$ 為學習率，決定每次更新的步伐。梯度 $\nabla L(\theta_t)$ 指示了損失函數在參數 θ_t 的位置下降最快的方向。重複執行梯度下降法以更新參數，直到本研究所設定之訓練次數(超參數)為止。

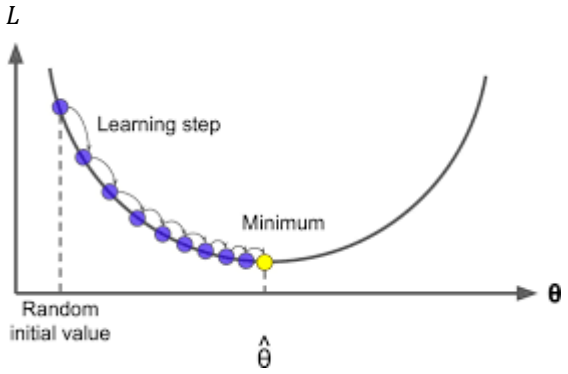

圖 6 梯度下降法示意圖

本研究訓練模型亦在 Keras 上進行，優化器使用 Adam(Adaptive Moment Estimation)方法，其由 Kingma 和 Ba 於 2014 年所提出，為一種基於梯度下降的最佳化算法。它的優勢

包括能夠根據梯度的特性自適應地調整學習率，適合不平穩的目標函數，且利用動量法加速訓練過程，再加上引入二階動量估計，有效地處理稀疏梯度，提升模型穩定性。

（三）挑選超參數（Hyperparameters）

超參數是在訓練模型開始之前設置的參數，這些參數不直接從數據中學習調整，而是通過外部調整來影響模型的訓練過程和性能。訓練過程相關超參數有學習率、批次大小(Batch Size)、期(Epoch)。超參數的選擇對模型性能有著至關重要的影響，因此需要通過系統化的方式進行挑選和優化。挑選的目標除了提高模型性能，亦防止過擬合，即避免模型過於貼合訓練數據，導致泛化能力下降，並且希望在有限資源下，找到表現最優的配置。本研究為降低調參時間成本，根據經驗和對模型的理解，手動階段化調參，先對最重要的超參數進行調整，逐步引入次要參數，以減少搜索空間。

本研究選擇學習率、批次大小及期，此三個超參數作調整，並以模型在調參集資料的準確度為準則，選出最佳模型，最後再使用最佳模型預測測試集及實際應用，觀察模型能力，此實驗將於第伍章節研究成果呈現。本研究調參步驟如下：

1. 固定批次大小=1024、期=60，設定學習率=0.0005，且與較大之學習率=0.005 模型比較，得學習率=0.0005 模型預測調參集的準確度較高，故改與較小之學習率=0.0001 模型比較，學習率=0.0005 模型準確度仍較高，故視學習率=0.0005 為模型的最佳配置。

2. 固定學習率=0.0005、批次大小=1024，設定期=60，同上步驟設定值(設定期=20, 40, 60)往上或往下得出模型的最佳配置，準確度最高模型期=40。

3. 固定學習率=0.0005、期=40，設定批次大小=1024，同上步驟設定值(設定批次大小=1024, 512, 256)往上或往下得出模型的最佳配置，準確度最高模型批次大小=512。

4. 本研究視學習率=0.0005、期=40、批次大小=512 為模型最佳配置，並對此模型進行最終的測試評估與實際應用。

（四）模型表現指標

在機器學習中，評估模型表現的指標至關重要，因為這些指標能夠幫助我們了解模型在特定任務中的效果。常見的模型表現指標包括準確度（Accuracy）、精確度（Precision）、召回率（Recall）、F1 分數等。在推論表現上，本研究使用了以下指標來總結模型配適的表現：

1. 準確度：準確度 = 正確預測的樣本數/總樣本數。表示模型正確預測的比例，但在類別不平衡情況下，單純依賴準確度可能會掩蓋少數類別的問題。

2. 宏平均 F1 分數（Macro-averaged F1 Score）：

$$\mathrm{F1} = 2 \times \frac{\text{精確度} \times \text{召回率}}{\text{精確度} + \text{召回率}}，\mathrm{F1}^{macro} = \frac{1}{A}\sum_{a=1}^{A} F1_a$$

其中宏平均指不考慮類別的樣本數，對所有類別的指標進行平均。精確度指預測為某一類別的所有樣本中，正確預測的樣本數比例。召回率指真實為某一類別的所有樣本中，預測正確的比例。宏平均 F1 分數能平衡不同類別的預測表現，特別適合處理多分類且類別不平衡的情況。

3. Top-K 準確度：指由高至低排序模型對每個類別的預測機率，真實的標籤類別若出現在前 K 個預測機率高的類別中，視為正確， K = 1 即為準確度。該指標在多分類問題中能提供

更有彈性的衡量標準，適用於推薦系統等問題上。本研究旨在檢誤統計編碼，不局限於模型給定一個標準答案，可參考複數個最可能的編碼，故使用此指標(設定 K=3)衡量模型表現。

4. Cohen’s Kapp：用於衡量模型預測和實際標籤之間的一致性，校正了隨機一致性的影響。類別不平衡時，模型可能傾向於預測為樣本數最多的類別，Cohen’s Kappa 通過考慮隨機預測的情形，能反映模型的實際性能，而非僅關注樣本多的類別。一般認為 $\kappa > 0.6$ 表示一致性良好，$\kappa > 0.8$ 表示高度一致。

$$\kappa = \frac{p_0 - p_e}{1 - p_e},$$

其中 p_0 為準確率，p_e 為隨機一致性，即模型將所有樣本預測為樣本數最多之類別的準確率。

伍、 研究成果

一、挑選超參數

承第肆章節第六之三小節說明，依模型預測調參集的準確度，進行階段化調參(①-③)，得準確度最高為③97.50%，最佳模型配置為學習率=0.0005、期=40、批次大小=512。接著，下小節對最佳模型進行最終模型性能測試。

表 7 調參集下，各個候選模型的表現

準確度	批次大小=1024		
超參數	學習率 =0.0001	學習率 =0.0005	學習率 =0.005
期 = 20	-	0.9688	-
期 = 40	-	②0.9738	-
期 = 60	0.9720	①0.9736	0.9474
準確度	學習率=0.0005, 期 = 40		
批次大小 =256	0.9728		
批次大小 =512	③0.9750		

二、模型表現

觀察最佳模型在本研究資料集上的表現，模型在訓練及驗證集(以下簡稱訓練集)上的所有指標皆達 98%以上，表示模型訓練的擬合程度不錯。而模型最終性能測試，可得準確率為 97.19%，雖略低於訓練集之 99.43%以及調參集之 97.50%，但若比較預測類別參考機率前 3 個類別的準確度(TOP-3)，在測試集上的 TOP-3 準確度為 98.38%，僅與訓練集(99.98%)差距約 1.6%，表示模型的泛化能力尚佳，並無明顯過擬合的問題。因為本研究資料有類別不平衡的問題，故參考 Cohen's Kappa 指標，反映模型的實際性能，得測試集的$\kappa = 0.9711 > 0.8$，表示模型的預測結果與實際標籤之間有高度一致性(表 8)。

另可觀察最佳模型在沒有參與訓練的資料集上(調參集、測試集)，宏精確度、宏召回率及宏 F1 分數皆低於 90%，究其原因為宏平均是無論樣本數量多少，每個類別的權重皆相等，故當模型對少數類別的預測表現較差時，宏平均 F1 分數會明顯降低。而模型為何對少數類別的表現較差，其可能在於資料上，第一個原因是某些科目編碼本身就非常少出現，因新北市家庭收支記帳調查支出項目細分為 808 個類別，故如娛樂用之錄音帶及錄影帶的科目編碼 85202，在 19 萬筆資料中僅有一筆，占比非常低；第二個原因則因 113 年新北市家庭收支記帳調查更新部分科目編碼，與 112 年以前的原始資料編碼稍有不同，故本研究僅蒐集 113 年調查原始資料，再加上按時間切分資料集為訓練及驗證集(1-7 月)、調參集(8 月)及測試集(9 月)，可能造成模型在受時間季節性影響較深的類別(如 9 月開學相關支出)上沒有經過一定的訓練，導致預測能力不佳。總結而言，尚需待時間蒐集資料(至少一年調查資料)，讓原始資料中的科目編碼完整，才能讓模型表現更佳。

表 8 在所有資料集下，最佳模型的表現

	1-7 月 （訓練及驗證集）	8 月 （調參集）	9 月 （測試集）
準確度	0.9943	0.9750	0.9719
TOP-3 準確度	0.9998	0.9868	0.9838
Cohen's Kappa(κ)	0.9942	0.9743	0.9711
宏精確度	0.9816	0.8533	0.8469
宏召回率	0.9816	0.8533	0.8416
宏 F1 分數	0.9809	0.8443	0.8358

三、實際應用

接著將模型實際應用於未經檢核的 10 月新北市家庭記帳調查資料，從 3 萬 3,038 筆編碼中抓出可能有誤之編碼計 1,162 筆，減少約 96.48%審核時間(假設每筆資料審核時間一樣)，有效減少編碼檢核之人力作業。

觀察模型抓出可能有誤之編碼(表 9)，第一筆資料顯示模型確實會根據解釋變數中是否有購自餐飲服務的編碼規則預測為 900 系列(購自餐飲服務之編碼)；而第二筆資料解釋變數消費品項中文字為身體用肥皂，非單寫肥皂，屬自然語言的呈現，表示模型具抓出關鍵字肥皂，並精準預測編碼為 92499 之能力；第三筆資料則可看出解釋變數購買地點為便利商店，非屬餐飲服

務之地點，因此模型正確預測編碼為 400 系列(非購自餐飲服務之食品)。

表 9 最佳模型實際應用於新北市家庭記帳調查資料審核範例

	消費品項	購自餐飲服務	購買地點	實際科目編碼	預測科目編碼	問題
1	便當	是	8.其他商店	37101	90301	購自餐飲服務之便當歸90301，非則歸37101
2	身體用肥皂	否	3.量販店	35108	92499	肥皂歸92499
3	咖啡	否	4.連鎖便利商店	90401	41110	在連鎖便利商店購買不為餐飲服務歸41110

陸、 結論與未來展望

一、結論

（一）本研究之最佳模型在測試集上有不錯的表現，準確度達 97.17%，若像推薦系統提供最有可能之 3 個編碼作檢誤，準確度提高至 98.38%。且模型的預測結果與實際標籤之間有高度一致性，並無傾向於預測某一類別。

（二）實際應用於檢誤中，模型抓出可能有誤之編碼計 1,162 筆，相較原本需檢誤筆數之 3 萬 3,038 筆，大幅減少 96.48%，確實降低編碼檢核之負擔。

（三）雖本研究之模型已確實在今年運用於實務上，並獲得良好成效。但家庭記帳調查之科目代碼經常更新編碼規則，一經更新現有模型將不合使用，需重新依新設置之科目代碼訓練，亦表示需重新收集資料，而為讓模型泛化能力佳，資料則必須收集完整，收集的時間期就需拉長，可能又會發生編碼規則更新使模型不合之情形。因此，為確保模型效果穩定，建議定期檢查資料品質並及時更新訓練詞庫，使模型能適應最新的資料特徵變化。

二、未來展望

（一）本研究資料有類別不平衡的問題，故單純依賴準確度作為指標挑選超參數，可能會有誤導的情形發生，因為模型可以通過簡單地偏向多數類別來達到高準確度，但實際性能可能較差。未來針對類別不平衡的問題，可以考慮如 F1 分數等指標來挑選超參數，得出更準確之最佳模型。

（二）本研究為降低調參時間成本，採階段化調參方式，可能僅獲次佳解模型，未來可嘗試使用網格搜索法，即在預先設定的範圍內，對每個超參數進行窮舉搜索，嘗試所有可能的組合。亦或更進一步使用自適應方法，其基於學習過程中的數據動

態調整超參數，例如學習率調整（Learning Rate Scheduler）。良好的設置參數，可有效提升模型性能。

（三）本研究之神經網絡架構選擇，為符合 CPU 運算能力，故選擇 LSTM 作主要模型。未來可往需 GPU 運算能力之模型作選擇，如 Transformer 模型中的 BERT 模型，同第參章節文獻探討所說，BERT 在自然語言處理中，有著更強大的性能，能夠更準確地分類編碼。

（四）本研究因時間限制，研究資料初步蒐集 113 年 1 至 10 月新北市家庭記帳調查原始資料，未來可持續蒐集長期資料，盡可能解決類別不平衡、季節性支出等資料問題後，重新訓練模型，使模型受更完整訓練以獲得更佳的泛化能力。

柒、 參考文獻

一、 李文德，2021，政府統計與機器學習的距離-從國際應用實例談起，主計月刊，第 781 期。

二、 趙明光，2024，機器學習於政府統計之應用，主計月刊，第 818 期。

三、 宋方捷、黃河川，2024，以 LINE 建構 AI 行職業統計編碼機器人，主計業務創新變革精進項目。

四、 中研院資訊所，CKIP Lab 中文詞知識庫小組，https://ckip.iis.sinica.edu.tw/。

五、 Meng Lee，2018，進入 NLP 世界的最佳橋樑：寫給所有人的自然語言處理與深度學習入門指南，https://leemeng.tw/shortest-path-to-the-nlp-world-a-gentle-guide-of-natural-language-processing-and-deep-learning-for-everyone.html。

六、 Meng Lee，2019，進擊的 BERT：NLP 界的巨人之力與遷移學習，https://leemeng.tw/attack_on_bert_transfer_learning_in_nlp.html。

七、 Simplilearn，AI & Machine Learning，https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn。

八、 UNECE，2021，Machine Learning for Official Statistics.

九、Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

十、全民瘋 AI 系列 [經典機器學習]，2021，交叉驗證 Cross-Validation 簡介，https://andy6804tw.github.io/crazyai-ml/25.%E4%BA%A4%E5%8F%89%E9%A9%97%E8%AD%89%20Cross-Validation%20%E7%B0%A1%E4%BB%8B/ 。