國立臺灣大學管理學院會計學研究所

碩士論文

Graduate Institute of Accounting
College of Management
National Taiwan University

Master Thesis

以文字探勘技術分析致股東報告書與績效間之關聯性

Application of Text Mining Technology on the Relationship between Report to Shareholders and Performance

王奕涵

I-Han Wang

指導教授：蔡彥卿 博士

劉心才 博士

Advisor: Yann-Ching Tsai, Ph.D.

Hsin-Tsai Liu, Ph.D

中華民國 111 年 6 月

June, 2022

doi:10.6342/NTU202201111

謝辭

時光荏苒，碩士人生在論文撰寫完成後告一段落，而在寫下謝辭前，我把這兩年的點點滴滴好好地回憶一遍，發現在這看似有些顛簸的道路上，能夠順利將它走完，真的須要感謝太多協助過我的人了。

首先無疑要感謝的是我的指導教授彥卿老師及心才老師，老師從一開始對於我有興趣的研究領域就常以專業角度來提供建議，後來在我遇到瓶頸時也總是給予幫助，才使我最終如期產出這份嘔心瀝血之作。同時，我也要感謝口試委員東吳會計雪芳老師及北大會計淑華老師，在學位考試時針對我未思考周全的層面給予提點，讓這份論文內容能更加完整。

另外也想謝謝鈺清，這兩年來不論是論文或課堂報告，每次程式上遇到什麼困難都跑去麻煩你，而你無論多忙都很樂意地幫助我，就這樣我們一起煎熬奮戰了好多個夜晚，沒有你的話實在很難想像現在我的程式能力會停滯在哪個程度。還有謝謝我的課堂好夥伴薇妮、信翔、家芸、定宜陪我走過這兩年永難忘懷的光陰，雖然我們膩在一起的時間常常開玩笑地嫌棄對方，但卻也扎扎實實陪伴彼此遨遊在會計與資料分析的領域中，真的很慶幸有你們讓我的碩士生活增添許多快樂的色彩。

其實還想感謝我身邊所有的朋友，偶爾陪我天南地北的亂聊天，或是帶我去趟無憂無慮的放鬆旅行，每次總讓我充滿電更有動力繼續前進，在此誠摯地謝謝各位，也希望你們一切都好。而在最後我想感謝我的家人，謝謝他們從小到大對我不求回報的付出，而且總是比我相信自己，家人的這份支持在很多次徬徨無助時，給了我很大的勇氣向前衝，如今我要邁向人生下一階段了，未來的我一定會更加努力，做個把自己照顧好且快樂的人。

王奕涵 謹誌於

國立臺灣大學會計學研究所

中華民國一百一十一年六月

doi:10.6342/NTU202201111

摘要

本研究以 2018 年問世之自然語言處理方法中的 BERT (Bidirectional Encoder Representation from Transformers) 模型為基礎，將其所衍生出的機器學習方法用以剖析臺灣上市櫃科技產業致股東報告書之資訊價值，並探討其文字內容與財務績效間之關聯性，看能否藉由致股東報告書內容及語調中分析出其與同產業績效平均值的高低關係及其對於企業本身未來績效成長的影響力。

研究中所使用之機器學習方法有兩種，一為將致股東報告書經前處理後轉換成 BERT CLS 向量及經 LDA 後向量，再把兩向量接起來，並以此一新向量形式放入傳統機器學習模型訓練與預測；另一為將致股東報告書經前處理後直接放入 BERT 模型中進行微調並產出預測結果。

實證結果發現，以營業收入淨額作為定義財務績效分類之依據，且與致股東報告書一同進行訓練的情況下，模型預測準確度十分理想，故推論出致股東報告書與財務績效間確實存在關聯性，且致股東報告書內容及語調可用於預測其現在或未來績效是否會高於同產業平均及與自身相比成長與否。

關鍵詞：致股東報告書、BERT、文字探勘、機器學習、績效預測

doi:10.6342/NTU202201111

ABSTRACT

This study analyzed the information value of Report to Shareholders covering all technology companies listed in Taiwan with machine learning methods primarily based on BERT, and explored the impact of text content and tone of Reports to Shareholders on a company's financial performance, including the company's financial position when compared with the average performance in the same industry and the future performance growth of the company.

In addition, two machine learning methods were utilized in the study. One of the method converted Reports to Shareholders into BERT CLS embedding and vector learned by LDA model, combined the two into a new vector form, and fed the new vector into traditional machine learning models for training and prediction; the other method generated prediction results by directly employing BERT to analyze pre-processed Reports to Shareholders and fine-tuning BERT parameters.

The empirical result showed that when using the net operating revenue to define the classification of the financial performance, the accuracy of the model results was great. It could be inferred that there is indeed a correlation between Reports to Shareholders and a company's financial performance. The content and tone of Reports to Shareholders could thus be used to predict whether a company's performance will be higher than the industry average and whether it will outgrow its present net operating revenue.

Keywords: Letter to Shareholders, BERT, Text Mining, Machine Learning, Performance Forecasting

doi:10.6342/NTU202201111

目錄

doi:10.6342/NTU202201111

doi:10.6342/NTU202201111

圖目錄

doi:10.6342/NTU202201111

表目錄

doi:10.6342/NTU202201111

第一章 緒論

第一節 研究背景及動機

因美國證券管理委員會 (United States Securities and Exchange Commission，SEC) 指出僅透過財務報表及其附註，無法充分讓年報資訊使用者就此判斷盈餘品質及用以預測公司未來財務績效 (SEC, 1987)，故規定公開發行公司須於年報中揭露 Management Discussion and Analysis (MD&A)，且後續亦發布 MD&A 之揭露指引 (SEC, 2003)，使年報得傳達來自管理階層觀點的分析資訊，達到資訊使用者可更加理解企業現在和未來整體營運狀況之目的。

臺灣雖無年報須揭露 MD&A 之規定，惟年報中致股東報告書之揭露目的卻是如出一轍，於民國 77 年 06 月 07 日所發布之「公開發行公司年報應行記載事項準則」[1]中即規定致股東報告書為臺灣公開發行公司年報編制內容應記載事項，經幾度修正後在第 8 條規範其內容應包含前一年度營業結果、本年度營業計畫概要、未來公司發展策略、受到外部競爭環境、法規環境及總體經營環境之影響。

再者，由於過去與企業攸關之資訊多以財務報表上量化之財務數字進行迴歸分析，實際以非量化之財務文字揭露做分析的研究寥寥可數，且除財報附註揭露外之文字揭露因毋須經過獨立第三方提供確信服務，其攸關性是以管理當局利益或是外部資訊使用者利益為首還有待證實，因此較難賦予文字揭露一個確切的定位。然而，近來文字探勘 (text mining) 工具快速發展，處理文字揭露資訊之技術已相當成熟，加上 2018 年提出之 BERT 自然語言模型又提升文字揭露資訊分析之準確度，文字探勘顯然是日後企業財務與營運分析必備工具之一。

因此本研究選擇臺灣公開發行上市櫃公司之致股東報告書作為文本分析之對象，欲探討是否能藉此來證明致股東報告書存在之價值，並讓資訊接收者擁有更多決策參考來源。

[1]法條來源：https://law.moj.gov.tw/LawClass/LawAll.aspx?pcode=G0400022

doi:10.6342/NTU202201111

第二節 研究目的與貢獻

現今國外已有愈來愈多財務會計領域人員使用文字探勘相關工具探討 MD&A 資訊價值，Li (2010) 就曾證明 MD&A 與企業未來財務績效呈正向顯著，然而僅有少數研究對致股東報告書進行文本分析，更遑論是以繁體中文書寫之致股東報告書，故本研究承接 Devlin et al. (2018) 所提出的 BERT (Bidirectional Encoder Representations from Transformers) 模型及 Peinelt, Nguyen and Liakata (2020) 提出於特定領域上將 LDA (Latent Dirichlet Allocation) 模型主題萃取功能添加到 BERT 模型上可提升模型訓練效果之概念，並衍生出不同的兩種機器學習模型流程，欲藉由不同的文本分析方法，證明於使用繁體中文的會計領域中，致股東報告書也存在與 MD&A 類似的資訊內涵，並可使用其來分析企業財務績效。

最終實證顯示若能正確以營業收入淨額定義財務績效之分類，模型預測分類結果皆能達到理想之準確度，驗證了致股東報告書的資訊價值。綜上所述，本研究之貢獻有兩點，第一為使用新興的 BERT 機器學習模型對繁體中文致股東報告書進行文本分析，一反過去財務會計領域多以傳統迴歸模型對財務數字或比率加以分析及探討，本研究實證顯示將機器學習模型應用於繁體中文之會計領域具可行性，因此未來可以此為分析對象，繼續從事相關研究分析，豐富其應用多元性；第二為使用企業文字揭露資訊來預測財務績效之分類，實證顯示文字揭露與一般量化財務數字相同，皆蘊含資訊價值，該結果可增加企業內部對其致股東報告書寫作內容之重視，亦可使企業外部資訊使用者（如：股票分析師、潛在投資者、股東等）將致股東報告書作為相關決策之重點參考來源。

doi:10.6342/NTU202201111

第三節 研究流程

本研究流程如圖 1-1 所示，將致股東報告書與財務績效之關聯性分為五章來探討，除本章說明研究動機與目的外，其餘章節之安排依序為：探討相關文獻、建立研究方法並介紹研究設計、顯示實證結果，最後則說明研究結論及限制。

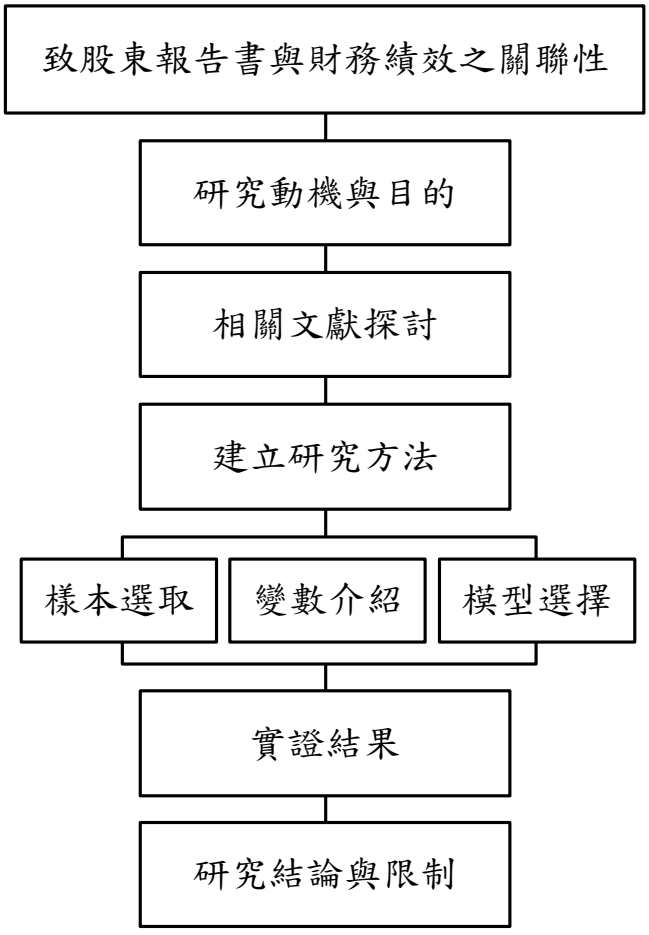

圖 1-1 研究流程圖

doi:10.6342/NTU202201111

第二章 文獻探討

本文以科技產業企業年報中致股東報告書作為研究對象，試圖分析其與現在及未來財務績效間是否具關聯性。由於研究方向會涉及到三種現有之研究議題，分別為文字揭露所隱含之資訊重要性、致股東報告書的預測價值、文字探勘技術對非財務資訊的分析方法與應用，因此於此章節會進行相關文獻敘述及探討。

第一節 文字揭露資訊重要性

許多研究顯示資訊揭露可以有效降低管理當局與外部股東間資訊不對稱及代理問題 (Glosten and Milgrom, 1985; Healy and Palepu, 2001)，而在資訊揭露的涵蓋範圍中，按 Petersen (2004) 所提出的可依照量化與非量化來區分，結構化的量化資訊被稱為硬資訊 (hard information)，例如會計科目數字、財務比率等；非結構化的非量化資訊則被稱為軟資訊 (soft information)，例如致股東報告書、管理階層的討論與分析 (MD&A)、財報附註揭露等。

有別於經會計師審計後報表的量化財務數字等硬資訊，屬於軟資訊的非量化財務資訊之文字揭露撰寫空間通常較廣，且會因企業管理者之個人特質及背景差異，造成不同企業間隨資源多寡決定其欲達成的公司目標 (Hambrick and Mason, 1984; Eisenhardt and Schoonhoven, 1996)，進而導致每間企業想傳達給資訊接收者的重點訊息不同、公開揭露資訊內容差異甚大。

再者，這些軟資訊通常由管理階層對與企業相關的利害關係人說明，此時揭露可讀性就成了一大重點。Schroeder and Gibson (1990) 就曾經以消極語態的使用、單字長度、句子長度三種層面去分析 40 家企業的 MD&A、致股東報告書及財報附註揭露的可讀性，並分析出其閱讀難易度尚有進步空間，未能達到管理階層欲與財報使用者溝通的程度。然而，上述雖顯示軟資訊編制存在相對的資訊品質風險，惟其中含有的資訊量依然高於一般硬資訊的訊息量 (Rogers and Grant, 1997)。

doi:10.6342/NTU202201111

對於軟資訊的重要性，有愈來愈多學者進行相關之研究，分析結果也印證文字揭露確實具有鑑往知來的信息量，像是企業發布之財經新聞稿中語調積極與否與市場異常收益有所關聯 (Henry, 2008)、企業 10-K 年度報告用詞也引起股票交易量的變化 (Loughran and Mcdonald, 2011)、企業社會責任報告書與盈利能力之間具正向關係 (Waddock and Graves, 1997; McGuire et al., 1988, 1990; Auperle et al., 1985) 等研究，除此之外還有多項與 MD&A 相關的研究也都說明軟資訊會影響企業未來財務狀況表現 (Clarkson et al., 1999; Cole and Jones, 2004; Bryan, 1997; Sun, 2010)，以上皆證明文字揭露資訊對評估企業營運狀況的重要性是無庸置疑。

doi:10.6342/NTU202201111

第二節 致股東報告書之預測價值

致股東報告書所涵蓋的資訊相當多樣，除簡述過去年度歷史性的營業狀況和結果外，也揭露了企業未來發展策略、外部競爭環境與總體經營環境等具前瞻性的內容，因此被認為是企業數一數二重要的軟資訊 (Arnold and Moizer, 1984)。

然而，因為致股東報告書不須經由會計師審核即可公開，很難確認其是否有虛報或誇大事實之行為，故有學者對於該文字揭露的資訊品質表示質疑，認為企業可能刻意操弄其措辭 (Courtis, 1998)。Hildebrandt and Snyder (1981) 即使用正負面詞頻的方法，以小樣本數發現企業不論當年度財務績效如何，所編製的致股東報告書正面詞語皆比負面詞語更頻繁地出現，語調皆為正面，並提出「Pollyanna Hypothesis」的說法來解釋此現象。

反觀來說，也有不少學者始終相信致股東報告書的資訊含量，以其來與企業財務狀況進行研究分析 (Smith and Taffler, 2000)。Kohut and Segars (1992) 提出績效不佳的公司傾向在致股東報告書中強調未來擁有的機會，而非過去財務表現；Abrahamson and Amir (1996) 使用致股東報告書中負面涵義的字相對於總數之指標 (RBADWD) 來證實信中負面涵義的字與未來收益有關；Frazier et al. (1984) 發現從致股東報告書和 MD&A 中，管理階層對企業現狀與未來展望的描述可區分出企業是否有破產可能性的徵兆。

另有研究從外部報告接收者的角度來看致股東報告書，Swales (1988) 發現投資人可使用致股東報告書的內容來識別市場股票「贏家」和「輸家」，並從管理階層的情緒區分高績效公司和低績效公司；Previts et al. (1994) 以金融分析師發表文件時所依據的資料來源做研究，推論出分析師在做出投資建議時會參考的資訊類型中，致股東報告書與 MD&A 占非常重要的部分。

雖然與硬資訊或 MD&A, CSR 等軟資訊相比，致股東報告書相關的研究實屬不多，惟仍可藉由上述研究看出，致股東報告書有預測企業財務績效的資訊價值。

doi:10.6342/NTU202201111

第三節 文字探勘分析方法與應用

文字探勘是資料探勘的一種，其囊括了資訊檢索、計算語言學、自然語言處理、機器學習等多種跨領域的知識，屬於非結構化及半結構化的文本分析，主要處理流程包括斷字斷詞與斷句、詞幹提取與還原、停用字刪除、特徵值萃取、分類與集群等。Sullivan (2001) 將文字探勘定義為「一種編輯、組織及分析大量文件的過程，發現資訊特徵及其間的關聯性，以提供給特定資訊使用者」。

文字探勘技術可應用於會計領域中的非結構化文本，Qiu, Srinivasan and Street (2006) 就選擇 30 家企業各 10 年，共 300 份的年報樣本作為研究對象，再使用年報對應年度的 ROE 以自行定義的方式將每一樣本分類為 positive, negative 和 neutral 後，放入支持向量機 (SVM) 進行訓練，結果證實年報使用文字探勘去預測下一年度財務績效的方法是可行的。

因此，近年來文字探勘在會計領域的應用有不少研究與討論 (Antweiler and Frank, 2004; Feldman et al., 2010; Balakrishnan, Qiu and Srinivasan, 2010)。Li (2008) 利用計算語言學中的 Fog Index 和文件字數兩項衡量指標，發現收益較低和短期收益為正的企業年報較難閱讀；後續研究 Li (2010) 按正負音調及盈利能力、運營、流動性等內容，手動對從 MD&A 中隨機提取的 30,000 個前瞻性聲明 (FLS) 的句子進行分類，並放入單純貝氏分類器進行訓練，結果發現企業 MD&A 中 FLS 的平均基調與未來收益呈正相關。

而以下會針對本研究中使用的文字探勘重要程序與機器學習分類方法加以說明：

（一）中文斷詞處理

中文能表現出意義的基本單位為詞，相較於英文可以固定方式使用空格來進行斷字，中文並沒有一定的模式，須要借助建詞典與演算法才能進行斷詞。近期較多人使用的中文斷詞系統有兩種，分別為中國百度的一名開發者所寫的結巴 (Jieba) 中文斷詞及中央科學研究院開發的 CKIP (Chinese Knowledge and Information

doi:10.6342/NTU202201111

Processing) 中文斷詞系統，兩者不僅支援繁體中文的斷詞處理，且目前皆於網路上開放原始碼供使用者研究，並讓使用者可以自行新增詞典。

本研究使用結巴中文斷詞，其處理流程是基於 Trie 字典樹找出文句中所有可能切成的組合，並將其繪製成有向無環圖 (Directed Acyclic Graph, DAG)，再以動態規劃 (Dynamic programming) 算法計算出最佳的切分組合。如果是沒有定義在詞典上的未知詞，結巴則採用基於中文成詞能力所建立的隱藏式馬可夫模型 (Hidden Markov Model, HMM) 模型加上維特比演算法 (Viterbi algorithm)，來辨識相鄰之單字是否可以組合成一個新的詞。

（二）向量空間模型－ Frequency-based vectors

向量空間模型 (Vector Space Model, VSM) 最早由 Salton et al. (1975) 所提出，後來成功應用在資訊檢索領域，其概念就是將文本內容轉以空間向量的形式表示，最常被用於計算文本間的相似度。

本研究所使用的 Frequency-based vectors 即為一種向量空間模型，亦可稱為字頻法 (Term Frequency, TF)。方法是先將整個文本集中所出現過的字詞建成一個字典庫 (dictionary)，再以字典庫中每一字詞出現在該文本中的頻率計算而成。TF 的值愈大說明字詞對此文本就可能愈重要，惟因為 TF 也可能被 Zipf's Law 所解釋的現象和該文本集的特性而影響，故此方法有其缺失存在。

（三）隱含狄利克雷分布

隱含狄利克雷分布 (Latent Dirichlet Allocation, LDA) 由 Blei et al. (2003) 所提出，為一基於機率分布的主題模型，並採非監督學習算法，僅須要指定主題數量即可將模型放入文本集資料進行訓練，不須先對文本集人工標記就能萃取出重要語義資訊，在大量文本的分析上十分有效率。

doi:10.6342/NTU202201111

LDA 的概念假設參雜兩種機率分布，一為一個文本通常涉及多個主題，故每一文本都應有其主題分布；另一為一個主題有偏好的用字傾向，主題間的字詞並非相互獨立，而每一主題都有對應的字詞分布，因此文本間的差異性來自於文本中主題的混合比例與該主題下的字詞分布機率。另外，LDA 基於「詞袋」(bag-of-words)的概念並不考慮文本中字詞的順序。

LDA 採用狄利克雷分布將研究者的先備知識納入模型估計過程，讓字詞與主題並非平均分布，因此分布考量了先備知識和觀察樣本兩種指標，為多項式分布。其概念示意圖如圖 2-1，字詞分布參數 φ 由狄利克雷超參數 (Dirichlet hyper-parameters) β 所控制，主題分布參數 θ 由狄利克雷超參數控制 α 所控制。其中，k 表示主題數量，M 表示文本數量，N 表示文本集中出現過的字詞總數，w 表示觀察到的字詞，z 則表示 w 對應的主題。

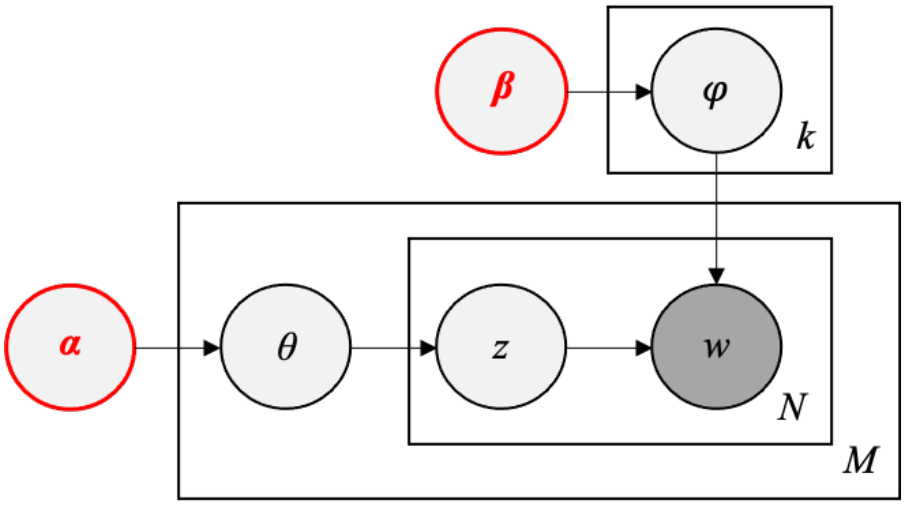

圖 2-1 LDA 概念示意圖

因此，透過 LDA 模型一篇文本生成的概念是：根據主題分布 θ 抽取主題 z，再依據該主題 z 的字詞分布 φ，從中抽出一個單詞 w，將這個過程反覆多次，即可產生一篇新的文本。

doi:10.6342/NTU202201111

（四）BERT

BERT (Bidirectional Encoder Representations from Transformers) 是 Google 在 2018 年發表以無監督的方式利用大量無標註文本預訓練而成的雙向語言模型 (Devlin et al., 2018)，為文本分析下游的任務先學習到語言學中的句法、詞義等資訊。

Google 在預訓練 BERT 時，讓模型同時進行兩項任務，分別是 Masked Language Model (MLM) 和 Next Sentence Prediction (NSP)。MLM 是以克漏字填空的概念訓練語意相關性，即把句子中的任一個字蓋住，讓模型預測可填入哪一最符合上下文語意的字；NSP 為訓練語意的接續性，給定模型上一句的句子，讓模型預測可接續哪一最符合語意的句子。

當預訓練的 BERT 模型建置好後，也可依照下游任務的需求在模型最後加上新的 layer 進行微調 (Fine-tune)，此時原本預訓練 BERT 模型的權重會被更新，讓模型更符合預測任務。

（五）羅吉斯迴歸

羅吉斯迴歸 (Logistic Regression) 是迴歸分析的一種，惟有別於一般線性迴歸的依變數為連續型變數，羅吉斯迴歸的依變數為類別變數，而因類別通常為是二元的，數值為 1 或 0，故此時可稱之為二元羅吉斯迴歸。一般而言，會將欲研究的現象設定成 1，為實驗組，反之設定成 0 則為對照組。

羅吉斯迴歸概念是將類別目標依變數的發生機率經過 log 函數先轉換成 log odds ($\frac{p}{1-p}$)，才對其與自變數間的關係進行線性預測，找到線性預測的那條線或平面後，就可輕易將分類為 1 和 0 的資料區隔開來，用於未來進行分類。另外，羅吉斯迴歸對於各項自變數的參數估計是使用「最大概似估計法 (Maximum Likelihood Estimation, MLE)」。

doi:10.6342/NTU202201111

（六）支持向量機

支持向量器 (Support Vector Machine，SVM) 最早在 1963 年被 Vladimir Naumovich Vapnik 和 Alexey Yakovlevich Chervonenkis 提出，是一種監督式的機器學習方法，可用於線性及非線性分類。其基本概念為找到一個決策邊界 (decision boundary)，讓邊界兩邊離樣本的距離 (margins) 最大化，使其可以完美對資料做有效的切割，這樣當有新的樣本要做分類時，分類成功的機率會較高。而其中離決策邊界最近的點被稱為支持向量 (support vector)。

圖 2-2 以樣本分布來解釋樣本可線性分類與不可線性分類。面對較複雜的樣本導致線性不可分類時，就要依靠非線性分類方法，非線性運作原理是將低微度的資料樣本經過轉換投映到高維度空間，在較高維度空間用同樣的方式去成功找尋能對各類別間樣本做有效切割的超平面。

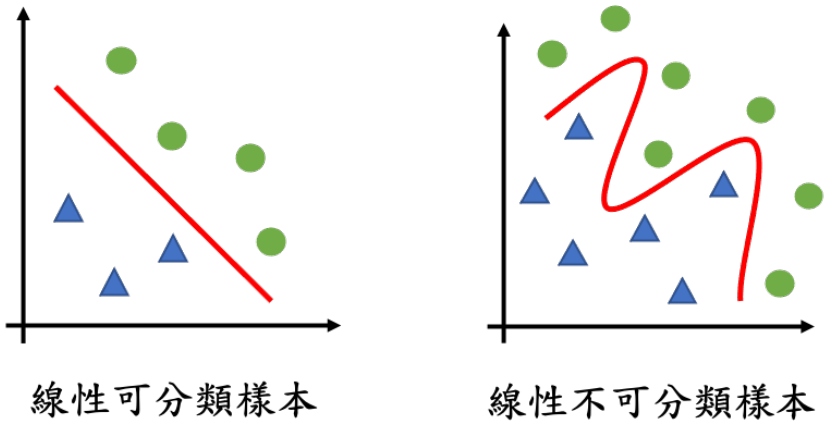

圖 2-2 線性可分類與不可分類樣本示意圖

（七）隨機森林

隨機森林 (Random Forest) 由 Leo Breiman (2001) 所提出，由很多棵決策樹所組成，為進階版的決策樹。決策樹是依據訓練資料所產生的樹，建立時為能使訊息增益最大化，會使用亂度來評估每一分枝下用以切分資料的變數，而最終找出所有合適的規則後，就可生出一棵決策樹來進行分類決策。

doi:10.6342/NTU202201111

隨機森林簡單來說就是多棵決策樹的集合 (ensemble)，其中每棵決策樹所挑選的訓練樣本都是採取後放回的隨機抽樣方式，加上變數的選取也是隨機的，故每棵決策樹用於訓練的樣本和變數不同，各自產生的規則不一樣，生成的決策樹也就具多元性。於最後預測時，隨機森林就將樣本放進每棵規則不同的決策樹並得出結果，再將所有結果綜合起來作為最終分類結果。

此方法解決單棵決策樹因為要讓模型效能變高而完美適應訓練資料，卻無法預測未知資料的過度擬合 (over-fitting) 情況，進而也提升模型分類預測的效能。

doi:10.6342/NTU202201111

第三章 研究方法

第一節 研究架構

本研究以科技產業的上市櫃公司作為研究對象，並將樣本分成科技產業、半導體業和電子零組件業的致股東報告書，以進行其與當年度 (t) 和次年度 (t+1) 財務績效關聯性之深度分析。

首先，取得致股東報告書樣本與財務績效資料後，會經過資料前處理將股東報告書文字內容轉換為向量形式，及將財務績效資料以自行定義的衡量方式轉換為兩元分類，接著依照 90/10 法則將整個樣本資料集切分成訓練資料集與測試資料集，分別放入不同的模型訓練流程中後，依照各模型的財務績效預測表現來得出最後的結論。研究架構圖如圖 3-1。

其中傳統機器學習模型中為避免拆分訓練與測試資料集有任何資料上的偏誤，故採用 10-fold 交叉驗證。概念是將整個資料集先分成十等份，每一等份的資料都會輪流用於測試，當其中一份當測試資料時，其餘九份就是訓練資料，此方法下每一模型都會產生出十次的預測結果，而各模型最終的表現是以平均值來做比較。

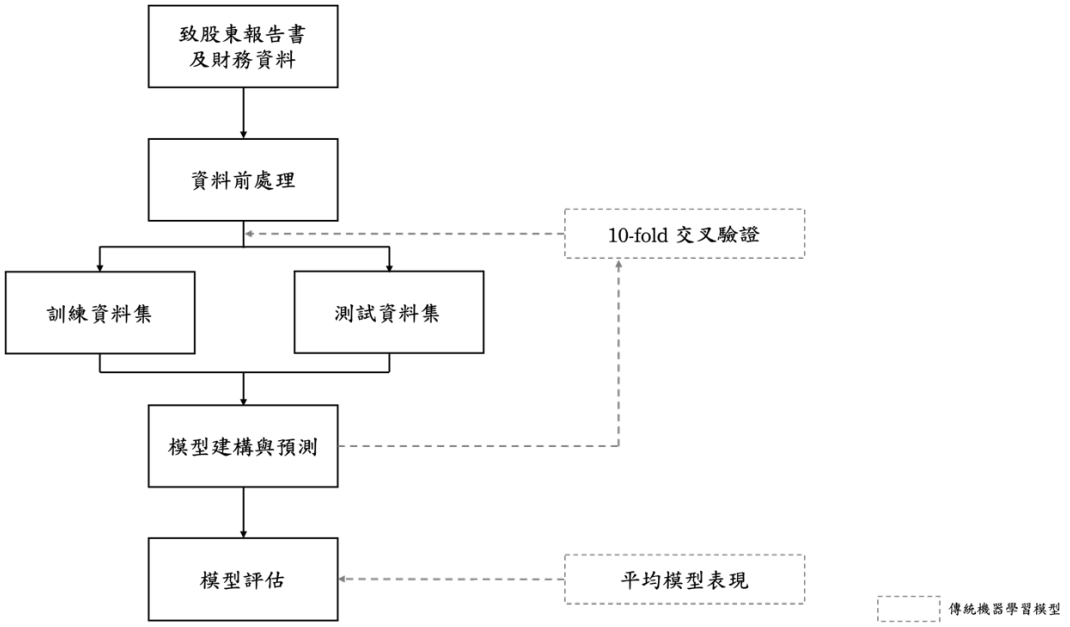

圖 3-1 研究架構圖

doi:10.6342/NTU202201111

第二節 樣本選取

本研究以 2022 年 1 月 15 日為基準，選取該日已於臺灣上市、上櫃與興櫃之科技產業公司來進行致股東報告書與當年度 (t) 或次年度 (t+1) 財務績效間的分析，並因為科技產業於臺灣已發展多年，穩定且持續精進技術、增加產值，故為讓本研究的樣本數量愈多以增加深度學習的效果，設定分析的時間區間為民國 100 年至 109 年度，共十年。

致股東報告書的資料來源為公開資訊觀測站，從該網站下載各公司民國 100 年至 109 年度之 10 年的股東會年報（年報會公告於下一年度），再擷取每個年度報告中第一部分的致股東報告書。

衡量財務績效所需要的財務數字資料來源則為台灣經濟新報資料庫系統 (TEJ+)，雖欲分析的是當年或次年度的財務績效，惟目前民國 110 年資料無法取得，因此僅取民國 100 年至 109 年度共 10 年的資料。而因財務績效可以不同的層面分析，本研究選取變數之概念是以投資者角度出發，且認為投資者可能較重視財務績效中的獲利能力，故選取能判斷獲利能力之相關數字，包含營業收入淨額、報酬率、稅前淨利、股東權益、每股盈餘、稀釋每股盈餘、調整因子及年底收盤價，其中報酬率計算方式是以投資報酬率 (ROI)，即投資淨損益除以總投入資金，為依據。變數皆為年資料，且後續會藉不同組合及運算方法來定義最終財務績效分類。

本研究分析了三種產業別，分別為科技產業、半導體業、電子零組件業，其中半導體業與電子零組件業皆為公開資訊觀測站所分類的產業別之一，而科技產業則為本研究自行定義之產業，將於分類為電子零組件業、電腦及週邊設備、半導體業、通信網路業、光電業、電子通路業、資訊服務業、電子工業和其他電子業的此九個產業別都視為廣義的科技產業，各自的企業數量如表 3-1 所示。

其中，各產業別選取樣本的方法一致，符合上市櫃標準且分類為科技產業、半導體業、電子零組件業之企業分別有 1,259、190 與 224 家，應有 12,590、1,900 與 2,240 筆樣本資料，扣除企業當年度本身沒有公開年度報告和技術性問題而無

doi:10.6342/NTU202201111

法成功獲取致股東報告書之內容者，再扣除企業當年度無法取得所有財務數字而無法計算財務績效者，最終剩餘有效樣本各有 5,629、833 與 1,372 筆。詳細的資料彙整如表 3-2 所示。

表 3-1 科技產業各細項產業別之企業數量

產業別	家數
電子零組件業	224
電腦及週邊設備	119
半導體業	190
通信網路業	97
光電業	148
電子通路業	37
資訊服務業	51
電子工業	292
其他電子業	101

表 3-2 樣本篩選流程彙整

	科技產業	半導體業	電子零組件業
符合臺灣上市、上櫃與興櫃之特定產業樣本數	12,590	1,900	2,240
扣除企業本身無公開年報	(2,884)	(376)	(198)
扣除企業年報為圖片檔或無法複製	(3,226)	(547)	(590)
致股東報告書有效樣本數	6,480	977	1,452
扣除企業本身財務數字無法取得	(851)	(144)	(80)
剩餘有效樣本數	5,629	833	1,372

doi:10.6342/NTU202201111

第三節 變數定義

致股東報告書為研究分析對象，其內含資訊屬非結構化的文字內容，因此在轉換為向量形式前，毋須對其先行定義衡量的方式。而預測標的為財務績效，因研究欲分析致股東報告書與財務績效間之關聯性有兩種，分別是能否從企業致股東報告書看出財務績效中獲利能力「高於同產業平均與否」及「與自身相比是否成長」，故以兩個不同的角度來定義及選擇所需之財務數字。另外在分析過程中，致股東報告書又分為與當年度 (t) 財務績效及與次年度 (t+1) 財務績效間之探討，故最終共使用了一種財務數字來衡量當年度 (t) 及六種不同的財務數字來衡量次年度 (t+1) 企業財務績效分類，各項財務績效分類於自行定義後皆僅用了二分法表示，預測為二元分類問題。以下會各自說明財務績效分類的衡量定義方式。

首先，先說明兩種角度下所選擇之次年度六種財務績效。「高於同產業平均與否」角度下有營業收入淨額與同產業中位數之相比，定義方式為先計算各產業下各年度營業收入淨額的中位數，再將企業次年度營業收入淨額與次年度同產業的中位數相比，如果次年度營業收入淨額高於中位數，則認為企業未來財務績效高於同產業平均，定義為 1，反之為 0。表 3-3 彙整了各產業別下每年度的營業收入淨額中位數。而「與自身相比是否成長」角度下則有營業收入淨額與同產業中位數之差距比例、報酬率、稅前淨利除以股東權益、每股盈餘除以年底收盤價、稀釋每股盈餘除以年底收盤價等六種績效分類，其中每股盈餘和稀釋每股盈餘皆有依照調整因子轉換為以符合民國 109 年為基準之財務數字，定義方式皆為同企業間該財務數字當年度與次年度的比較，如果次年度財務數字大於當年度，則認為企業未來財務績效成長，定義為 1，反之為 0。

然而，因為次年度財務績效須要拿當年度與次年度相比，加上 110 年資料無法取得，故 109 年致股東報告書無法進行訓練，資料訓練樣本數因此減少，科技產業、半導體業、電子零組件業之最終訓練樣本數分別僅有 4,687、691 與 1,175 筆。未來績效用於訓練的最終樣本數彙整如表 3-4 所示。

doi:10.6342/NTU202201111

至於當年度財務績效僅以「高於同產業平均與否」的角度使用營業收入淨額來與同產業中位數相比，定義方式為先計算各產業下各年度營業收入淨額的中位數，再將企業當年度營業收入淨額與當年度同產業的中位數相比，如果當年度營業收入淨額高於中位數，則認為企業財務績效高於同產業平均，定義為 1，反之為 0。再者，致股東報告書與當年績效之關聯性訓練樣本數不變。

表 3-3 各年度營業收入淨額中位數

（單位：新台幣仟元）	科技產業	半導體業	電子零組件業
100	2,458,824	2,329,664	2,842,277
101	2,423,557	2,772,462	2,436,453
102	2,556,807	3,035,778	2,975,194
103	2,546,716	2,693,040	2,728,475
104	2,332,165	2,262,919	2,311,901
105	2,358,363	2,161,784	2,553,081
106	2,330,824	2,397,581	2,856,395
107	2,446,252	2,048,722	2,820,470
108	2,220,677	2,250,506	2,669,131
109	2,306,140	1,926,113	2,883,247

表 3-4 未來績效用於訓練的樣本彙整

	科技產業	半導體業	電子零組件業
有效樣本數	5,629	833	1,372
扣除因 110 年資料無法取得而 109 年致股東報告書無法進行訓練	(942)	(142)	(197)
可用於訓練有效樣本數	4,687	691	1,175

doi:10.6342/NTU202201111

第四節 模型選擇

本研究使用兩種完全不同的模型訓練流程，一為將致股東報告書轉換成 BERT CLS 接上 LDA 的向量形式，再分別放入四種機器學習模型中訓練，最後得出分類預測結果；另一為將致股東報告書放入 BERT 模型，並輔以 Fine-Tuning 進行訓練，直接得出分類預測結果。以下會闡述兩種流程，而轉換為向量形式表達前的資料前處理詳細過程會於下節說明。

首先，因 Peinelt, Nguyen and Liakata (2020) 證明於特定領域上，將 LDA 主題式的萃取概念添加到 BERT 模型上可以在訓練時提高一系列語義相似性，進而提升預測的效能，故本研究承接 Peinelt, Nguyen and Liakata (2020) 提出的概念，並經過一些調整後得出一種新的訓練方式，期待能於會計領域上也對上下文的學習能力獲得相同的加乘效果。

上述說明的想法即應用在第一種方式，為將 BERT CLS 接上 LDA 的向量形式放入三加一種不同的機器學習模型中，其中機器學習模型包含羅吉斯迴歸、支持向量機、隨機森林和混合投票制，建立好模型後就可得出最終的預測結果，流程圖如圖 3-2 所示。

此處因致股東報告書文本主要內容為繁體中文，故使用繁體中文 BERT-BASE 預訓練模型，利用以大量文本預訓練的語言模型，計算出每個文本經 BERT 模型後所得出的 CLS Embeddings 作為其代表向量。另外，LDA 的部分則是透過將致股東報告書先以 TF 的向量形式呈現，再經過 LDA 的主題萃取後得出最終 LDA 向量，此方法下也讓原本以 TF 計算的向量達到降維效果，減少後續運算時間與資源。

而用於分類預測的四種機器學習模型中，前三種是既有模型，最後一種為本研究自行定義的混合投票制，其概念為將資料分別放入羅吉斯迴歸、支持向量機和隨機森林模型後，將三種模型各自產出的預測結果進行投票，因分類結果只有兩種，故哪一類獲得大於等於 2 的票數，就為混合投票制下的預測結果。

doi:10.6342/NTU202201111

最後，對模型除使用 10-fold 交叉驗證以降低資料拆分所導致的偏誤外，訓練的超參數調整有兩處，一為設定每一種模型的 random_state 皆固定，讓每次模型訓練的過程一樣，以利後續進行模型評估和分析；另一為在隨機森林模型特別使用 1000 棵決策樹，以期許能降低損失 (loss) 提高效能，其餘超參數皆使用套件本身預設。

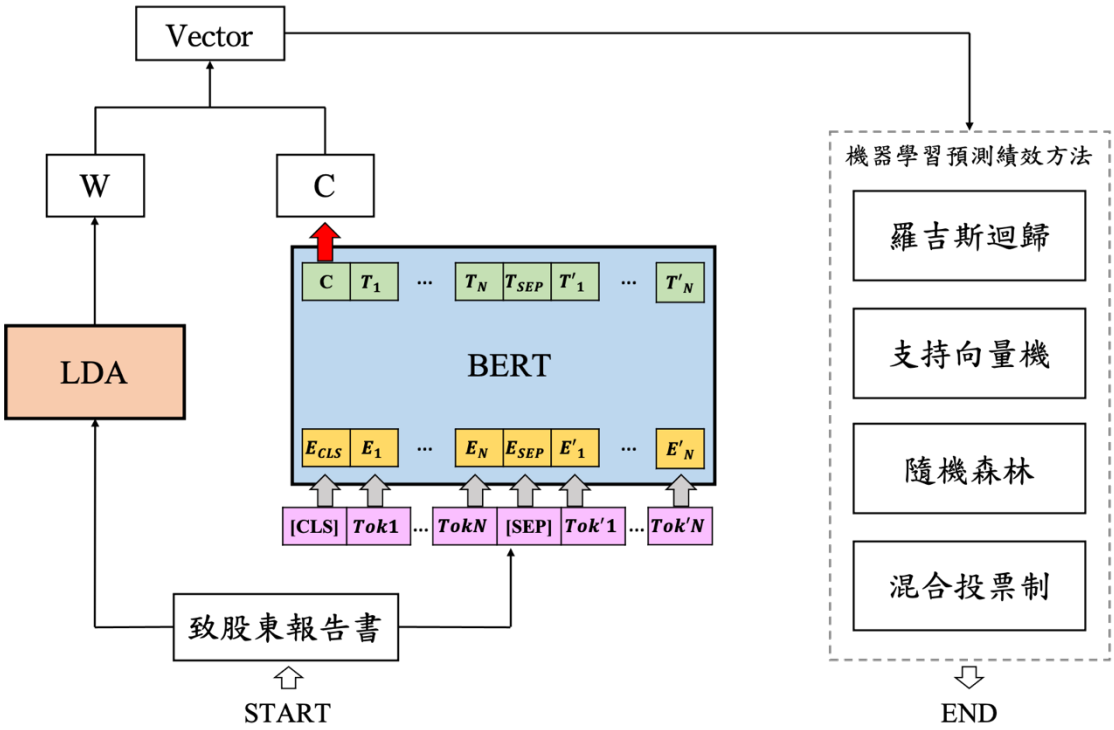

圖 3-2 BERT+LDA 模型流程圖

doi:10.6342/NTU202201111

再來，第二種方式是直接放入 BERT Fine-Tuning 的模型進行預測，流程圖如圖 3-3 所示。因文本主要內容為繁體中文，故此方法使用的也是繁體中文 BERT-BASE 預訓練模型，利用以大量文字文本預訓練的語言模型為基礎，再另外接上符合下游分類任務之全連接神經網路 (Fully-connect Neural Network, FCN) 層進行訓練，就可得出最終分類預測結果。

此處使用一層全連接神經網路層，訓練的 Batch Size 定為 8，以每 8 筆樣本資料進行一次訓練； Learning Rate 定為 10^{-6}，讓每次訓練可做些微而非大幅度的調整； Epochs 則定為 10，讓最終模型是看過全部樣本 10 次才建立出來，期許模型能達到預期學習效果且不會發生過度擬合的情況。

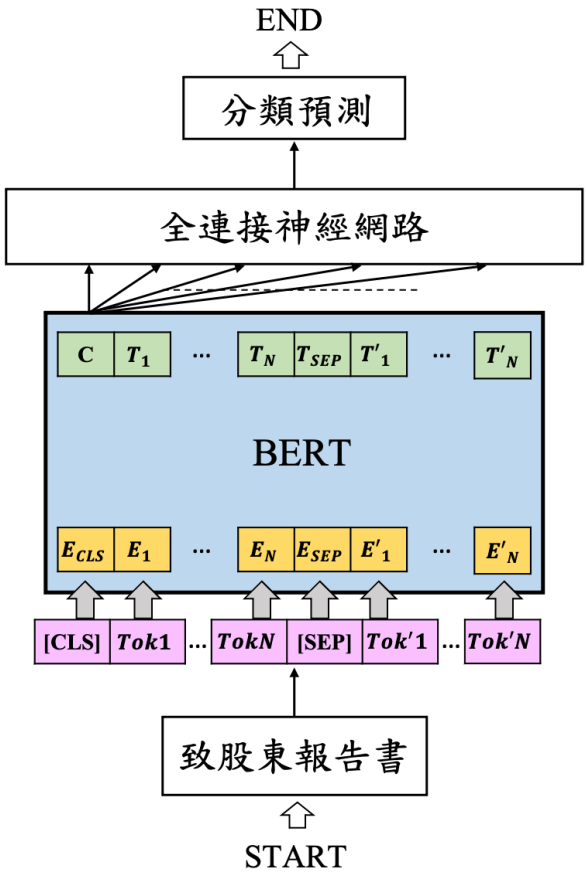

圖 3-3 BERT Fine-Tuning 模型流程圖

doi:10.6342/NTU202201111

第五節 資料前處理

在使用模型訓練致股東報告書前，為符合可放入模型的資料型態，須先對其文字原始內容進行前處理，而前處理的過程，本研究使用了刪除停用字詞、中文斷詞、TF Vectorizer、LDA、BERT CLS 等多種程序，此節會對各程序加以說明。

（一）刪除停用字詞

一般統稱經常出現卻無助於文本分析的字為停用字，並會從原文本中刪除，不讓其特性影響分析的過程。於本研究中，為能使後續以致股東報告書的文字內容進行有效分析，故初步觀察各家致股東報告書後所發現之用字傾向、搜集資料時文檔轉換出現的特殊字型，及可能與過去盈餘表現相關的數字後，將自行定義之欲刪除停用字從文本中排除，停用字統整如表 3-5 所示。其種類大致可分為：(1) 全形和半形標點符號、(2) 特殊字型、(3) 數字和國字數字

表 3-5 停用字一覽表

全形標點符號	【	】	「	」	（	）	〔	〕	〖	〗	『	』
	。	，	、	：	；	～	！	@	#	%	＿	&
	＝	＋	－	＊	／	＼	__	？	〈	〉	《	》
半形標點符號	[	]	{	}	<	>	(	)	‘	’	.	,
	:	;	`	~	!	@	#	%	%	^	&	_
	=	+	-	*	/	\	_	?				
特殊字型	★	◎	\n	\t	元	cid	\\uf06a		致股東報告書			
數字	0	1	2	3	4	5	6	7	8	9		
國字數字	壹	貳	參	肆	伍	陸	柒	捌	玖	拾	佰	仟
	萬	億	兆	零	一	二	三	四	五	六	七	八
	九	十	○									

doi:10.6342/NTU202201111

（二）中文斷詞

致股東報告書主要內容以中文闡述為主，如何對中文語句進行斷字是至關重要的步驟。於此步驟下，先不考慮文中可能出現之英文語句的斷字處理，使用 Github 上開源之結巴 (Jieba) 中文斷詞工具，其所建立的結巴字典都儲存於名為「dict.txt.big」的文字檔 (txt) 中，不自行於字典中新增字詞，並讓此斷詞工具依照字典與處理概念作判斷，最終將各致股東報告書中的所有中文語句一一切成獨立的中文字詞。

（三）TF Vectorizer

因放入 TF Vectorizer 模型的資料格式為有語句的文章形式，故要先將每篇致股東報告書中文斷字的結果以空白鍵為區隔連結起來，才能放入既有的套件模型中，以利計算出每個文章以 Term Frequency 表示的向量。其中，本研究也透過此套件的功能，將套件中預先定義的英文停用字刪除，納入中文斷字步驟未考慮的部分，使停用字的刪除可以更加全面，其餘模型的訓練參數 (min_df, max_df) 經過多次嘗試不同組合時，發現對最終預測效果沒有太大差異，故最終對其餘訓練參數並無變動，皆使用預設值。

而經過 TF Vectorizer 的運算過後，科技產業、半導體業、電子零組件業中致股東報告書所出現過文字所產出的字典庫詞庫數量，也就是以向量形式表示的維度大小，當年度分別有 92,810、25,367、37,214，次年度則分別為 82,220、21,683、33,650。

（四）LDA

LDA 的計算方法要以 TF 為基礎，故將以 TF 表示的向量放入模型進行計算，此步驟重要的決定在於要設定多少的「主題」數量，也就是 LDA 後向量的維度大小。而於本研究下，經過多次不同主題數量的嘗試，最終選擇設為

doi:10.6342/NTU202201111

768，不僅讓原本的 TF 向量得到有效的降維效果，其中最大原因為 BERT BASE 模型最終產出的維度也為 768 ，研究期待最終放入訓練分類模型中的向量，BERT CLS 與 LDA 的影響力一致，可以為此綜合向量的表示分別帶來上下文語意與主題萃取兩種不同意義的貢獻。

另外，為了研究的一致性及便利性，將 LDA 模型訓練的 random_state 設為固定，使每次經過 LDA 模型產出向量的過程一致，而其餘模型訓練參數並無變動，皆使用預設值。

（五）BERT CLS

BERT Embeddings 指的是文章經過 BERT 預訓練模型提出特徵後，所轉換出的向量，而 CLS 則為在文章最一開始所新增空白 token 的 Embedding，即為 Embeddings 中的第一個 Embedding ，且本研究下所指的 BERT Embeddings 並未經過 Fine-tune 過程。

本研究因分析標的致股東報告書文本主要內容為繁體中文，故使用 Google 於 2018 年所提出的 BERT-Base, Chinese 預訓練模型，且其預訓練模型為公開資源，已將模型所需的資料儲存於名為「chinese_L-12_H-768_A-12」的資料夾中，只須於 Github 下載即可使用。

再者，之所以僅使用 CLS Embedding 代表 BERT 訓練後的文章向量，原因有二：一為如使用所有 Embeddings 來進行後續訓練，則可能使用大量時間與資源；另一為在預訓練階段的 NSP 任務中，CLS 會被拿來當句子是否相接的預測，故即便 CLS Token 本身不帶有任何具體意義，且未進行 Fine-tune，本研究也相信 CLS Embedding 就已涵蓋了基於下一句預測的語意表達，可以客觀地表示整個致股東報告書的意思。

doi:10.6342/NTU202201111

上述介紹完五種於本研究所使用到的資料前處理方式，然而進行兩種不同的模型訓練流程前，所使用的資料前處理方式不同，因此以下來說明此兩種模型訓練流程所進行的前處理步驟。

第一種將 BERT CLS 接上 LDA 的向量放入不同機器學習模型的方法中，其中 BERT CLS 和 LDA 分別使用不同的前處理。要取得 BERT CLS 前原始的致股東報告書須經過刪除停用字詞及 BERT 未經 Fine-tune 的預訓練模型轉換方能成功得到 CLS Embedding；LDA 則為經過刪除停用字詞、中文斷字、TF Vectorizer 和 LDA 模型轉換才可取得。

分別獲得 CLS Embedding 和 LDA 後，最終將此二向量接在一起，會變成 768 維加 768 維，共 1,536 維的新向量。BERT CLS 接上 LDA 向量之資料前處理步驟如圖 3-4 所示。

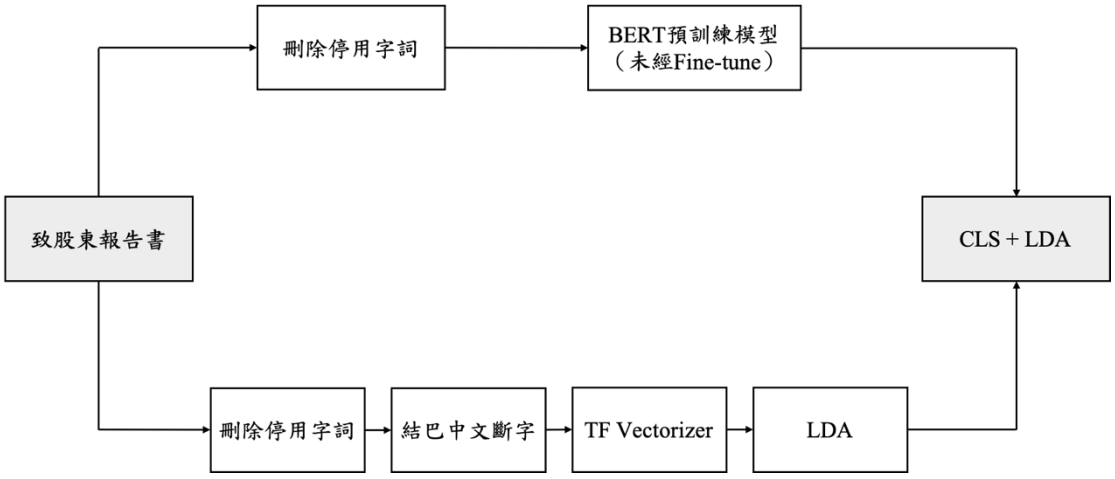

圖 3-4 BERT CLS 接上 LDA 向量之資料前處理步驟

第二種是將文章直接放入 BERT Fine-Tuning 模型的方法，此方法因 BERT Fine-Tuning 模型就是最終的分類預測模型，故僅須要將原始的致股東報告書先經過刪除停用字詞，即可直接放入 BERT Fine-Tuning 模型中進行訓練。BERT Fine-Tuning 模型所需的資料前處理步驟如圖 3-5 所示。

doi:10.6342/NTU202201111

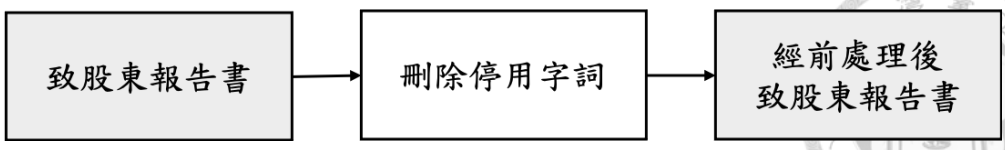

圖 3-5 BERT Fine-Tuning 模型所需之資料前處理步驟

致股東報告書經過資料前處理取得向量 (X 值) 後，即可與對應的財務績效 (y 值) 一同放入訓練模型中進行訓練，而因本研究欲分析致股東報告書與財務績效間的關聯性，且預期致股東報告書能有效捕捉當年度績效分類及預測次年度績效分類，故本研究期待模型訓練效果愈佳愈好，即愈足以支持本研究之預期假設。

doi:10.6342/NTU202201111

第四章 實驗結果

第一節 敘述性統計

由於致股東報告書的長度可能影響資訊量蘊含的多寡，進而導致不同的分類效果，故為能將此因素列入考量，首先先對各產業別下致股東報告書的平均字數進行比較，如表 4-1 所示，並發現總體來說平均字數大約落在 2,000 字上下，而其中電子零組件業是最多的，科技產業次之，半導體業則是最少。

表 4-1 致股東報告書平均字數

	科技產業	半導體業	電子零組件業
當年	2069.89	1845.30	2135.68
次年	2054.25	1791.03	2125.43

再者，表 4-2 為統計各指標下樣本類別之數量與比例，用以觀察是否存在「樣本不平衡」的問題。「樣本不平衡」指資料集中各類別的樣本數量極端不均衡，於二元分類問題之下，代表分類為 1 與分類為 0 的樣本數量比例過於懸殊，且如果此狀況存在，則必須於後續訓練前先對資料集進行處理，否則將導致分類模型為符合樣本特性而有嚴重的偏向性。

從表 4-2 可看出，不論各產業別樣本數量多寡，其各指標下之類別比例普遍平均。「高於同產業平均與否」中當年度、次年度營業收入淨額與同產業中位數之相比，及「與自身相比是否成長」中次年度營業收入淨額與同產業中位數之差距比例（科技產業、電子零組件業）、報酬率、稅前淨利除上股東權益之樣本類別比例差距皆於 5% 之內，可謂幾乎零差距；至於「與自身相比是否成長」中次年度營業收入淨額與同產業中位數之差距比例（半導體業）、每股盈餘除上收盤價及稀釋每股盈餘除上收盤價雖然樣本類別比例差距稍大，為 5 至 10% 之間，惟仍落於

doi:10.6342/NTU202201111

本研究的接受範圍內，且預期此差距不至於影響模型分類結果，故最終選擇以此資料集進行後續模型訓練，毋須另外擴增資料樣本或重新採樣。

表 4-2 各指標下樣本類別數量與比例

			科技產業		半導體業		電子零組件業	
			1	0	1	0	1	0
當年	高於同產業平均與否	營業收入淨額	2,822 (50.13%)	2,807 (49.87%)	417 (50.06%)	416 (49.94%)	684 (49.85%)	688 (50.15%)
次年	高於同產業平均與否	營業收入淨額	2,400 (51.21%)	2,287 (48.79%)	351 (50.80%)	340 (49.20%)	600 (51.06%)	575 (48.94%)
	與自身相比是否成長	營業收入淨額	2,460 (52.49%)	2,227 (47.51%)	406 (58.76%)	285 (41.24%)	643 (54.72%)	532 (45.28%)
		報酬率	2,554 (54.49%)	2,133 (45.51%)	377 (54.56%)	314 (45.44%)	645 (54.89%)	530 (45.11%)
		稅前淨利/股東權益	2,273 (48.50%)	2,414 (51.50%)	342 (49.49%)	349 (50.51%)	571 (48.60%)	604 (51.40%)
		每股盈餘/收盤價	2,022 (43.14%)	2,665 (56.86%)	309 (44.72%)	382 (55.28%)	507 (43.15%)	668 (56.85%)
		稀釋每股盈餘/收盤價	2,023 (43.16%)	2,664 (56.84%)	310 (44.86%)	381 (55.14%)	508 (43.23%)	667 (56.77%)

doi:10.6342/NTU202201111

第二節 BERT CLS + LDA 放入機器學習模型結果

本節先介紹本研究所使用於評估模型之衡量指標，接著顯示將各產業別產出之 BERT CLS 接上 LDA 向量後放入不同機器學習模型訓練，且經過 10-Fold 交叉驗證的模型分類表現，及圖示化後的分類結果。

首先，表 4-3 為混淆矩陣 (confusion matrix)。TP 表示模型預測分類與實際分類一致為 1；FP 表示模型預測分類為 1 但實際分類為 0，也稱為型一錯誤；FN 表示模型預測分類為 0 但實際分類為 1，也稱為型二錯誤；TN 表示模型預測分類與實際分類一致為 0。

表 4-3 混淆矩陣

	實際為 1	實際為 0
預測為 1	TP (**T**rue **P**ositive)	FP (**F**alse **P**ositive)
預測為 0	FN (**F**alse **N**egative)	TN (**T**rue **N**egative)

於本研究下，評估模型預測表現會以上述混淆矩陣為基礎，使用其衍生的不同的公式計算出多種衡量指標，包括：Precision、Recall、F1-Score、Accuracy 和 AUC。Precision 代表模型預測分類為 1 的樣本中預測正確的比例，其公式為 $\frac{TP}{TP+FP}$；Recall 代表實際分類為 1 的樣本中預測正確的比例，其公式為 $\frac{TP}{TP+FN}$；F1-Score 代表綜合考量 Precision 與 Recall 後的調和平均數，其公式為 $\frac{2}{\frac{1}{\text{Precision}}+\frac{1}{\text{Recall}}}$；Accuracy 代表整個樣本中預測正確的比例，其公式為 $\frac{TP+TN}{TP+FP+TN+FN}$。另外，ROC 曲線用於比較各種決策門檻下 TP rate ($\frac{TP}{TP+FN}$) 與 FP rate ($\frac{FP}{FP+TN}$) 間的變化，圖形中縱軸為 TP rate、橫軸為 FP rate，而此 ROC 曲線下所覆蓋的面積稱作 AUC，其值位於 0 至 1 之間，用於判別模型的鑑別力，值愈高表示模型效能愈好。

doi:10.6342/NTU202201111

（一）科技產業

將屬於科技產業的致股東報告書代表向量與其對應之當年度績效分類放入四個 10-fold 交叉驗證模型訓練後，即可探討致股東報告書與當年度財務績效間之關聯性，且因為是對當年度績效的深度分析，績效已為過去之商業行為結果，故本研究著重的地方在於能否由致股東報告書用詞看出當年度財務績效與同產業平均間之關係，進而判斷企業是否將財務狀況如實透露於報告書的字裡行間，使企業之利益關係者能更了解企業現況。此情況下對於績效分類沒有不平等的期待，因此以下會以模型平均 Accuracy 來辨別致股東報告書與當年度績效分析之最合適模型。

科技產業分析致股東報告書與當年度績效關聯性之實證結果如表 4-4 。從表中可發現四種機器學習模型訓練效果良好，不論何種衡量指標的值皆高於 0.70，顯示科技產業致股東報告書在以「高於同產業平均與否」的角度分析下，其用詞確實與當年度財務績效相關，可從中判別當年度企業營業收入淨額是否優於同業。

再者，四種機器學習模型中又以混合投票制之分類平均 Accuracy 最高，為 0.7234，加上其餘衡量指標並無怪異之處（如：透露模型有偏頗性等），故綜合考量後推論出此分析使用混合投票制進行模型訓練是最合適的，圖 4-1 即為此模型其中一次訓練之各年度分類結果圖。

表 4-4 科技產業分析當年度績效模型效果（機器學習模型）

訓練 n = 5,067 / 測試 n = 562					
	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（產業平均）				
羅吉斯迴歸	0.7103	0.7023	0.7061	0.7070	0.7068
支持向量機	0.7114	0.7019	0.7062	0.7076	0.7075
隨機森林	0.7109	0.7456	0.7275	0.7206	0.7204
混合投票制	0.7251	0.7224	0.7236	0.7234	0.7232

doi:10.6342/NTU202201111

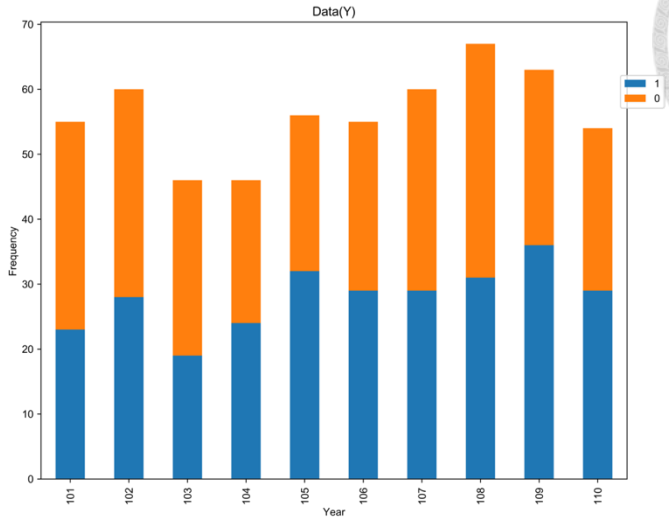

圖 4-1 科技產業混合投票制下當年度營業收入淨額（產業平均）分類圖

接著將屬於科技產業的致股東報告書代表向量與其對應之六種次年度績效分類分別放入四個 10-fold 交叉驗證模型訓練後，即可探討致股東報告書與次年度財務績效間之關聯性。次年度績效之深度分析與當年度不同的是，績效會是未來之商業行為結果，故本研究著重的地方在於能否從致股東報告書用詞預測出次年度財務績效與同產業平均間之關係及與自身相比是否成長，進而判斷企業於報告書的字裡行間是否透露未來資訊，讓企業之利益關係者對企業的投資或合作決策能更精準。此情況下對於績效分類雖然沒有存在完全不平等的期待，惟仍然希望實際為 1 的樣本可以有效被預測出，因此以下除了以模型平均 Accuracy 來辨別致股東報告書與次年度績效分析之最合適模型外，也會使用平均 Recall 來輔助挑選。

科技產業分析致股東報告書與次年度績效關聯性之實證結果如表 4-5。從表中可發現六種績效的預測標的中，僅有以「高於同產業平均與否」角度下定義之營業收入淨額能使四種機器學習模型訓練效果良好，不論何種衡量指標的值皆高於 0.70；而以「與自身相比是否成長」角度下定義之營業收入淨額表現亦落於本研究尚可接受的範圍內，不論何種衡量指標的值四捨五入後皆高於 0.55；其餘以「與自身相比是否成長」角度分析下之報酬率、稅前淨利除以股東權益、每股盈餘除以

doi:10.6342/NTU202201111

年底收盤價、稀釋每股盈餘除以年底收盤價的訓練效果都非常不理想，平均 Accuracy 都位於 0.5 上下，與隨機猜測的結果相似。然而，值得一提的是，使用報酬率作為預測標的的隨機森林，平均 Recall 達到 0.78，表示其可有效預測出未來分類實際為 1 的樣本，有利使用於相關的研究中；相較之下，每股盈餘除以年底收盤價和稀釋每股盈餘除以年底收盤價作為預測標的的隨機森林，平均 Recall 幾乎為 0，表示未來分類實際為 1 被預測出的可能微乎其微。

以上解釋可推論出科技產業致股東報告書用詞並不與每一種次年度財務績效的預測標的相關，如欲進行分析僅能從企業營業收入淨額分析，以得知是否可從其中預測次年度企業營業收入淨額是否優於同業及與自身相比是否成長。另外，兩種角度下定義之營業收入淨額之四種模型又皆以隨機森林之分類平均 Accuracy 最高，分別為 0.7250 及 0.5688，且平均 Recall 也最高，分別為 0.7699 及 0.6858，其餘衡量指標無怪異之處，故綜合考量後推論出此二分析角度下，使用隨機森林進行模型訓練皆是最合適的。圖 4-2 即為與自身相比是否成長之角度下，隨機森林模型其中一次訓練之各年度分類結果圖。

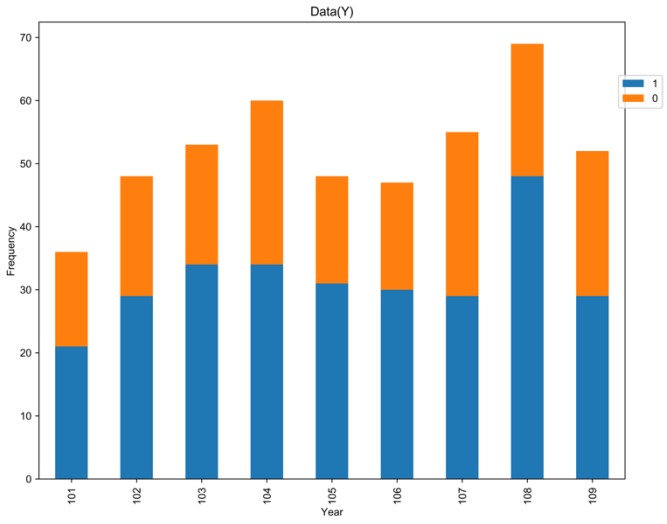

圖 4-2 科技產業隨機森林下次年度營業收入淨額（自身成長）分類圖

doi:10.6342/NTU202201111

表 4-5 科技產業分析次年度績效模型效果（機器學習模型）

訓練 n = 4,219 / 測試 n = 468										
	營業收入淨額（產業平均）					營業收入淨額（自身成長）				
	Precision	Recall	F1-score	Accuracy	AUC	Precision	Recall	F1-score	Accuracy	AUC
羅吉斯迴歸	0.7190	0.7087	0.7135	0.7090	0.7090	0.5726	0.5766	0.5739	0.5513	0.5506
支持向量機	0.7153	0.7131	0.7138	0.7073	0.7073	0.5686	0.5745	0.5707	0.5477	0.5469
隨機森林	0.7161	0.7699	0.7411	0.7250	0.7247	0.5762	0.6858	0.6249	0.5688	0.5640
混合投票制	0.7306	0.7290	0.7294	0.7233	0.7233	0.5747	0.5954	0.5841	0.5560	0.5546
	報酬率					稅前淨利/股東權益				
羅吉斯迴歸	0.5468	0.5933	0.5683	0.5103	0.5026	0.4814	0.4662	0.4731	0.4980	0.4973
支持向量機	0.5389	0.5613	0.5491	0.4992	0.4936	0.4812	0.4760	0.4781	0.4971	0.4969
隨機森林	0.5269	0.7799	0.6283	0.4980	0.4709	0.4438	0.3195	0.3705	0.4747	0.4708
混合投票制	0.5372	0.6145	0.5725	0.5010	0.4906	0.4736	0.4446	0.4582	0.4914	0.4902
	每股盈餘/年底收盤價					稀釋每股盈餘/年底收盤價				
羅吉斯迴歸	0.4438	0.3698	0.4026	0.5274	0.5087	0.4517	0.3795	0.4118	0.5328	0.5146
支持向量機	0.4507	0.4001	0.4229	0.5304	0.5153	0.4537	0.4113	0.4307	0.5319	0.5178
隨機森林	0.4077	0.0875	0.1436	0.5517	0.4959	0.3953	0.0870	0.1420	0.5485	0.4932
混合投票制	0.4504	0.3166	0.3710	0.5381	0.5117	0.4544	0.3310	0.3825	0.5394	0.5145

doi:10.6342/NTU202201111

（二）半導體業

將屬於半導體業的致股東報告書代表向量與其對應之當年度績效分類放入四個 10-fold 交叉驗證模型訓練後，即可探討致股東報告書與當年度財務績效間之關聯性，此部分欲著重的當年度分析概念與科技產業相同，對於績效分類沒有不平等的期待，因此以下會以模型平均 Accuracy 來辨別致股東報告書與當年度績效分析之最合適模型。

半導體業分析致股東報告書與當年度績效關聯性之實證結果如表 4-6。從表中可發現四種機器學習模型訓練效果良好，除了隨機森林的少數衡量指標低於 0.80 外，其餘模型中之任一衡量指標的值皆高於 0.80，顯示半導體業致股東報告書在以「高於同產業平均與否」的角度分析下，其用詞確實與當年度財務績效相關，可從中判別當年度企業營業收入淨額是否優於同業。

再者，四種機器學習模型中又以混合投票制之分類平均 Accuracy 最高，高達 0.8511，加上其餘衡量指標並無怪異之處，故綜合考量後推論出此分析使用混合投票制進行模型訓練是最合適的，圖 4-3 即為此模型其中一次訓練之各年度分類結果圖。

表 4-6 半導體業分析當年度績效模型效果（機器學習模型）

訓練 n = 750 / 測試 n = 83					
	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（產業平均）				
羅吉斯迴歸	0.8473	0.8374	0.8404	0.8403	0.8411
支持向量機	0.8435	0.8310	0.8354	0.8380	0.8373
隨機森林	0.7892	0.8153	0.7999	0.7959	0.7945
混合投票制	0.8554	0.8509	0.8513	0.8511	0.8511

doi:10.6342/NTU202201111

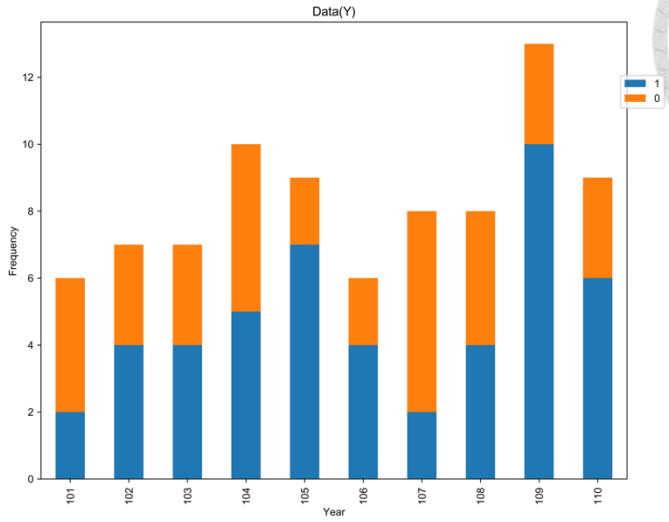

圖 4-3 半導體業混合投票制下當年度營業收入淨額（產業平均）分類圖

接著將屬於半導體業的致股東報告書代表向量與其對應之六種次年度績效分類分別放入四個 10-fold 交叉驗證模型訓練後，即可探討致股東報告書與次年度財務績效間之關聯性，此部分欲著重的次年度分析概念與科技產業相同，對於績效分類雖然沒有存在完全不平等的期待，惟仍然希望實際為 1 的樣本可以有效被預測出，因此以下除了以模型平均 Accuracy 來辨別致股東報告書與次年度績效分析之最合適模型外，也會使用平均 Recall 來輔助挑選。

半導體業分析致股東報告書與次年度績效關聯性之實證結果如表 4-7。從表中可發現六種績效的預測標的中，僅有以「高於同產業平均與否」角度下定義之營業收入淨額能使四種機器學習模型訓練效果良好，不論何種衡量指標的值都位於 0.80 上下；而以「與自身相比是否成長」角度下定義之營業收入淨額表現亦落在本研究尚可接受的範圍內，除 AUC 外其餘衡量指標的值皆高於 0.55；其餘以「與自身相比是否成長」角度分析下之報酬率、稅前淨利除以股東權益、每股盈餘除以年底收盤價、稀釋每股盈餘除以年底收盤價的訓練效果都非常不理想，平均 Accuracy 都位於 0.5 上下，與隨機猜測的結果相似。然而與科技產業相異的是，使用報酬率作為預測標的的隨機森林，Recall 降至 0.69，雖非不甚理想，惟表示

doi:10.6342/NTU202201111

其相對無法有效預測出未來分類實際為 1 的樣本；另外，每股盈餘除以年底收盤價和稀釋每股盈餘除以年底收盤價作為預測標的的隨機森林，Recall 雖從 0 提升至 0.23，惟仍表示正確預測出未來分類實際為 1 的樣本的可能性低。

以上解釋可推論出半導體業致股東報告書用詞並不與每一次年度財務績效的預測標的相關，如欲進行分析僅能從企業營業收入淨額分析，以得知是否可從其中預測次年度企業營業收入淨額是否優於同業及與自身相比是否成長。另外，以「高於同產業平均與否」角度定義之營業收入淨額下之四種模型又以混合投票制之分類平均 Accuracy 最高，為 0.8553，且平均 Recall 也最高，為 0.8601，其餘衡量指標無怪異之處，故綜合考量後推論出此分析使用混合投票制進行模型訓練是最合適的；而以「與自身相比是否成長」角度定義之營業收入淨額下之四種模型又以隨機森林之分類綜合表現最佳，平均 Accuracy 為 0.5600，平均 Recall 為 0.8421，其餘衡量指標無怪異之處，故綜合考量後推論出此分析使用隨機森林進行模型訓練是最合適的。圖 4-4 即為與自身相比是否成長之角度下，隨機森林模型其中一次訓練之各年度分類結果圖，各年度樣本多傾向預測分類為 1 。

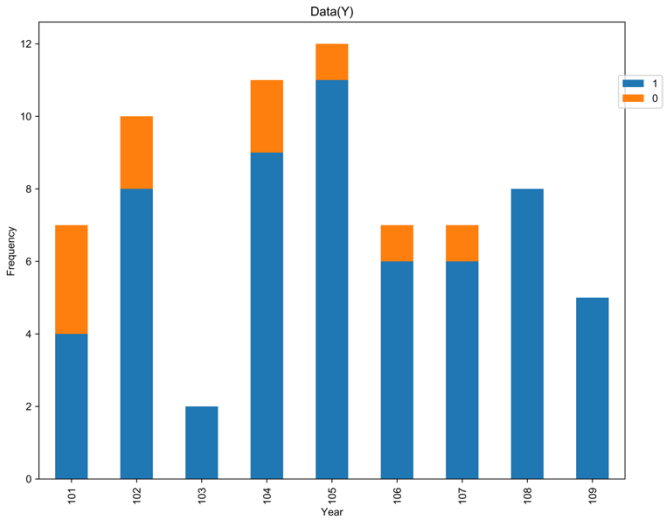

圖 4-4 半導體業隨機森林下次年度營業收入淨額（自身成長）分類圖

doi:10.6342/NTU202201111

表 4-7 半導體業分析次年度績效模型效果（機器學習模型）

訓練 n = 622 / 測試 n = 69										
	Precision	Recall	F1-score	Accuracy	AUC	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（產業平均）					營業收入淨額（自身成長）				
羅吉斯迴歸	0.8534	0.8510	0.8511	0.8509	0.8512	0.6232	0.6644	0.6398	0.5642	0.5458
支持向量機	0.8464	0.8461	0.8440	0.8437	0.8449	0.6307	0.6170	0.6193	0.5585	0.5523
隨機森林	0.7801	0.8287	0.7987	0.7902	0.7945	0.5901	0.8421	0.6905	0.5600	0.5042
混合投票制	0.8545	0.8601	0.8560	0.8553	0.8564	0.6212	0.6809	0.6457	0.5657	0.5455
	報酬率					稅前淨利/股東權益				
羅吉斯迴歸	0.5262	0.5384	0.5303	0.4834	0.4794	0.5121	0.5089	0.5084	0.5152	0.5148
支持向量機	0.5110	0.5137	0.5104	0.4674	0.4642	0.5051	0.5194	0.5094	0.5066	0.5069
隨機森林	0.5160	0.6923	0.5881	0.4761	0.4575	0.4694	0.4270	0.4457	0.4747	0.4751
混合投票制	0.5057	0.5321	0.5165	0.4616	0.4563	0.5009	0.4918	0.4936	0.5037	0.5035
	每股盈餘/年底收盤價					稀釋每股盈餘/年底收盤價				
羅吉斯迴歸	0.4411	0.4276	0.4309	0.5037	0.4947	0.4572	0.4500	0.4503	0.5153	0.5073
支持向量機	0.4684	0.4701	0.4646	0.5225	0.5183	0.4748	0.4882	0.4770	0.5282	0.5248
隨機森林	0.4451	0.2283	0.2979	0.5211	0.4935	0.4536	0.2377	0.3077	0.5254	0.4988
混合投票制	0.4441	0.3987	0.4164	0.5095	0.4977	0.4547	0.4138	0.4286	0.5167	0.5069

doi:10.6342/NTU202201111

（三）電子零組件業

將屬於電子零組件業的致股東報告書代表向量與其對應之當年度績效分類放入四個 10-fold 交叉驗證模型訓練後，即可探討致股東報告書與當年度財務績效間之關聯性，此部分欲著重的當年度分析概念與科技產業和半導體業相同，對於績效分類沒有不平等的期待，因此以下會以模型平均 Accuracy 來辨別致股東報告書與當年度績效分析之最合適模型。

電子零組件業分析致股東報告書與當年度績效關聯性之實證結果如表 4-8。從表中可發現四種機器學習模型訓練效果良好，不論何種衡量指標的值都位於 0.75 至 0.80 之間，顯示電子零組件業致股東報告書在以「高於同產業平均與否」的角度分析下，其用詞確實與當年度財務績效相關，可從中判別當年度企業營業收入淨額是否優於同業。

再者，四種機器學習模型中又以隨機森林之分類平均 Accuracy 最高，為 0.7938，加上其餘衡量指標並無怪異之處，故綜合考量後推論出此分析使用隨機森林進行模型訓練是最合適的，圖 4-5 即為此模型其中一次訓練之各年度分類結果圖。

表 4-8 電子零組件業分析當年度績效模型效果（機器學習模型）

訓練 n = 1,235 / 測試 n = 137					
	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（產業平均）				
羅吉斯迴歸	0.7809	0.7907	0.7823	0.7843	0.7876
支持向量機	0.7637	0.7523	0.7550	0.7595	0.7621
隨機森林	0.7873	0.8089	0.7950	0.7938	0.7974
混合投票制	0.7867	0.7949	0.7876	0.7894	0.7921

doi:10.6342/NTU202201111

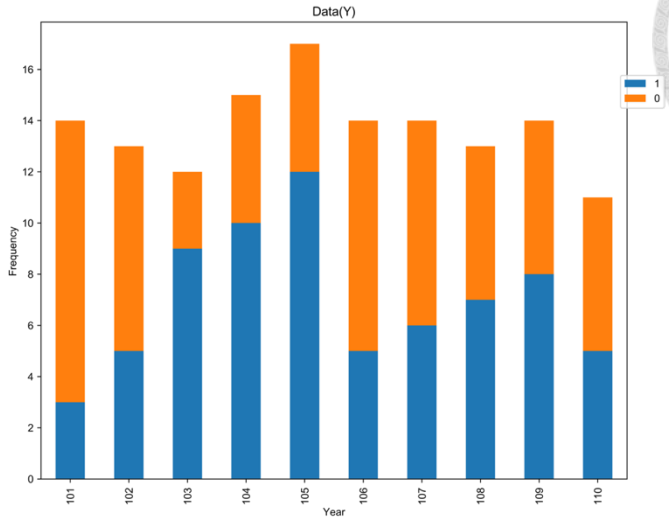

圖 4-5 電子零組件業隨機森林下當年度營業收入淨額（產業平均）分類圖

接著將屬於電子零組件業的致股東報告書代表向量與其對應之六種次年度績效分類分別放入四個 10-fold 交叉驗證模型訓練後，即可探討致股東報告書與次年度財務績效間之關聯性，此部分欲著重的次年度分析概念與科技產業和半導體業相同，對於績效分類雖然沒有存在完全不平等的期待，惟仍然希望實際為 1 的樣本可以有效被預測出，因此以下除了以模型平均 Accuracy 來辨別致股東報告書與次年度績效分析之最合適模型外，也會使用平均 Recall 來輔助挑選。

電子零組件業分析致股東報告書與次年度績效關聯性之實證結果如表 4-9。從表中可發現六種績效的預測標的中，僅有以「高於同產業平均與否」角度下定義之營業收入淨額能使四種機器學習模型訓練效果良好，不論何種衡量指標的值都位於 0.77 上下；而以「與自身相比是否成長」角度下定義之營業收入淨額表現亦落於本研究尚可接受的範圍內，衡量指標的值四捨五入後大多高於 0.55；其餘以「與自身相比是否成長」角度分析下之報酬率、稅前淨利除以股東權益、每股盈餘除以年底收盤價、稀釋每股盈餘除以年底收盤價的訓練效果都非常不理想，平均 Accuracy 都位於 0.5 上下，與隨機猜測的結果相似。再者，與半導體業相同的是，使用報酬率作為預測標的的隨機森林，Recall 依然維持 0.69，表示其依舊相對無

doi:10.6342/NTU202201111

法有效預測出未來分類實際為 1 的樣本；另外，每股盈餘除以年底收盤價和稀釋每股盈餘除以年底收盤價作為預測標的的隨機森林，Recall 依舊極低，僅約 0.15，表示正確預測出未來分類實際為 1 的樣本的可能性極低。

以上解釋可推論出電子零組件業致股東報告書用詞並不與每一次年度財務績效的預測標的相關，如欲進行分析僅能從企業營業收入淨額分析，以得知是否可從其中預測次年度企業營業收入淨額是否優於同業及與自身相比是否成長。另外，以「高於同產業平均與否」角度定義之營業收入淨額下之四種模型又以隨機森林之分類平均 Accuracy 最高，為 0.7882，且平均 Recall 也最高，為 0.8366，其餘衡量指標無怪異之處，故綜合考量後推論出此分析使用隨機森林進行模型訓練是最合適的；而以「與自身相比是否成長」角度定義之營業收入淨額下之四種模型也以隨機森林之分類綜合表現最佳，平均 Accuracy 為 0.5608，平均 Recall 為 0.7621，其餘衡量指標無怪異之處，故綜合考量後推論出此分析使用隨機森林進行模型訓練是最合適的。圖 4-6 即為與自身相比是否成長之角度下，隨機森林模型其中一次訓練之各年度分類結果圖，各年度樣本多傾向預測分類為 1。

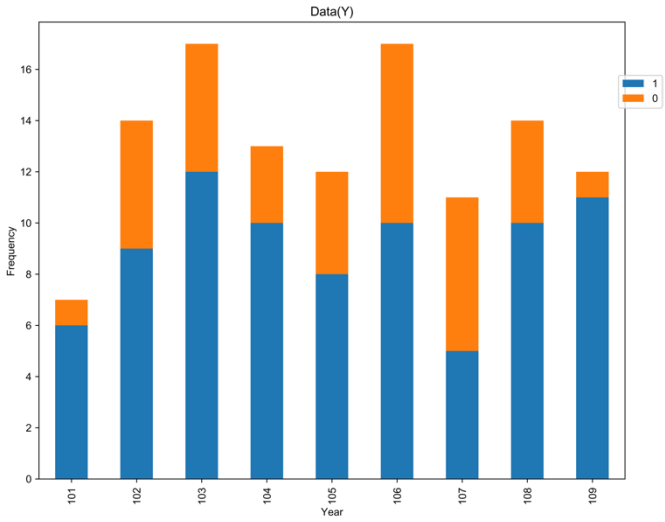

圖 4-6 電子零組件業隨機森林下次年度營業收入淨額（自身成長）分類圖

doi:10.6342/NTU202201111

表 4-9 電子零組件業分析次年度績效模型效果（機器學習模型）

訓練 n = 1,058 / 測試 n = 117										
	Precision	Recall	F1-score	Accuracy	AUC	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（產業平均）					營業收入淨額（自身成長）				
羅吉斯迴歸	0.7706	0.7921	0.7783	0.7728	0.7745	0.5984	0.6195	0.6071	0.5642	0.5596
支持向量機	0.7543	0.7596	0.7540	0.7506	0.7515	0.5713	0.5635	0.5652	0.5267	0.5251
隨機森林	0.7695	0.8366	0.7997	0.7882	0.7893	0.5765	0.7621	0.6535	0.5608	0.5435
混合投票制	0.7832	0.7951	0.7866	0.7830	0.7848	0.5867	0.6215	0.6017	0.5523	0.5467
	報酬率					稅前淨利/股東權益				
羅吉斯迴歸	0.5295	0.5427	0.5344	0.4860	0.4802	0.4836	0.4778	0.4787	0.4979	0.4981
支持向量機	0.5131	0.5011	0.5053	0.4663	0.4633	0.4914	0.5024	0.4946	0.5038	0.5039
隨機森林	0.5206	0.6925	0.5939	0.4817	0.4582	0.4682	0.4059	0.4310	0.4876	0.4866
混合投票制	0.5196	0.5500	0.5329	0.4749	0.4664	0.4799	0.4708	0.4733	0.4945	0.4944
	每股盈餘/年底收盤價					稀釋每股盈餘/年底收盤價				
羅吉斯迴歸	0.4514	0.4247	0.4353	0.5286	0.5162	0.4334	0.3995	0.4139	0.5166	0.5025
支持向量機	0.4239	0.4329	0.4267	0.5030	0.4941	0.4133	0.4320	0.4207	0.4894	0.4822
隨機森林	0.4227	0.1577	0.2275	0.5430	0.4960	0.3877	0.1371	0.2001	0.5294	0.4823
混合投票制	0.4365	0.3812	0.4045	0.5218	0.5048	0.4133	0.3619	0.3836	0.5056	0.4881

doi:10.6342/NTU202201111

第三節 BERT Fine-Tune 模型結果

本節顯示將各產業別致股東報告書經資料前處理後放入 BERT 機器學習模型並經過 Fine-Tune 後的模型分類表現，所使用之模型評估衡量指標與上節一致。

（一）科技產業

將屬於科技產業的致股東報告書與其對應之當年度績效分類放入 BERT Fine-Tuning 模型訓練後，即可探討致股東報告書與當年度財務績效間之關聯性，此部分欲著重的當年度分析概念如上節所述，對於績效分類沒有不平等的期待，因此以下會以模型最終 Accuracy 來辨別致股東報告書與當年度績效分析之最合適模型。

科技產業分析致股東報告書與當年度績效關聯性之實證結果如表 4-10。從表中可發現模型訓練效果良好，除 Precision 為 0.69 外，其餘衡量指標的值都高於 0.70。而因為最終 Accuracy 為 0.7123，故顯示科技產業致股東報告書在以「高於同產業平均與否」的角度分析下，其用詞確實與當年度財務績效相關，可從中判別當年度企業營業收入淨額是否優於同業。

接著將屬於科技產業的致股東報告書與其對應之六種次年度績效分類放入 BERT Fine-Tuning 模型訓練後，即可探討致股東報告書與次年度財務績效間之關聯性，此部分欲著重的次年度分析概念也與上節所述相同，對於績效分類雖然沒有存在完全不平等的期待，惟仍然希望實際為 1 的樣本可以有效被預測出，因此以下除了以模型最終 Accuracy 來辨別致股東報告書與次年度績效分析之最合適模型外，也會使用最終 Recall 來輔助挑選。

科技產業分析致股東報告書與次年度績效關聯性之實證結果如表 4-11。從表中可發現六種績效的預測標的中，僅有以「高於同產業平均與否」角度下定義之營業收入淨額為預測標的時，BERT Fine-Tuning 模型訓練最終 Accuracy 有 0.7271，達到理想的訓練效果；其餘以「與自身相比是否成長」角度下定義之營業收入淨額、報酬率、稅前淨利除以股東權益、每股盈餘除以年底收盤價、稀釋每股盈餘除以年

doi:10.6342/NTU202201111

底收盤價的 BERT Fine-Tuning 模型訓練效果儘管部分衡量指標的值有達 0.55，惟如以全部衡量指標來綜合評估，其結果依然不甚理想。以上解釋可推論出科技產業致股東報告書用詞並不與每一次年度財務績效的預測標的相關，如欲得知是否可從其中預測次年度企業營業收入淨額是否優於同業，對企業營業收入淨額分析會達到較好的結果；然而，如欲得知與自身相比是否成長之資訊，則在此模型下各預測標的皆不合適。

然而，使用「高於同產業平均與否」角度以營業收入淨額作為預測標的之 BERT Fine-Tuning 模型，最終 Accuracy 雖然最高，惟其最終 Recall 只有 0.5848，對於欲找尋出次年度分類為 1 的樣本並不太合適，無法從中獲得哪一樣本次年度績效會在同產業平均之上；反觀使用「與自身相比是否成長」角度以報酬率作為預測標的之 BERT Fine-Tuning 模型，最終 Recall 達到 0.7479，表示其對於未來分類實際為 1 的樣本能有效預測出來，故如欲找尋次年度績效確實成長之樣本以利制定投資決策，此預測標的會較合適。

doi:10.6342/NTU202201111

表 4-10 科技產業分析當年度績效關聯性實證結果（BERT Fine-Tuning）

訓練 n = 5,067 / 測試 n = 562					
	Precision	Recall	F1-score	Accuracy	AUC
營業收入淨額（產業平均）	0.6916	0.7607	0.7245	0.7123	0.7125

表 4-11 科技產業分析次年度績效關聯性實證結果（BERT Fine-Tuning）

訓練 n = 4,219 / 測試 n = 468					
	Precision	Recall	F1-score	Accuracy	AUC
營業收入淨額（產業平均）	0.7892	0.5848	0.6718	0.7271	0.7210
營業收入淨額（自身成長）	0.5748	0.4980	0.5336	0.5416	0.5440
報酬率	0.5443	0.7479	0.6301	0.5544	0.5514
稅前淨利/股東權益	0.5163	0.3559	0.4213	0.5373	0.5281
每股盈餘/年底收盤價	0.5276	0.4236	0.4699	0.5864	0.5671
稀釋每股盈餘/年底收盤價	0.5208	0.3676	0.4310	0.5778	0.5536

doi:10.6342/NTU202201111

（二）半導體業

將屬於半導體業的致股東報告書與其對應之當年度績效分類放入 BERT Fine-Tuning 模型訓練後，即可探討致股東報告書與當年度財務績效間之關聯性，此部分對於績效分類沒有不平等的期待，因此以下會以模型最終 Accuracy 來辨別致股東報告書與當年度績效分析之最合適模型。

半導體業分析致股東報告書與當年度績效關聯性之實證結果如表 4-12。從表中可發現模型訓練效果良好，不論何種衡量指標的值都高於 0.75，顯示半導體業致股東報告書在以「高於同產業平均與否」的角度分析下，其用詞確實與當年度財務績效相關，可從中判別當年度企業營業收入淨額是否優於同業。

接著將屬於半導體業的致股東報告書與其對應之六種次年度績效分類放入 BERT Fine-Tuning 模型訓練後，即可探討致股東報告書與次年度財務績效間之關聯性，此部分對於績效分類雖然沒有存在完全不平等的期待，惟仍然希望實際為 1 的樣本可以有效被預測出，因此以下除了以模型最終 Accuracy 來辨別致股東報告書與次年度績效分析之最合適模型外，亦使用最終 Recall 來輔助挑選。

半導體業分析致股東報告書與次年度績效關聯性之實證結果如表 4-13。從表中可發現六種績效的預測標的中，僅有以「高於同產業平均與否」角度下定義之營業收入淨額為預測標的時，BERT Fine-Tuning 模型訓練最終 Accuracy 有 0.7857，且 Recall 更高達 0.8947，達到理想的訓練效果；而以「與自身相比是否成長」角度下定義之每股盈餘除以年底收盤價、稀釋每股盈餘除以年底收盤價的 BERT Fine-Tuning 模型訓練效果亦落於本研究尚可接受的範圍內，衡量指標的值大多高於 0.55；其餘以「與自身相比是否成長」角度下定義之營業收入淨額、報酬率的 BERT Fine-Tuning 模型訓練效果看似不甚理想，最終 Accuracy 都位於 0.50 至 0.60 之間，與隨機猜測的結果相似，惟其最終 Recall 分別為 0.8000 及 0.7273，對於找尋次年度績效確實成長之樣本十分合適；最後以「與自身相比是否成長」角度下定義之稅前淨利除以股東權益的 BERT Fine-Tuning 模型訓練效果中，最終

doi:10.6342/NTU202201111

Accuracy 僅有 0.3857，比隨機猜測的結果還不準確。以上解釋可推論出半導體業致股東報告書用詞並不與每一次年度財務績效的預測標的相關，如欲得知是否可從其中預測次年度企業營業收入淨額是否優於同業，對企業營業收入淨額分析會達到較好的結果；然而，如欲得知與自身相比是否成長之資訊，則在此模型下使用每股盈餘除以年底收盤價及稀釋每股盈餘除以年底收盤價會較合適。

doi:10.6342/NTU202201111

表 4-12 半導體業分析當年度績效關聯性實證結果（BERT Fine-Tuning）

訓練 n = 750 ／ 測試 n = 83					
	Precision	Recall	F1-score	Accuracy	AUC
營業收入淨額（產業平均）	0.8500	0.7556	0.8000	0.7976	0.8009

表 4-13 半導體業分析次年度績效關聯性實證結果（BERT Fine-Tuning）

訓練 n = 622 ／ 測試 n = 69					
	Precision	Recall	F1-score	Accuracy	AUC
營業收入淨額（產業平均）	0.7556	0.8947	0.8193	0.7857	0.7755
營業收入淨額（自身成長）	0.5357	0.8000	0.6417	0.5180	0.4938
報酬率	0.4800	0.7273	0.5783	0.5000	0.5123
稅前淨利/股東權益	0.3778	0.5313	0.4416	0.3857	0.3972
每股盈餘/年底收盤價	0.5455	0.5806	0.5625	0.6000	0.5980
稀釋每股盈餘/年底收盤價	0.5128	0.6667	0.5797	0.5857	0.5958

doi:10.6342/NTU202201111

（三）電子零組件業

將屬於電子零組件業的致股東報告書與其對應之當年度績效分類放入 BERT Fine-Tuning 模型訓練後，即可探討致股東報告書與當年度財務績效間之關聯性，此部分對於績效分類沒有不平等的期待，因此以下會以模型最終 Accuracy 來辨別致股東報告書與當年度績效分析之最合適模型。

電子零組件業分析致股東報告書與當年度績效關聯性之實證結果如表 4-14。從表中可發現模型訓練效果良好，除最終 Precision 較低外，其餘衡量指標的值都位於 0.75 上下，顯示電子零組件業致股東報告書在以「高於同產業平均與否」的角度分析下，其用詞確實與當年度財務績效相關，可從中判別當年度企業營業收入淨額是否優於同業。

接著將屬於電子零組件業的致股東報告書與其對應之六種次年度績效分類放入 BERT Fine-Tuning 模型訓練後，即可探討致股東報告書與次年度財務績效間之關聯性，此部分對於績效分類雖然沒有存在完全不平等的期待，惟仍然希望實際為 1 的樣本可以有效被預測出，因此以下除了以模型最終 Accuracy 來辨別致股東報告書與次年度績效分析之最合適模型外，也會使用最終 Recall 來輔助挑選。

電子零組件業分析致股東報告書與次年度績效關聯性之實證結果如表 4-15。從表中可發現六種績效的預測標的中，以「高於同產業平均與否」角度下定義之營業收入淨額，及以「與自身相比是否成長」角度下定義之營業收入淨額、報酬率為預測標的時，BERT Fine-Tuning 模型訓練最終 Accuracy 有 0.55 以上，且其三最終 Recall 更分別有 0.7600、0.6250 與 0.8429，達到良好的訓練效果；其餘以「與自身相比是否成長」角度下定義之稅前淨利除以股東權益、每股盈餘除以年底收盤價、稀釋每股盈餘除以年底收盤價的訓練效果皆不甚理想，最終 Accuracy 都位於 0.50 上下，與隨機猜測的結果相似。以上解釋可推論出電子零組件業致股東報告書用詞並不與每一次年度財務績效的預測標的相關，如欲得知是否可從其中預測次年度企業營業收入淨額是否優於同業，對企業營業收入淨額分析會達到較好的

doi:10.6342/NTU202201111

結果；然而，如欲得知與自身相比是否成長之資訊，則在此模型下使用營業收入淨額及報酬率會較合適。

doi:10.6342/NTU202201111

表 4-14 電子零組件業分析當年度績效關聯性實證結果（BERT Fine-Tuning）

訓練 n = 1,235 / 測試 n = 137					
	Precision	Recall	F1-score	Accuracy	AUC
營業收入淨額（產業平均）	0.6795	0.8413	0.7518	0.7464	0.7540

表 4-15 電子零組件業分析次年度績效關聯性實證結果（BERT Fine-Tuning）

訓練 n = 1,058 / 測試 n = 117					
	Precision	Recall	F1-score	Accuracy	AUC
營業收入淨額（產業平均）	0.5588	0.7600	0.6441	0.6441	0.6594
營業收入淨額（自身成長）	0.5682	0.6250	0.5952	0.5660	0.5647
報酬率	0.6211	0.8429	0.7152	0.6017	0.5464
稅前淨利/股東權益	0.5098	0.4194	0.4602	0.4831	0.4865
每股盈餘/年底收盤價	0.4444	0.3019	0.3596	0.5169	0.4971
稀釋每股盈餘/年底收盤價	0.4091	0.1698	0.2400	0.5169	0.4849

doi:10.6342/NTU202201111

第五章 結論及建議

第一節 研究結論與建議

近年來隨著科技發展，文字探勘與自然語言處理的方法日益多元化，引起財務會計領域人士致力將其應用於財會議題上，加上 2018 年 BERT 問世，更有效捕捉文本間的語意關係與結構，故本研究兩種模型訓練過程皆以 BERT 為基礎加以衍生，以多面向之財務指標探討企業致股東報告書與財務績效之關聯性，並以不同機器學習方法進行模型實證分析。

研究結果發現，不論使用任一模型訓練流程，所分析的三種產業別中致股東報告書皆顯示確實與財務績效相關，且特別適合用於分析當年度 (t) 營業收入淨額與預測次年度 (t+1) 營業收入淨額是否高於同產業平均之議題上，其預測結果之 Accuracy 最高可達到 0.70 以上；而如欲以績效與過去自身相比是否成長來進行分析，則可使用營業收入淨額與同產業中位數之差距比例作為預測標的，其預測結果之 Accuracy 於研究中大多高於 0.56。另外以報酬率來看績效是否成長的話，也得出可從致股東報告書內容準確預測出次年度財務績效為樂觀樣本之結論。

另外研究亦發現，所分析的產業別中整體來說 BERT CLS 接上 LDA 的向量形式放入傳統機器學習模型比 BERT Fine-Tuning 模型的訓練效果更佳，印證本研究基於 Peinelt, Nguyen and Liakata (2020) 所做的預期，而 BERT CLS 接上 LDA 的向量形式放入之四種不同機器學習模型中，又以隨機森林和混和投票制模型訓練效果較其他兩種機器學習方法佳。

進一步探討研究所分析之三種產業別致股東報告書的訓練效果，則發現不論是當年度或次年度分析中，半導體業皆為最佳，電子零組件次之，科技產業則是最後，其效果排序與描述統計中平均字數多寡之排序不一致，推論出因模型經訓練後所捕捉之語意關係與結構主要由本身內容主導，已將字數因素的影響性淡化，才導致不同產業發布之致股東報告書與財務績效存在不同程度的關聯性。

doi:10.6342/NTU202201111

第二節 研究限制

此研究主要有兩大限制，一是研究樣本不齊全，在蒐集時因部分企業本身並無公開年報、企業年報為圖片檔或無法複製及企業本身財務數字無法取得，故剩餘有效樣本數約為應有樣本數的一半，然而機器學習模型的優勢即為愈大的數據庫可能有愈佳的訓練效果，樣本數不齊全的情況導致在訓練部分尚有很大的進步空間；二是產業別僅限縮在科技產業，而非全方面將現有公開發行公司之所屬產業別皆納入考量，故本研究所做之結論只適用於科技產業，如欲探討其他研究中未提及之產業，則需另外進行實證分析。

doi:10.6342/NTU202201111

參考文獻

Abrahamson, E. and Amir, E.,1996, “The information content of the president’s letter to shareholders.”, *Journal of Business, Finance and Accounting* 23(8), 1157-1182

Antweiler, W., and Frank, M. Z., 2004, “Is all that talk just noise? The information content of internet stock message boards.”, *The Journal of Finance* 59(3), 1259-1294

Arnold, J. and Moizer, P., 1984, “A survey of the methods used by UK investment analysts to appraise investments in ordinary shares”, *Accounting and Business Research* 14(55), 195-207

Aupperle, K. E., Carroll, A. B. and Hatfield J. D., 1985, “An empirical investigation of the relationship between corporate social responsibility and profitability.”, *Academy of Management Journal* 28, 446-463

Balakrishnan, R., Qiu, X. Y. and Srinivasan, P., 2010, “On the predictive ability of narrative disclosures in annual reports.”, *European Journal of Operational Research* 202(3), 789-801

Blei, D. M., Ng, A. Y. and Jordan, M. I., 2003, “Latent dirichlet allocation.”, *Journal of Machine Learning Research* 3, 993-1022

Breiman, L., 2001, “Random forests.”, *Machine Learning* 45, 5-32

Bryan, S. H., 1997, “Incremental information content of required disclosures contained in management discussion and analysis.”, *The Accounting Review* 72(2), 285-301

Clarkson, P. M., Kao, J. L. and Richardson, G. D., 1999, “Evidence that management discussion and analysis (MD&A) is a part of a firm's overall disclosure package.”, *Contemporary Accounting Research* 16(1), 111-134

Cole, C. J. and Jones, C. L., 2004, “The usefulness of MD&A disclosures in the retail industry.”, *Journal of Accounting, Auditing and Finance* 19, 361-388

Courtis, J. K., 1998, “Annual report readability variability: tests of the obfuscation hypothesis.”, *Accounting, Auditing and Accountability Journal* 11(4), 459- 472

doi:10.6342/NTU202201111

Delvin, J., Chang, M. W., Lee, K. and Toutanova, K., 2018, “BERT: Pre-training of deep bidirectional transformer for language understanding.”, arXiv:1810.04805

Eisenhardt, K. M. and Schoonhoven, C. B., 1996, “Resource-based view of strategic alliance formation: Strategic and social effects in entrpreneurial firms.”, *Organization Science* 7(2), 136-150

Feldman, R., Govindaraj, S., Livnat, J. and Segal B., 2010, “Management’s tone change, post earnings announcement drift and accruals.”, *Review of Accounting Studies* 15(4), 915-953

Frazier, K. B., Ingram, R. W. and Tennyson, B. M., 1984, “A methodology for the analysis of narrative accounting disclosures.”, *Journal of Accounting Research* 22(1): 318-331

Glosten, L. R. and Paul, R. M., 1985, “Bid, ask and transaction prices in a specialist market with heterogeneously informed traders.”, *Journal of Financial Economics* 14, 71-100

Hambrick, D. C. and Mason, P. A., 1984, “Upper echelons: The organization as a reflection of its top managers.”, *Academy of Management Review* 9, 193-206

Healy, P. M. and Palepu, K. G., 2001, “Information asymmetry, corporate disclosure, and the capital markets: A review of the empirical disclosure literature.”, *Journal of Accounting and Economics* 31, 405-440

Henry, E., 2008, “Are investors influenced by the way earnings press releases are written?”, *The Journal of Business Communication* 45, 363-407

Hildebrandt, H. W. and Snyder, R. D., 1981, “The Pollyanna hypothesis in business writing: initial results, suggestions for research.”, *Journal of Business Communication* 18(1), 5-15

Kohut, G. F. and Segars, A. H., 1992, “The president’s letter to stockholders: An examination of corporate communication strategy.”, *Journal of Business Communication* 29, 7-21

Li, F., 2008, “Annual report readability, current earnings, and earnings persistence.”, *Journal of Accounting and Economics* 45(2-3), 221-247

doi:10.6342/NTU202201111

Li, F., 2010, “The information content of forward-looking statements in corporate filings—A naïve Bayesian machine learning approach.”, *Journal of Accounting Research* 48(5), 1049-1102

Loughran, T. and McDonald, B., 2011, “When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks.”, *The Journal of Finance* 66(1), 35-65

McGuire, J. B., Sundgren, A., and Schneeweiss, T., 1988, “Corporate social responsibility and firm financial performance.”, *Academy of Management Journal* 31(4), 854-872

McGuire, J.B., Schneeweiss, T. and Branch, B., 1990, “Perceptions of firm quality: A cause or result of firm performance.”, *Journal of Management* 16(1), 167-180

Peinelt, N., Nguyen, D. and Liakata, M., 2020, “tBERT: Topic models and BERT joining forces for semantic similarity detection.”, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7047-7055

Petersen, M. A., 2004, “Information: Hard and Soft.”, Working paper, Northwestern University.

Previts, G. J., Bricker, R. J., Robinson, T. R. and Young, S. J., 1994, “A content analysis of sell-side financial analyst company reports.”, *Accounting Horizons* 8(2), 55-70

Qiu, X. Y., Srinivasan, P. and Street, N., 2006, “Exploring the forecasting potential of company annual reports.”, *Proceedings of the American Society for Information Science and Technology* 43(1), 1-15

Rogers, R. K. and Grant, J., 1997, “Content analysis of information cited in reports of sell-side financial analysts.”, *The Journal of Financial Statement Analysis* 3(1), 17-31

Salton, G., Wong, A. and Yang, C. S., 1975, “A vector space model for automatic indexing.”, *Communications of the ACM* 18(11), 613-620

Schroeder, N. and Gibson, C., 1990, “Readability of management’s discussion and analysis.”, *Accounting Horizons* 4(4), 78–87

doi:10.6342/NTU202201111

Smith, M. and Taffler, R. J., 2000, “The chairman’s statement- A content analysis of discretionary narrative disclosures.”, *Accounting, Auditing and Accountability Journal* 13(5), 624-646

Sun, Y., 2010, “Do MD&A disclosures help users interpret disproportionate inventory increases?”, *The Accounting Review* 85(4), 1411-1440

Swales, Jr., G. S., 1988, “Another Look at the President's Letter to Stockholders.”, *Financial Analysts Journal*, 71-73

U.S. Securities and Exchange Commission (SEC), 1987, “Concept release on management's discussion and analysis of financial condition and results of operations.”, *Securities Act Release* No. 6711.

U.S. Securities and Exchange Commission (SEC), 2003, “Interpretation: commission guidance regarding management’s discussion and analysis of financial condition and results of operations.”, *Securities Act Release* No. 8350.

Waddock, S. A. and Graves, S. B., 1997, “The corporate social performance-financial performance link.”, *Strategic Management Journal* 18, 303-319

doi:10.6342/NTU202201111

附錄一 BERT CLS 向量形式實證結果

附錄一顯示將各產業別之致股東報告書經 BERT 處理後得出之 CLS Embedding 放入四種機器學習模型訓練之實證結果，評估衡量數據可參以下各表。

科技產業					
訓練 n = 5,067 / 測試 n = 562					
	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（當年度產業平均）				
羅吉斯迴歸	0.5581	0.5537	0.5554	0.5570	0.5570
支持向量機	0.5490	0.5526	0.5505	0.5483	0.5484
隨機森林	0.5414	0.5451	0.5428	0.5407	0.5409
混合投票制	0.5573	0.5559	0.5560	0.5562	0.5563

半導體業					
訓練 n = 750 / 測試 n = 83					
	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（當年度產業平均）				
羅吉斯迴歸	0.5549	0.5371	0.5395	0.5491	0.5541
支持向量機	0.5478	0.5243	0.5278	0.5419	0.5480
隨機森林	0.5557	0.5501	0.5467	0.5515	0.5560
混合投票制	0.5691	0.5408	0.5466	0.5610	0.5668

電子零組件業					
訓練 n = 1,235 / 測試 n = 137					
	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（當年度產業平均）				
羅吉斯迴歸	0.5489	0.5602	0.5500	0.5452	0.5495
支持向量機	0.5390	0.5462	0.5382	0.5358	0.5399
隨機森林	0.5386	0.5448	0.5357	0.5336	0.5391
混合投票制	0.5380	0.5536	0.5414	0.5358	0.5400

doi:10.6342/NTU202201111

科技產業										
訓練 n = 4,219 / 測試 n = 468										
	營業收入淨額（次年度產業平均）					營業收入淨額（自身成長）				
	Precision	Recall	F1-score	Accuracy	AUC	Precision	Recall	F1-score	Accuracy	AUC
羅吉斯迴歸	0.5581	0.5537	0.5554	0.5570	0.5570	0.5302	0.6304	0.5747	0.5110	0.5060
支持向量機	0.5490	0.5526	0.5505	0.5483	0.5484	0.5292	0.6504	0.5815	0.5099	0.5044
隨機森林	0.5414	0.5451	0.5428	0.5407	0.5409	0.5145	0.5468	0.5291	0.4901	0.4881
混合投票制	0.5573	0.5559	0.5560	0.5562	0.5563	0.5270	0.6363	0.5748	0.5069	0.5018
	報酬率					稅前淨利/股東權益				
羅吉斯迴歸	0.5533	0.6060	0.5776	0.5180	0.5100	0.5012	0.4845	0.4921	0.5159	0.5153
支持向量機	0.5485	0.5945	0.5698	0.5118	0.5041	0.4979	0.4682	0.4822	0.5133	0.5120
隨機森林	0.5471	0.6495	0.5932	0.5159	0.5032	0.4986	0.4647	0.4805	0.5133	0.5122
混合投票制	0.5506	0.6155	0.5803	0.5159	0.5068	0.5038	0.4795	0.4910	0.5184	0.5175
	每股盈餘/年底收盤價					稀釋每股盈餘/年底收盤價				
羅吉斯迴歸	0.4301	0.3490	0.3842	0.5193	0.4991	0.4287	0.3463	0.3819	0.5182	0.4978
支持向量機	0.4273	0.3571	0.3884	0.5155	0.4965	0.4145	0.3429	0.3746	0.5067	0.4872
隨機森林	0.4266	0.2918	0.3449	0.5236	0.4965	0.4216	0.2952	0.3457	0.5197	0.4936
混合投票制	0.4281	0.3243	0.3680	0.5208	0.4975	0.4204	0.3181	0.3612	0.5155	0.4920

doi:10.6342/NTU202201111

半導體業										
訓練 n = 622 / 測試 n = 69										
	Precision	Recall	F1-score	Accuracy	AUC	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（次年度產業平均）					營業收入淨額（自身成長）				
羅吉斯迴歸	0.5413	0.5268	0.5304	0.5312	0.5331	0.5769	0.8266	0.6774	0.5411	0.4811
支持向量機	0.5573	0.5418	0.5439	0.5443	0.5487	0.5732	0.7754	0.6573	0.5310	0.4780
隨機森林	0.5517	0.5371	0.5358	0.5355	0.5432	0.6170	0.7368	0.6696	0.5775	0.5439
混合投票制	0.5639	0.5502	0.5511	0.5501	0.5549	0.5825	0.8019	0.6725	0.5455	0.4917
	報酬率					稅前淨利/股東權益				
羅吉斯迴歸	0.5146	0.5454	0.5247	0.4718	0.4627	0.5189	0.5188	0.5160	0.5268	0.5248
支持向量機	0.5248	0.5455	0.5305	0.4848	0.4766	0.5433	0.5049	0.5207	0.5471	0.5436
隨機森林	0.5399	0.6262	0.5748	0.5037	0.4918	0.5323	0.5478	0.5376	0.5383	0.5369
混合投票制	0.5261	0.5662	0.5405	0.4848	0.4755	0.5337	0.5181	0.5224	0.5398	0.5369
	每股盈餘/年底收盤價					稀釋每股盈餘/年底收盤價				
羅吉斯迴歸	0.4937	0.4531	0.4706	0.5457	0.5330	0.4916	0.4580	0.4722	0.5413	0.5300
支持向量機	0.4690	0.4523	0.4579	0.5254	0.5142	0.4807	0.4677	0.4708	0.5327	0.5232
隨機森林	0.5068	0.4372	0.4683	0.5558	0.5410	0.5118	0.4479	0.4760	0.5587	0.5453
混合投票制	0.4900	0.4484	0.4657	0.5428	0.5287	0.4950	0.4639	0.4762	0.5442	0.5328

doi:10.6342/NTU202201111

電子零組件業										
訓練 n = 1,058 / 測試 n = 117										
	Precision	Recall	F1-score	Accuracy	AUC	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（次年度產業平均）					營業收入淨額（自身成長）				
羅吉斯迴歸	0.5806	0.5682	0.5717	0.5677	0.5694	0.5500	0.7179	0.6204	0.5226	0.5033
支持向量機	0.5693	0.5397	0.5495	0.5515	0.5556	0.5467	0.7224	0.6198	0.5191	0.4994
隨機森林	0.5730	0.5667	0.5645	0.5575	0.5615	0.5493	0.5926	0.5669	0.5089	0.5023
混合投票制	0.5825	0.5638	0.5688	0.5668	0.5701	0.5478	0.7129	0.6172	0.5191	0.5002
	報酬率					稅前淨利/股東權益				
羅吉斯迴歸	0.5362	0.5702	0.5497	0.4928	0.4843	0.4399	0.4172	0.4254	0.4570	0.4567
支持向量機	0.5200	0.5396	0.5264	0.4740	0.4674	0.4333	0.4077	0.4154	0.4502	0.4507
隨機森林	0.5383	0.6278	0.5772	0.4996	0.4857	0.4444	0.4282	0.4330	0.4604	0.4612
混合投票制	0.5283	0.5653	0.5430	0.4834	0.4751	0.4419	0.4167	0.4247	0.4587	0.4590
	每股盈餘/年底收盤價					稀釋每股盈餘/年底收盤價				
羅吉斯迴歸	0.4274	0.3939	0.4070	0.5150	0.5001	0.4232	0.3916	0.4040	0.5099	0.4955
支持向量機	0.4399	0.4099	0.4208	0.5218	0.5083	0.4385	0.4038	0.4168	0.5184	0.5048
隨機森林	0.4259	0.3249	0.3652	0.5150	0.4932	0.4327	0.3254	0.3683	0.5209	0.4984
混合投票制	0.4331	0.3844	0.4041	0.5184	0.5021	0.4243	0.3742	0.3943	0.5107	0.4947

doi:10.6342/NTU202201111

附錄二 LDA 向量形式實證結果

附錄二顯示將各產業別之致股東報告書單單經 LDA 處理後直接放入四種機器學習模型訓練之實證結果，評估衡量數據可參以下各表。

科技產業					
訓練 n = 5,067 / 測試 n = 562					
	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（當年度產業平均）				
羅吉斯迴歸	0.6179	0.5966	0.6061	0.6120	0.6125
支持向量機	0.6261	0.5702	0.5960	0.6134	0.6138
隨機森林	0.5747	0.5708	0.5721	0.5729	0.5732
混合投票制	0.6212	0.5866	0.6026	0.6129	0.6133

半導體業					
訓練 n = 750 / 測試 n = 83					
	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（當年度產業平均）				
羅吉斯迴歸	0.6389	0.6302	0.6319	0.6338	0.6332
支持向量機	0.6361	0.6232	0.6274	0.6314	0.6296
隨機森林	0.5244	0.5536	0.5350	0.5247	0.5250
混合投票制	0.6328	0.6251	0.6266	0.6277	0.6267

電子零組件業					
訓練 n = 1,235 / 測試 n = 137					
	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（當年度產業平均）				
羅吉斯迴歸	0.5997	0.5554	0.5738	0.5940	0.5948
支持向量機	0.6027	0.5400	0.5675	0.5948	0.5945
隨機森林	0.5562	0.5599	0.5551	0.5583	0.5591
混合投票制	0.6021	0.5469	0.5702	0.5948	0.5954

doi:10.6342/NTU202201111

科技產業										
訓練 n = 4,219 / 測試 n = 468										
	營業收入淨額（次年度產業平均）					營業收入淨額（自身成長）				
	Precision	Recall	F1-score	Accuracy	AUC	Precision	Recall	F1-score	Accuracy	AUC
羅吉斯迴歸	0.6185	0.6638	0.6400	0.6183	0.6171	0.5505	0.5795	0.5645	0.5310	0.5282
支持向量機	0.6178	0.6559	0.6356	0.6155	0.6151	0.5527	0.5687	0.5602	0.5325	0.5305
隨機森林	0.5851	0.5953	0.5896	0.5761	0.5758	0.5359	0.6014	0.5665	0.5170	0.5128
混合投票制	0.6238	0.6620	0.6418	0.6221	0.6215	0.5474	0.5803	0.5631	0.5281	0.5250
	報酬率					稅前淨利/股東權益				
羅吉斯迴歸	0.5233	0.7463	0.6138	0.4899	0.4664	0.4811	0.3879	0.4286	0.4990	0.4964
支持向量機	0.5319	0.8510	0.6541	0.5103	0.4772	0.4848	0.3431	0.4008	0.5037	0.4996
隨機森林	0.5371	0.6040	0.5675	0.4995	0.4902	0.4955	0.4800	0.4866	0.5099	0.5096
混合投票制	0.5261	0.7923	0.6313	0.4969	0.4691	0.4822	0.3656	0.4148	0.5010	0.4977
	每股盈餘/年底收盤價					稀釋每股盈餘/年底收盤價				
羅吉斯迴歸	0.4335	0.1361	0.2050	0.5492	0.4998	0.4222	0.1341	0.2019	0.5460	0.4969
支持向量機	0.4235	0.0787	0.1309	0.5547	0.4977	0.4120	0.0767	0.1276	0.5528	0.4959
隨機森林	0.4316	0.3592	0.3916	0.5197	0.5002	0.4314	0.3541	0.3886	0.5200	0.4998
混合投票制	0.4293	0.1117	0.1751	0.5505	0.4979	0.4222	0.1096	0.1723	0.5485	0.4961

doi:10.6342/NTU202201111

半導體業										
訓練 n = 622 / 測試 n = 69										
	Precision	Recall	F1-score	Accuracy	AUC	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（次年度產業平均）					營業收入淨額（自身成長）				
羅吉斯迴歸	0.5868	0.6423	0.6097	0.5876	0.5874	0.6214	0.6503	0.6325	0.5616	0.5430
支持向量機	0.5868	0.6199	0.6006	0.5861	0.5849	0.6267	0.6712	0.6449	0.5717	0.5518
隨機森林	0.5764	0.6114	0.5915	0.5730	0.5714	0.6043	0.7167	0.6539	0.5587	0.5251
混合投票制	0.5895	0.6325	0.6069	0.5890	0.5887	0.6300	0.6851	0.6533	0.5789	0.5565
	報酬率					稅前淨利/股東權益				
羅吉斯迴歸	0.5448	0.7201	0.6173	0.5180	0.4984	0.4734	0.4520	0.4613	0.4805	0.4788
支持向量機	0.5446	0.7210	0.6176	0.5180	0.4971	0.4756	0.4257	0.4483	0.4834	0.4813
隨機森林	0.5468	0.6139	0.5757	0.5079	0.4984	0.4909	0.4832	0.4863	0.4949	0.4940
混合投票制	0.5436	0.7140	0.6140	0.5151	0.4952	0.4889	0.4491	0.4668	0.4949	0.4933
	每股盈餘/年底收盤價					稀釋每股盈餘/年底收盤價				
羅吉斯迴歸	0.4451	0.2551	0.3193	0.5224	0.4985	0.4640	0.2724	0.3384	0.5282	0.5056
支持向量機	0.4401	0.2543	0.3166	0.5195	0.4958	0.4577	0.2662	0.3306	0.5253	0.5027
隨機森林	0.4618	0.4120	0.4287	0.5180	0.5115	0.4834	0.4219	0.4449	0.5325	0.5254
混合投票制	0.4451	0.2581	0.3212	0.5209	0.4977	0.4637	0.2730	0.3382	0.5267	0.5048

doi:10.6342/NTU202201111

電子零組件業										
訓練 n = 1,058 / 測試 n = 117										
	Precision	Recall	F1-score	Accuracy	AUC	Precision	Recall	F1-score	Accuracy	AUC
	營業收入淨額（次年度產業平均）					營業收入淨額（自身成長）				
羅吉斯迴歸	0.6355	0.6429	0.6338	0.6247	0.6274	0.5537	0.5765	0.5639	0.5165	0.5086
支持向量機	0.6474	0.6174	0.6275	0.6298	0.6309	0.5670	0.5724	0.5686	0.5268	0.5215
隨機森林	0.5737	0.5806	0.5752	0.5659	0.5650	0.5660	0.6467	0.6021	0.5361	0.5244
混合投票制	0.6443	0.6267	0.6309	0.6298	0.6315	0.5629	0.6025	0.5811	0.5276	0.5183
	報酬率					稅前淨利/股東權益				
羅吉斯迴歸	0.6355	0.6429	0.6338	0.6247	0.6274	0.4695	0.4367	0.4484	0.4842	0.4855
支持向量機	0.6474	0.6174	0.6275	0.6298	0.6309	0.4644	0.4380	0.4466	0.4774	0.4790
隨機森林	0.5737	0.5806	0.5752	0.5659	0.5650	0.4832	0.4822	0.4801	0.4987	0.4990
混合投票制	0.6443	0.6267	0.6309	0.6298	0.6315	0.4783	0.4502	0.4599	0.4901	0.4915
	每股盈餘/年底收盤價					稀釋每股盈餘/年底收盤價				
羅吉斯迴歸	0.3735	0.1782	0.2384	0.5132	0.4736	0.3755	0.1815	0.2418	0.5114	0.4726
支持向量機	0.3911	0.1817	0.2449	0.5191	0.4789	0.3835	0.1835	0.2456	0.5149	0.4757
隨機森林	0.4223	0.3356	0.3718	0.5148	0.4915	0.4314	0.3546	0.3875	0.5182	0.4969
混合投票制	0.3859	0.1838	0.2457	0.5166	0.4770	0.3874	0.1894	0.2514	0.5157	0.4772

doi:10.6342/NTU202201111

附錄三 致股東報告書範例

附錄三為台灣積體電路製造股份有限公司民國一百零八年度致股東報告書內容，資料取自該企業民國一百零八年度年報，為本研究中科技產業及半導體業樣本之一。

各位股東女士、先生：

民國一百零八年是台積公司持續達成許多里程碑的一年。儘管面臨國際間貿易緊張局勢所帶來業務上的逆風，台積公司的營收依舊連續十年創下紀錄。貿易緊張局勢也帶給我們客戶更高的不確定性，同時影響產品的終端需求。然而，受惠於客戶對我們領先業界的 7 奈米（N7）製程技術的強勁需求，台積公司民國一百零八年的營收，若以美元計算，相較於民國一百零七年增加了 1.3%，而全球半導體產業較前一年則減少了 12%。

民國一百零八年，我們見證了 5G 網路和智慧型手機在全球幾個主要市場的加速運用。我們預計未來幾年，晶片內含量更高的 5G 智慧型手機將更快普及於全球。5G 智慧型手機要求更高效能、更快速及更複雜功能，將導致增加採用台積公司領先業界的技術。為滿足此需求量提升的情況，台積公司將民國一百零八年的資本支出提高至 149 億美元，我們預期此一需求將持續成長。

台積公司於民國一百零八年持續致力於強化業務的基本體質，藉由提升品質系統以提供客戶更好的服務，擴充我們的研發基礎架構，增強資訊架構和資訊安全，以及加速我們的技術差異化。

台積公司持續提供業界最先進的技術，使所有產品創新者得以成功應用，從而不斷擴大產品創新者的規模，以推動半導體業的成長。

我們的 N7 製程技術在民國一百零八年邁入量產的第二年，在行動裝置、高效能運算（HPC）、物聯網（IoT）和車用電子等眾多產品中持續被廣泛採用。我們新推出的 7 奈米強效版（N7+）製程技術亦領先全球導入極紫外光（EUV）微影技

doi:10.6342/NTU202201111

術進行量產。民國一百零八年，我們的 N7 家族，包括 N7 及 N7+ 製程技術的營收佔全年晶圓銷售金額的 27%。我們的 6 奈米（N6）製程技術剛於民國一百零九年第一季進入試產階段，並成功進一步擴展了 N7 家族的未來性。

我們的 5 奈米（N5）製程技術已廣泛採用 EUV 技術，將於民國一百零九年上半年開始量產。作為業界最先進的解決方案，N5 製程技術已進一步擴大我們的客戶產品組合，同時擴增我們的潛在市場。

我們的 3 奈米（N3）製程技術將是繼 N5 後另一全節點提升的製程，並將於推出時提供業界最佳的功耗/效能/面積（PPA）的製程技術。

台積公司獨有的晶圓級封裝解決方案，包括整合型扇出（Integrated Fan-Out, InFO）和 CoWoS®（Chip on Wafer on Substrate）持續保持強勁成長。我們正在開發三維晶片堆疊解決方案，例如系統整合晶片（System on Integrated Chip, SoIC），以提供業界系統級解決方案。

台積公司民國一百零八年的主要成就包括：

● 晶圓出貨量達 1,010 萬片 12 吋晶圓約當量，民國一百零七年為 1,080 萬片 12 吋晶圓約當量。

● 先進製程技術（16 奈米及以下更先進製程）的銷售金額佔整體晶圓銷售金額的 50%，高於民國一百零七年的 41%。

● 提供 272 種不同的製程技術，為 499 個客戶生產 10,761 種不同產品。

● 在專業積體電路製造服務領域之佔有率達 52%，高於民國一百零七年的 51%。

財務表現

台積公司民國一百零八年全年合併營收為新台幣 1 兆 699 億 9,000 萬元，較前一年的 1 兆 314 億 7,000 萬元增加了 3.7%；稅後淨利為新台幣 3,452 億 6,000 萬元，每股盈餘為新台幣 13.32 元，較前一年稅後淨利 3,511 億 3,000 萬元及每股盈餘 13.54 元均減少了 1.7%。

doi:10.6342/NTU202201111

若以美元計算，台積公司民國一百零八年全年合併營收為 346 億 3,000 萬美元，稅後淨利為 111 億 8,000 萬美元，較前一年度的全年合併營收 342 億美元增加 1.3%，較前一年度的稅後淨利 116 億 4,000 萬美元則減少了 4.0%。

台積公司民國一百零八年毛利率為 46.0%，前一年為 48.3%；營業利益率為 34.8%，前一年則為 37.2%。稅後純益率為 32.3%，較前一年的稅後純益率 34.0% 減少了 1.7 個百分點。

為使公司獲利得以提前回饋予股東，台積公司於民國一百零八年將現金股利由年度配發過渡為季度配發，並將現金股利配發總額由前一年度的每股新台幣 8 元進一步提高至每股新台幣 10 元。

技術發展

民國一百零八年，台積公司持續增加研發費用，達到 29 億 6,000 萬美元的歷史新高，以滿足客戶的需求，並延續技術上的領導地位。

民國一百零八年，我們的 N5 製程技術進入試產，預計於民國一百零九年上半年開始量產。N5 製程技術可望拓展我們客戶的產品組合，並且隨著客戶尋求建立其產品領導地位時，擴大我們的潛在市場。

台積公司在 N7 製程技術量產的第二年，即於民國一百零八年底前取得超過 100 件客戶產品設計定案，同時，採用 EUV 的 N7+ 製程技術也開始量產。我們的 N6 製程技術符合進度，預計於民國一百零九年底前進入量產，為下一波 N7 產品提供了一個明確的升級路徑。

民國一百零八年，我們利用在 28 奈米製程技術上的領先地位開始量產 22 奈米超低功耗（22ULP）及 22 奈米超低漏電（22ULL）製程技術，22 奈米超低漏電技術支援物聯網與穿戴式裝置應用，22 奈米超低功耗技術則支援影像處理、數位電視、機上盒、以及其他消費性電子產品。民國一百零八年，我們也擴展 16 奈米技術組合到 12 奈米精簡型強效版（12FFC+）製程技術及 16 奈米精簡型

doi:10.6342/NTU202201111

強效版（16FFC+）製程技術，以支援客戶在超低功耗應用上的需求。

台積公司藉由無縫整合的前段晶圓製程與後段晶片封裝，提供先進封裝解決方案，實現了晶圓級製程的系統整合。民國一百零八年，我們推出具備更精細的連結線寬與間距的第五代 InFO 解決方案，來支援行動裝置及高效能運算產品。台積公司 CoWoS® 技術持續更大尺寸中介層的異質整合。台積公司也開發了系統整合晶片（System-on-Integrated Chips, TSMC-SoIC®），此項領先業界的三維晶片堆疊解決方案能夠整合多個非常鄰近的晶片，提供最佳的系統效能。

台積公司的開放創新平台（Open Innovation Platform®, OIP）設計生態系統協助 499 家客戶釋放創新，將產品快速上市。民國一百零八年，台積公司持續增加開放創新平台雲端聯盟的合作夥伴，使得客戶能在一個安全可靠的雲端環境進行晶片設計，顯著提升了客戶的設計生產力。我們也持續與設計生態系統的夥伴合作，於民國一百零八年將資料庫與矽智財組合擴增到超過 26,000 個項目。台積公司已在 TSMC-Online 線上提供自 0.5 微米至 5 奈米超過 10,600 個技術檔案及超過 360 個製程設計套件，民國一百零八年客戶下載使用技術檔案與製程設計套件已超過 10 萬次。

企業社會責任

台積公司致力於健全的公司治理，追求獲利成長，同時也重視環境、社會，以及所有利害關係人的利益平衡。健全的公司治理奠基於我們的企業核心價值之上，也是台積公司企業社會責任的基石。身為全球半導體產業重要的成員之一，我們將日益挑戰的全球環境視為自身責任，並且身先士卒，有所作為。

民國一百零八年，我們成立了企業社會責任執行委員會，由董事長擔任主席，與多位不同領域的高階主管及既有的企業社會責任委員會一同訂定公司企業社會責任的策略方針，並呼應聯合國永續發展目標。我們著重於推動綠色製造、打造包容職場、培育人才、建立責任供應鏈，以及關懷弱勢。台積公司將致力實踐企業公民角色，追求永續未來。

doi:10.6342/NTU202201111

榮譽與獎項

民國一百零八年，台積公司在創新、公司治理、永續發展、投資人關係、資訊揭露以及整體傑出經營管理方面，獲得來自富比世雜誌、財富雜誌、日本經濟新聞社、天下雜誌、資誠聯合會計師事務所、RobecoSAM（S&P Global）、台灣證交所等頒發的多項榮譽與獎項。在技術創新方面，台積公司於美國專利商標局的專利申請數量排名第十，亦獲得本國法人專利申請百大排名第一名的榮譽。在永續發展方面，台積公司再次獲選道瓊永續世界指數的組成企業，是全球唯一連續 19 年入選的半導體公司，亦獲企業騎士評選為 2019 全球百大最佳永續發展企業第十名。此外，台積公司持續獲選為 MSCI 全球 ESG 領導者指數以及 FTSE4Good 新興市場指數之重要成分股。在投資人關係方面，台積公司持續獲得來自 Institutional Investor 雜誌的多項榮譽。

未來展望

我們相信 5G 網路在通訊上帶來的顯著進展，將開啟許多不同類型的終端連結裝置間嶄新的應用模式，並且驅動數據量呈指數成長。隨著演算法的持續創新，一個更聰明且更智慧化的社會也應運而生。數位運算如今已逐漸變得無所不在，同時需要大量的運算能力。因此，我們期待 5G 相關及高效能運算應用的發展，將會在未來幾年帶動對於我們先進技術的強勁需求。憑藉著最先進的技術與產能，以及最廣泛的客戶群，台積公司正處於最佳的位置引領業界掌握市場的成長。

民國一百零九年，國際間貿易緊張局勢造成總體經濟的不確定性持續存在，台積公司將保持靈活的應變能力，致力於業務的基本體質，進一步加速技術的差異化。我們是「大家的晶圓技術產能提供者」（everyone's foundry），公平且公正的對待所有客戶，我們會盡全力保護智慧財產，秉持最高的誠信正直原則經營業務，並且堅守技術領先、卓越製造及客戶信任的三位一體競爭優勢。

台積公司的專業積體電路製造服務商業模式、開放創新平台，以及涵蓋誠信正直、承諾、創新與客戶信任的四大核心價值，使我們成為「大家的晶圓技術產能提供者」。進入嶄新的數位化時代，我們將持續與全世界的積體電路創新者緊密合作，

doi:10.6342/NTU202201111

創造價值並且為股東賺取優良的報酬。我們致力於健全的公司治理，善盡全球企業公民的社會責任，追求永續發展。感謝各位股東對台積公司的信任與支持，我們期待與您攜手共同邁向繁榮的未來。

劉德音　魏哲家

董事長　總裁

doi:10.6342/NTU202201111