

## 應用遷移學習與文字探勘分析致股東報告書

陳予得<sup>1</sup> 林嬋娟<sup>2</sup>

<sup>1</sup> 新加坡商蝦皮娛樂電商

<sup>2</sup> 國立臺灣大學會計學系

通訊作者：林嬋娟

通訊地址：10617 臺北市羅斯福路四段 1 號

E-mail: cjlin@ntu.edu.tw

投稿日期：2021 年 7 月 19 日；2 審後接受，接受日期：2022 年 3 月 28 日

### 摘 要

本研究採自然語言處理方法 (Bidirectional Encoder Representation from Transformers, BERT) 建立文字探勘模型，以經標記之國內半導體業致股東報告書訓練 BERT。本研究亦分析 BERT 是否改善過往文字探勘方法的缺點，最後以情緒分析剖析致股東報告書語調對公司未來績效的影響。實證結果顯示，經驗證資料集表現篩選超參數 (hyperparameter) 後，BERT 測試資料集分類準確率高達 0.86。透過 BERT 視覺化釋例，本文發現其能捕捉否定詞修飾的詞彙，且能捕捉形容詞所修飾的名詞。惟與 Li (2010a) 使用 MD&A 語調之研究結果不同，本文實證並未發現當年致股東報告書情緒與下一年盈餘及盈餘變動具顯著正關聯，推論原因可能是國內投資人結構或資訊透明度特性，導致致股東報告書與美國 MD&A 資訊內涵不同。

**關鍵詞：**遷移學習、情緒分析、盈餘預測、致股東報告書

---

作者感謝領域主編以及兩位匿名評審委員之寶貴意見，文中言論由作者自行負責。

數據可用性：本文使用的數據可從公開資料來源取得。



 東華書局  
Tung Hua Book Co., Ltd.

## 1. 前言

美國 Securities and Exchange Commission (SEC) 要求公開發行公司於年報中揭露管理階層討論與分析 (Management Discussion and Analysis, MD&A)，其認為財務報表及其附註無法提供充分資訊，以協助財務報表使用者判斷公司盈餘品質及預測公司未來績效 (SEC 1987)。因此 SEC 要求公司提供 MD&A，使讀者瞭解可能重大影響公司現在或未來營運、流動性或資本之市場趨勢、需求、事件或不確定性 (SEC 2003)。<sup>1</sup>

過去許多文獻顯示，MD&A 在討論未來營運方向或資本支出未來規劃的確會幫助財務報表使用者預測公司未來表現。例如，Bryan (1997) 調查 MD&A 是否能提供其他額外的資訊，其研究 MD&A 語調與未來銷售、資本支出、營運現金流量以及盈餘變化之關聯，結果發現 MD&A 的確與未來財務數據相關。Cole and Jones (2004) 發現在零售業裡，有關 MD&A 揭露收入變化之原因 (如既有店面銷售成長或擴增新店面，以及揭露未來資本支出)，有助於投資者預測未來銷貨收入及盈餘。Sun (2010) 發現，MD&A 存貨相關揭露可幫助使用者解讀存貨大量增加的現象以及預測未來公司表現。Bochkay and Levine (2019) 透過字典法與文字頻率矩陣建立文字資訊相關變數，發現財務數據結合 MD&A 之文字資訊相較於僅財務數據的模型，能更準確預測未來一年的股東權益報酬率。然而前述研究須依賴人工進行分類，導致樣本數量過少，可能會造成推論無法普及的問題。

檢視文本分析相關文獻，發現許多不同文本分析方法試圖解決上段所述之問題，例如，Li (2010a) 從 MD&A 中的前瞻性敘述 (Forward-Looking Statement, FLS) 隨機抽取三萬行句子訓練朴素貝氏分類器 (Naïve Bayesian Classifier)，並對一千多萬行句子進行語調之分類，其發現 FLS 與未來盈餘及盈餘變動呈現顯著正相關。Loughran and McDonald (2011) 利用字典法對文本進行分類，將文本語調與公司股價、交易量、舞弊等進行連結。Siano and Wysocki (2021) 利用 BERT (Bidirectional Encoder Representation from Transformers)、字典法及朴素貝氏分類器進行語調預測準確度比較，發現 BERT 表現得更為準確。該文認為，過去會計研究受限於較少資訊背景以及硬體設備，無法將機器學習廣泛地應用於文本分析上，但這些限制會隨著愈來愈方便使用的遷移學習 (transfer learning) 工具 (例如：BERT) 而解除。

---

<sup>1</sup> SEC Financial Reporting Manual 對 MD&A 有以下要求及目的：

(1) 提供關於公司財務報告之文字敘述，使投資者能從管理階層角度分析公司。

(2) 加強整體財務資訊之揭露，並提供應分析財務資訊之背景。

(3) 提供有關公司盈餘及現金流量之品質和潛在變化之資訊，以便投資者能確信過去績效有代表未來績效之可能性。

資料來源：<https://www.sec.gov/rules/interp/33-8350.htm>

臺灣公司年報並不包含上述之 MD&A，惟年報中的致股東報告書內容與美國 SEC 規範的 MD&A 要求及目的十分相近。根據公開發行公司年報應行記載事項準則第 8 條規定，致股東報告書應包含前一年度之營業結果、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、法規環境及總體經濟環境之影響等。<sup>2</sup> 因此致股東報告書除應就上年度營業計畫實施成果、預算執行情形、財務收支及獲利能力分析、研究發展狀況等予以檢討，作成說明外，亦應說明公司當年度之經營方針、預期銷售數量及其依據及重要之產銷政策。亦即致股東報告書對公司產品開發、技術發展、國際市場開拓等皆有揭露，而這些揭露是公司管理階層對投資人溝通之重要資訊。故本文選擇致股東報告書作為本研究文本分析之標的，探索致股東報告書揭露之資訊是否幫助投資人評估公司未來營運績效。

過去國內有關致股東報告書之相關文獻有限，僅探討致股東報告書文字資訊是否具有資訊內涵（例如，劉妍伶，2011；黃娟娟，2012），亦尚無機器學習方法之應用，本研究可以彌補這方面文獻之不足。本研究以半導體產業為研究對象，主要係因臺灣領先全球半導體晶圓製造業與封裝測試，半導體類股市值在整體櫃買市場占比近四成，並時常受到三大法人關注及投資。而半導體業景氣亦隨著產品需求和趨勢變化而產生起伏，若能從管理階層的角度瞭解發展趨勢或是需求變化，則可以協助投資人預測未來公司績效。據此，本研究旨在介紹與應用最新的自然語言處理工具 BERT，探討 BERT 模型應用於致股東報告書之情緒分類表現、BERT 視覺化和語境測試，以及比較 BERT 與過去傳統文字探勘方法不同之處，並說明其如何克服傳統文字探勘方法的缺點。另外，本研究透過已經完成訓練的 BERT 模型，協助辨別致股東報告書之情緒，並進一步結合財務數據去預測未來盈餘。

本研究以 2011-2018 年國內上市（櫃）公司致股東報告書為研究樣本，實證結果顯示，適當處理致股東報告書的中英夾雜狀況後，能提高 BERT 模型情緒分析的準確率。在交叉驗證下，BERT 模型分類準確率高達 0.86，其表現優於傳統文字探勘方法（準確率大約 0.55），且模型在預測正向語調和負向語調的表現結果相差無幾，優於 Loughran and McDonald (2016) 和 Siano and Wysocki (2021) 的研究結果表現。本研究進一步使用文字資訊結合財務數據以預測公司未來盈餘，以驗證致股東報告書的資訊價值。研究結果顯示，當年致股東報告書情緒與下一年盈餘及盈餘變動並沒有呈現顯著關係。推測原因部分可能是臺灣的致股東報告書與美國的 MD&A 本身資訊內涵有差異，導致中文致股東報告書與下一年盈餘及盈餘變動無顯著關聯。

本研究之可能貢獻有以下三點：第一，率先應用 BERT 於致股東報告書進行情緒分

<sup>2</sup> 資料來源：<https://law.moj.gov.tw/LawClass/LawAll.aspx?pcode=G0400022>

析。過去會計領域分析 MD&A，常採字典法、樸素貝氏演算法 (Naïve Bayes) 或是其他傳統機器學習模型，而這些方法不僅沒有考慮語言結構或假設過於簡單，其準確率可能差強人意，進而導致推論錯誤。目前遷移學習被大量應用於自然語言處理，並且取得相當好的成果。本文實證結果顯示，BERT 同樣能成功應用在中文致股東報告書情緒分析上，且準確率高，因此可以增加未來會計領域研究方法的多樣性，並改善過去傳統文字探勘方法之缺點；第二，分析 BERT 與傳統文字探勘方法的差異，以視覺化和語境測試解釋差異，能夠更直觀地瞭解 BERT 模型相較於其他傳統文字探勘方法的優勢，未來會計研究能透過視覺化文字意涵以做更深入之探討；第三，透過 BERT 預測 2011 年至 2018 年的致股東報告書情緒，結合 Li (2010a) 盈餘預測模型，探討致股東報告書是否具資訊內涵，可作為會計研究的範例。

本研究共有 5 節，除本節前言外，其餘內容安排如下：第 2 節為文獻探討，第 3 節說明研究設計與流程，第 4 節說明實證結果，最後一節為結論以及研究限制。

## 2. 文獻探討

本節文獻探討首先回顧情緒分析於財金與會計領域之應用，第 2.2 節文字探勘方法之比較分析，第 2.3 節介紹自然語言處理方法——BERT，第 2.4 節回顧 MD&A 之資訊價值相關研究。

### 2.1 情緒分析 (sentiment analysis) 於財金與會計領域之應用

有關財金與會計領域之文字資訊，除了發佈於年報之文字資訊，尚包括分析師研究報告、法人說明會、新聞媒體之輿論等不同類型的資料。財會領域之文字探勘研究大致可分為情緒分析、可讀性、風險和競爭程度等。<sup>3</sup> 與本研究較攸關者為情緒分析，故本小節文獻回顧以情緒分析為主。情緒分析旨在探討文本想表達的觀點及態度，其可能是樂觀、悲觀或者是確定、不確定性組合而成，而中立語調可以視為第三種語調，因為大部分詞句既不樂觀也不悲觀。Li (2010a) 分析 MD&A 前瞻性敘述，其隨機選取三萬行句子訓練樸素貝氏演算法，讓模型理解句子的語調是正向、負向、中立或不確定，亦讓模型理解內容的主題，包括營業收入、成本、淨利等十二種類別。經過交叉驗證 (cross validation) 後，語調預測與主題預測分別獲得約 0.67 和 0.63 的準確率。Li (2010a) 進一步測試結果顯示，前瞻

<sup>3</sup> 可讀性 (readability) 主要探討讀者是否能良好理解文本企業年報表達之觀點或內容，例如，Li (2008) 探討企業年報之可讀性，Miller (2010) 分析年報長度及可讀性是否影響股票交易量。風險分析主要探討文本內容與公司面臨不同類型風險之關係，例如，Rogers, Van Buskirk, and Zechman (2011) 分析盈餘宣告語調與訴訟風險之關係。有關競爭環境，Li, Lundholm, and Minnis (2013) 分析年報呈現之競爭情況與公司績效之關係。

性敘述內容語調與未來盈餘及盈餘變動呈正向關聯。Loughran and McDonald (2011) 首先用 Harvard Dictionary (Harvard-IV-4 TagNeg, H4N) 分析 1994-2008 年之 10-K 文本，惟並未發現 H4N 負向詞彙與年報發佈後的股價超額報酬呈現負關聯，作者指出，近四分之三被 H4N 歸類為負向詞彙中，從財務觀點而言，其歸類不應該是負向。Loughran and McDonald (2011) 另外建立代表負向之專門字典，並再重新量化分析，結果發現負向詞彙與年報發佈後的股價超額報酬率、異常交易量和報酬波動性呈顯著關聯。Price, Doran, Peterson, and Bliss (2012) 除了利用 H4N 字典，同時結合自行建立盈餘相關詞彙的字典，以分析法說會資訊，研究結果顯示，法說會語調不但引起市場價格的短期變化，還影響股票交易量的變化。

## 2.2 文字探勘方法之比較分析

文字探勘最基本的問題是如何將非結構化的文件萃取出與研究主題相關的特徵。早期文獻採用人工閱讀的方式，近年文字探勘方法則以電腦計算為基礎，能將隱含於字裡行間的資訊轉換為數值型態，不僅能降低人工成本，亦能大量增加研究效率。以下將介紹不同文字探勘方法及其優缺點，並彙整成表 1。

### 2.2.1 人工判讀

早期有關文本分析之文獻使用大量人工閱讀，以理解文章的情緒或觀點，例如，Bryan (1997) 透過人工閱讀 250 篇 MD&A，Cole and Jones (2004) 閱讀 568 篇零售業 MD&A。此方法之優點為：(1) 方法簡單；(2) 準確率高；(3) 適用於大部分的分析，如情緒分析、文本內容分析或是命名實體識別 (Named Entity Recognition)<sup>4</sup>。人工閱讀主要缺點為：(1) 需要大量人力；(2) 樣本數量少，限制實證結果的可推論性 (generalizability)；(3) 判斷涉及主觀性，因此難以複製。

### 2.2.2 字典法

字典法是基於預設的字典和規則，將文本的詞彙逐一映射至預設的不同種類。字典可以分為通用性及專用，如 Harvard-IV-4 是常用的英文通用辭典，Henry (2008) 和 Loughran and McDonald (2011) 建立的辭典則屬於會計專用的辭典，具有更高的分類準確度。Hoberg and Phillips (2016) 透過字典法將 10-K 有關產品描述映射於向量空間中，並分析產品描述的文字相似度，藉此衡量公司間的競爭程度。字典法的優點有：(1) 方法簡單；(2) 容易應用於不同主題。而其缺點是：(1) 忽略研究主題須具備之先前知識，如在 MD&A 中，多數句子

<sup>4</sup> 其目標為擷取文字資料中指向實體的文字區塊，例如：人名、地名、組織名。資料來源：<https://ckip.iis.sinica.edu.tw/project/ner>



為中立語調，若僅用正負情緒的字典容易造成分類不準確的問題；(2) 詞彙存在多種意義，在不同情境有不同的語意，而字典法沒辦法捕捉此語境問題，會造成判斷不準確；(3) 若沒有透過人工標籤，沒有辦法驗證分類準確率；(4) 對語言依賴性大；(5) 不具學習能力。

2.2.3 機器學習

機器學習本質上是統計算法，具有類似人工智慧的學習能力。而被應用於會計文本分析上的方法包括 N-grams<sup>5</sup>、支撐向量機法 (Support Vector Machine)<sup>6</sup> 及廣為使用之樸素貝氏演算法。樸素貝氏演算法是假設辭彙之間獨立，運用貝氏定理而做成的分類器，其概念是假設一個句子中，每個詞彙都與其他詞彙無關，而由於獨立的性質，故將各個詞彙出現在特定類別的機率相乘，求出該句子出現在特定類別的機率。Antweiler and Frank (2004) 蒐集 Yahoo! 財經頻道及 Raging Bull 上 150 萬則股票貼文，隨機選取 1,000 則貼文訓練樸素貝氏演算法。儘管貼文對股票報酬影響有限，作者仍發現貼文的則數與股票報酬波動有關。Li (2010a) 透過樸素貝氏演算法訓練，發現 FLS 與後續盈餘及盈餘變動呈正相關。Huang, Zang, and Zheng (2014) 利用樸素貝氏演算法分析 1995-2008 年間 S&P 500 的 363,952 份分

表 1 文字探勘方法分析比較

方法類型	優點	缺點	相關文獻
人工判讀	1. 方法簡單 2. 準確率高 3. 適用於大部分的分析	1. 耗費大量人力與時間 2. 樣本數量少 3. 涉及主觀性，難以複製	Bryan (1997); Cole and Jones (2004)
字典法	1. 方法簡單 2. 容易應用於不同主題	1. 忽略研究主題須具備之先前知識 2. 未考慮詞彙存在多種意思 3. 若無人工標籤，無法驗證分類準確率 4. 對語言依賴性大 5. 不具有學習能力	Henry (2008); Loughran and McDonald (2011); Hoberg and Phillips (2016)
機器學習	1. 研究主題不需要有相關的字典 2. 考慮辭彙所處的語境 3. 能衡量模型準確率	1. 難度較大 2. 詞彙之間不一定獨立 3. 監督式學習需要預分類	Antweiler and Frank (2004); Li (2010a); Huang, Zang, and Zheng (2014)

<sup>5</sup> 指文本中連續出現的 n 個語詞。n 元語法模型是基於 (n-1) 馬可夫鏈的一種概率語言模型，通過 n 個語詞出現的概率來推斷語句的結構。資料來源：<https://zh.wikipedia.org/wiki/N%E5%85%83%E8%AF%AD%E6%B3%95>

<sup>6</sup> 是一種在分類與迴歸分析中分析資料的監督式學習模型，資料來源：<https://zh.wikipedia.org/wiki/%E6%94%AF%E6%8C%81%E5%90%91%E9%87%8F%E6%9C%BA>

析師報告，發現分析師報告可提供額外資訊內涵，投資人對負向語句之反應較正向強烈，且該報告對公司後續五年盈餘成長具預測價值。Li (2010b) 指出，樸素貝氏演算法相較字典法擁有下列優點：(1) 研究主題不需要有相關的字典；(2) 會考慮詞彙所處的語境；(3) 能衡量模型準確率。相較字典法，其缺點則為：(1) 難度較字典法大；(2) 詞彙之間不一定獨立；(3) 監督式學習需要預分類。

## 2.3 自然語言處理方法——BERT

因本文將採用 BERT，故此節將先介紹遷移學習，接著探討何謂 BERT 及應用 BERT 之文獻，最後介紹 BERT 如何進行預訓練。

### 2.3.1 遷移學習

Pan and Yang (2010) 指出，不論是機器學習或是深度學習在分類或迴歸的問題中，會有一個常見的假設是訓練資料和測試資料服從相同的分佈及特徵空間 (feature space)。當分佈改變時，大部分模型需要重新蒐集訓練資料去重建一個模型。實際應用中，資料蒐集耗時又複雜，要建置一個大量、高品質標註的數據庫極度困難，故不充足的訓練資料仍無可避免。因此若能減少蒐集資料的需求，將其他領域中的知識遷移至使用者研究之領域，以提高該領域之分類效果，遷移學習的概念因而產生。就會計研究而言，遷移學習可提供一個具「大數據」能力的「預訓練」模型，只需要一小部分的會計專業領域訓練資料，便能成功地將模型「微調」至會計專業領域並解決相關的研究問題。

近期關於自然語言處理領域中，許多預訓練模型相繼提出，如 ELMo (Peter, Neumann, Iyyer, Gardner, Clark, Lee, and Zettlemoyer, 2018)<sup>7</sup>、GPT (Radford, Narasimhan, Salimans, and Sutskever, 2018)<sup>8</sup>、BERT (Devlin, Chang, Lee, and Toutanova, 2019)，在以下的自然語言理解任務如 GLUE (General Language Understanding Evaluation)<sup>9</sup>、SQuAD (Stanford Question Answering Dataset)<sup>10</sup> 及 SWAG (Situations With Adversarial Generations)<sup>11</sup> 表現突出。這些不

<sup>7</sup> ELMo 模型架構採用雙層雙向 LSTM (Long short-term memory)，其預訓練輸入為從左到右及右到左之文本，學習辭彙語意。

<sup>8</sup> GPT 是由 OpenAI 開發，模型設計基於 Google 開發之 Transformer 架構，為全世界參數最多的神經網路模型。

<sup>9</sup> GLUE 是由紐約大學、華盛頓大學和 DeepMind 的團隊提出的數據集，其中包含 11 項常見的自然語言處理的任務，例如：CoLA (The Corpus of Linguistic Acceptability) 是辨認每個句子語法是否正確，或是 SST-2 (The Stanford Sentiment Treebank) 是關於句子的情緒分析。

<sup>10</sup> SQuAD 是由史丹佛大學開發的數據集，其目標為給定一篇文章，準備相關的問題，需要模型給出問題的答案。

<sup>11</sup> SWAG 任務是給定一個陳述句子和其他四個句子，模型判斷哪個句子最具有邏輯性，相當於閱讀理解的問題。

同的模型具有以下特點：(1) 模型結構非常大且深度非常深，如 BERT 參數有差不多 1.1 億個；(2) 用大量的資料預訓練，如 BERT-Chinese 是用中文維基百科預訓練而成的。而這些特點能夠讓模型捕捉到更良好的詞向量 (word embeddings)<sup>12</sup>，例如：同個詞彙其詞向量會隨著語境變化而變化、具有捕捉否定詞以及理解相隔較遠的詞彙意義之特色，可克服字典法、樸素貝氏演算法或其他傳統文字探勘方法的缺點。

上述提到許多其他不同的預訓練模型，本研究擬採 BERT 之主要原因如下：

- (1) BERT 是 Google 開發的免費開源軟體，任何人都可以使用。
- (2) Delvin et al. (2019) 指出，BERT 在不同自然語言處理任務中，表現得更為出色。
- (3) BERT 的操作簡單，且網路有非常多資源。

### 2.3.2 BERT 之應用

近年，BERT 開始應用於財金、會計及社會科學領域，例如 Hiew, Huang, Mou, Li, Wu, and Xu (2019) 以三家在香港交易所 (Hong Kong Stock Exchange) 上市且股票交易熱門的公司為研究對象，蒐集這三家公司的微博 (Weibo) 貼文，訓練五種不同的情緒分析模型，發現 BERT 從中脫穎而出。Elwany, Moore, and Oberoi (2019) 嘗試區分法律文件中協議條款是屬於自動更新還是屬於經過固定期限後會到期的類別，該文利用大量特定領域法律文件微調 BERT 以提高模型分類準確性，結果發現 Precision、Recall 和 F1-score 皆高達九成。Siano and Wysocki (2021) 探討 BERT 應用於季別盈餘報告中之情緒分析表現，該文比較 Loughran and McDonald 的字典法、樸素貝氏演算法、隨機分類及 BERT 的情緒分類模型，發現 BERT 較其他三個方法表現更為出色。該文進一步測試 BERT 是否只依高頻率的單字標記句子的情緒及是否考慮語境問題，結果顯示，儘管刪除高頻率單字會使 BERT 模型分類準確率些微下降，但仍較其他三個模型表現佳，且當句子被打亂時，BERT 表現顯著下降，因此佐證 BERT 考慮上下文，並非只依靠單詞而做判斷。Li, Li, Wang, Jia, and Rui (2020) 蒐集 Eastern Stock Exchange 的股票線上評論，將其分類為正向、負向及中立，並且細分為八種類別，發現 BERT 分類表現較支撐向量機及 LSTM<sup>13</sup> 模型來得更為準確。

### 2.3.3 BERT 訓練結構

BERT 是 Google 以無監督式學習訓練大量文本而建置的模型，再透過使用者以監督

---

<sup>12</sup> 詞向量係將抽象的辭彙之間的語意關係量化成向量形式。

<sup>13</sup> LSTM 是目前遞歸神經網路 (Recurrent neural network) 最常使用的模型，適合處理時間序列的問題，由於文字結構包含先後順序的問題，因此自然語言處理常常使用該模型。



式學習方式應用於不同領域中。其本身的概念是語言模型 (Language Model, LM) 的一種變形，語言模型是給定  $m$  個詞彙組成的句子，去預估下一個詞彙  $w_{m+1}$  出現的機率分佈，其數學形式如公式 (1)：

$$P(w) = P(w_{m+1} | w_1, w_2, w_3, \dots, w_m) \quad (1)$$

BERT 不同於以往訓練語言模型作法，只從句子左邊訓練到右邊或是從句子右邊訓練到左邊，Devlin et al. (2019) 證明 BERT 的雙向訓練語言模型比上述單向訓練的語言模型更能捕捉詞彙順序及上下文關係。BERT 訓練的任務有二：

- (1) 克漏字填空：隨機選取 15% 的辭彙將其遮蔽，模型將透過未遮蔽的詞彙去學習預測被遮蔽部分最有可能出現的辭彙。
- (2) 預測上下文：給定兩個句子，模型將學習預測這兩個句子是否為上下文或是沒有任何關係。

這兩項任務最大好處是其不需要透過標註過的資料訓練（無監督式學習），因此機器能夠自動學習大規模的資料以及詞向量。當預訓練的模型建置好後，研究者可以依不同任務需求，利用其專業領域的數據在 BERT 最後一層微調（監督式學習）。以情緒分析為例，將在 BERT 模型最後一層加上線性分類器進行微調，因此，從頭開始訓練的參數只有屬於線性分類器的參數而已，參見 BERT 訓練流程（圖 1）。

以下定義本研究在訓練 BERT 模型使用的專有名詞。首先介紹何謂過擬合 (overfitting) 以及欠擬合 (underfitting)。過擬合是指當模型過於精確地訓練特定資料集，導致其無法良好地調適其他資料或預測未來的結果，因此模型會在未學習過的資料集上表現不佳；欠擬合是指模型學習能力不足，無法學習到數據集中的「一般規律」，導致其不論在已經學習或未學習的資料集皆表現不佳。上述兩種情形是在機器學習領域中應當避免的情況。

而在涉及類神經網路訓練過程中，模型至少要包含下列超參數之調整：

- (1) 學習率 (learning rate)：係指控制當損失函數梯度下降時更新權重的速度，較高的學習率

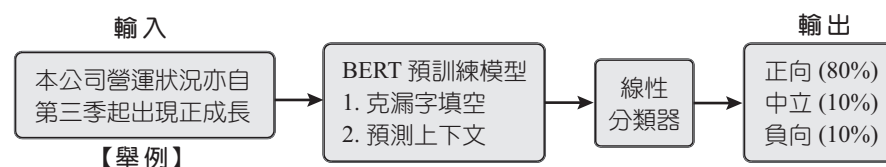


圖 1 BERT 訓練流程

會導致在訓練過程中無法收斂於局部極小值，較低的學習率可避免錯過局部極小值，但其所需要收斂的時間較長，且可能有欠擬合情形；

- (2) Epochs：係指神經網路遍歷一次完整的訓練資料並且返回了一次，太多 Epochs 會花費大量時間且造成模型過擬合，太少 Epochs 則會讓模型有欠擬合的情形；
- (3) Batch Size：係指一次訓練的樣本數目，Batch Size 太大可能導致過擬合情形，Batch Size 太小則可能導致欠擬合情形；
- (4) 有代表性的驗證資料集：驗證資料<sup>14</sup>集扮演的角色，是檢視模型在新數據上的表現，同時透過調整超參數，去選擇在驗證資料集表現最好的模型，最後再透過該模型在測試資料集的表現作為衡量標準。

其中學習率、Epochs 及 Batch Size 調整是為了達到最低的損失和最高的準確率，同時避免過擬合或欠擬合的情況發生。而將驗證資料集特別排除於訓練及測試資料集之外的原因是避免高估模型的準確率，因為若將驗證資料集與測試資料集混合在一起，並透過該資料集的表現去選擇超參數的話，會造成模型結果嚴重高估而失去參考依據。

## 2.4 MD&A 之資訊價值相關研究

由於本研究目的在透過 BERT 分析非結構化的致股東報告書文字內容後，探討該文本是否能提供財務數據以外的資訊，以幫助預測下一年盈餘及盈餘變動，因此以下將彙整研究美國企業 MD&A 文字資訊之資訊內涵相關文獻。

大部分文獻蒐集 MD&A 揭露的訊息，並探討這些訊息是否能幫助投資者更準確地預測公司未來盈餘。例如，Bryan (1997) 以人工閱讀 1990 年 250 篇 MD&A，將 MD&A 揭露的訊息區分為銷售價格、銷售量、營收、成本、流動性、資本支出和未來趨勢七大變數，並分別標記其情緒是正向、中立、負向或沒有揭露。該文發現，出現最多次的情緒類別是中立語調，進一步分析結果顯示，MD&A 有關銷售量變化和未來趨勢之揭露與未來一年的銷貨收入變化方向有顯著正相關；有關未來趨勢之揭露與未來一年的 EPS 變化方向同樣是顯著正相關。然而七大變數中沒有任何一個與未來現金流量變化方向有關，此與 SEC 預期相關揭露能預測營運現金流量的假設並不相符；有關公司財務流動性、已規劃資本支出及未來趨勢之揭露則與公司未來一年的資本支出變化方向有顯著相關。

---

<sup>14</sup> 訓練資料集主要用於模型擬合，直接參與參數訓練的過程；驗證資料集是在訓練過程中，評估模型能力與超參數選擇之依據；測試資料集是評估模型最終能力，因此測試資料集不應作為超參數選擇或特徵選擇之依據。

Cole and Jones (2004) 探討零售業於 MD&A 揭露不同銷貨收入變化之原因 (例如, 來自於同店銷貨收入之成長或營運規模之變化), 是否可預測公司未來營運和資本支出。該文預期同店銷售成長與新店面開幕應與未來銷貨收入成長呈現正相關, 且同店銷售成長可能與未來盈餘和股價呈現正相關, 因其比新店面開幕花更少成本但產生更多收入。經分析 160 家零售業 MD&A 結果支持其假說, 顯示 MD&A 揭露不同銷貨收入變化原因具有不同的資訊價值。

Sun (2010) 探討 MD&A 對存貨變化的解釋與公司未來財務績效是否相關。該文分析 568 篇 MD&A, 並將存貨變化的解釋分為有利、不利及中立。研究結果顯示, 有利的存貨變化解釋與未來的 ROA 和銷貨收入成長皆呈顯著正相關, 而不利的存貨變化解釋則與下一年的 ROA 呈顯著負相關。

Li (2010a) 研究 MD&A 中的前瞻性敘述是否包含公司未來獲利的資訊。該文利用樸素貝氏演算法將前瞻性敘述的句子分類為有利、不利及中立, 並加總為語調變數。研究結果顯示語調與未來盈餘及盈餘變動皆呈顯著正相關, 因此當管理階層於 MD&A 對公司未來表現愈樂觀和正向, 該公司下季盈餘愈高。

Bochkay and Levine (2019) 蒐集 1994 年至 2012 年 10-K 資料, 並利用 Loughran and McDonald (2011) 辭典, 計算詞彙出現的頻率。接著使用 Ridge Regression 篩選前五十個與預測未來盈餘顯著相關的詞彙, 發現超過一半以上的詞彙是負向詞彙, 而正向詞彙只占四分之一, 同時作者還發現若減少伴隨著成本出現, 模型會將減少視為正向。研究結果顯示, 結合財務資訊與 MD&A 文字資訊的模型預測未來一年的 ROE 較只有包含財務資訊的模型更為準確。Bochkay and Levine (2019) 同時發現, MD&A 資訊價值會隨著公司特性而不同, 未來績效起伏較大、未來績效較差、投資人監督較嚴格及較可能發生財務困境等的公司, 其 MD&A 具有更高的資訊內涵。<sup>15</sup>

<sup>15</sup> 國內類似 MD&A 之資訊內涵文獻非常有限, 僅有少數未出版碩博士論文以文本分析探討年報之資訊價值。例如, 石慧妤 (2009) 應用內容分析法將年報分為正面及負面情緒詞彙, 實證結果顯示, 加入文字資訊之盈餘變化預測模型較僅含量化資訊模型有較高之準確度。劉妍伶 (2011) 探討企業致股東報告書是否存在以印象管理為目的之揭露策略, 其以內容分析法對 2007 年度上市 (櫃) 公司致股東報告書進行訊息框架分類與編碼, 研究結果顯示, 管理當局會視營運結果來調整資訊提供的策略, 有利之訊息框架在盈餘宣告時對股票累積異常報酬具額外解釋力, 不利之訊息框架則對累積異常報酬無顯著影響。黃娟娟 (2012) 將年報文件分為致股東報告書、營運狀況、會計師查核報告與財務報表附註四個段落, 探討年報文字揭露各段落內容是否能解釋企業失敗情形, 研究結果顯示, 財務性比率結合致股東報告書之文字資訊, 有助於改善財務報表解釋企業失敗之預警效果。

### 3. 研究設計與流程

#### 3.1 研究流程

首先，本研究選取半導體業上市（櫃）公司之致股東報告書全文進行人工標記，接著拆分資料集，並依據驗證集的表現，選擇最佳模型並將其應用於測試資料集，作為比較基準。最後，將最終模型應用於致股東報告書情緒預測，並結合各年度的財務資訊對隔年的盈餘進行預測。參見圖 2 研究流程圖。

以下將依研究流程順序說明。第 3.2 節介紹何謂監督式學習，說明為什麼需要手動標記致股東報告書的情緒，第 3.3 節分別說明模型選擇相關議題，第 3.4 節介紹如何將 BERT 應用於盈餘預測，最後第 3.5 節說明本文之樣本選取。

#### 3.2 監督式學習

監督式學習係透過模型於訓練資料中獲得或建立一個假設，並依照該假設去預測新的資料，而訓練資料是由輸入物件和標籤化輸出所組成。<sup>16</sup> 本研究第一部分係透過自然語言處

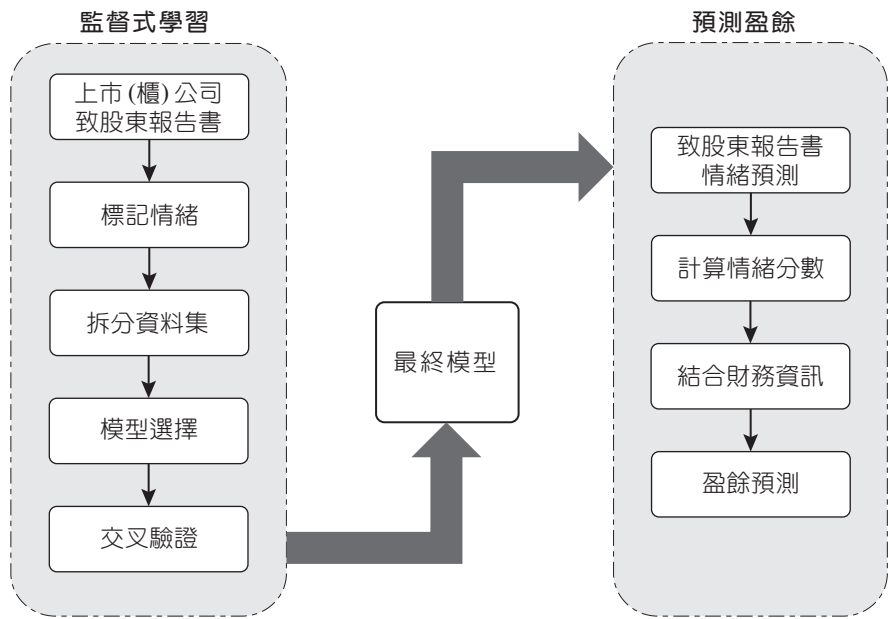


圖 2 研究流程圖

<sup>16</sup> 資料來源：[https://en.wikipedia.org/wiki/Supervised\\_learning](https://en.wikipedia.org/wiki/Supervised_learning)

理預訓練的模型，對致股東報告書進行文字情緒分析，目的是使機器能夠像人一般理解致股東報告書，以克服過往文字探勘方法的缺點，此類型分析屬於監督式學習範疇。

監督式學習通常包含四個主要部分，分別為標記資料、特徵、模型結構以及可區分的訓練及測試資料，以本研究資料為例說明如下。

- (1) 標記資料：標記致股東報告書句子為正向、負向或中立之情緒；
- (2) 特徵：致股東報告書的句子；
- (3) 模型結構：類神經網路；
- (4) 可區分的訓練及測試資料；

其中，標記資料耗費最多時間及人力，因其需要人工判讀文章，以正確地標記句子的情緒。本研究與 Siano and Wysocki (2021) 均採 BERT 模型訓練樣本，不同處是標記資料的方法。本研究係透過人工閱讀致股東報告書標記句子的情緒，但 Siano and Wysocki (2021) 係依據每年同季別的營收變化比率標記，若變化比率大於所有樣本變化比率的中位數，則標記文章為正向，其他則標記為負向，其作法之優點包括：(1) 大量地減少資料標記時間；(2) 取得平衡資料；(3) 仍達到非常高之準確率。

本研究標記之致股東報告書除總結去年營運績效外，包含揭露公司對未來發展的看法，因此即使該年營收成長，但公司管理階層可能對未來抱持悲觀的看法，而在致股東報告書揭露一些負面的消息。因此，本研究優點係人工標記資料較為準確，能捕捉管理階層對下一年營運預測與該年營運結果不同的情況，該狀況如採 Siano and Wysocki (2021) 的作法，可能使實際語調為樂(悲)觀的句子被標記成悲(樂)觀。

### 3.3 模型選擇

BERT 模型利用文字當作特徵以模擬語言結構，透過微調已經被訓練好的 BERT，讓模型能用更少時間、更少資料學習特定領域特性，以達到更好的效能。本研究參考 Siano and Wysocki (2021) 的作法，隨機選取占總樣本數 20% 句子當作驗證資料集 (validation)，並透過該驗證資料集的表現去選擇適當的超參數，而剩下句子會再依 80% / 20% 比例分為子訓練資料集 (sub-training) 和測試資料集 (test) 進行交叉驗證，旨在避免拆分訓練資料集及測試資料集產生之偏誤，導致結果不可靠。圖 3 為模型選擇流程圖。

#### 3.3.1 模型選擇

為了決定將使用哪些學習率、Epochs 數量及 Batch Size 大小，本研究訓練許多不同的模型，並評估其各自在驗證資料集的表現。



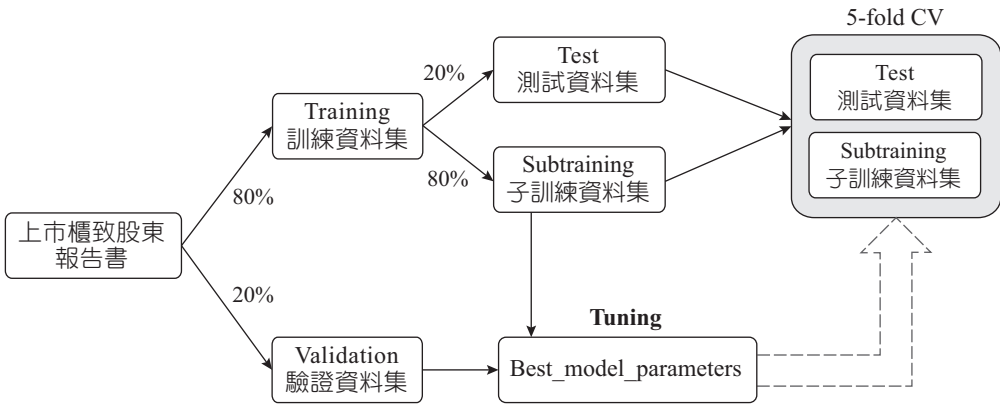


圖 3 模型選擇流程圖

關於 Epochs 數量，由於本研究使用 Early Stopping 的技巧<sup>17</sup>，故先選擇 5 個 Epochs，較 Delvin et al. (2019) 推薦的 Epochs 數量再多一個，接著嘗試 Delvin et al. (2019) 推薦不同的學習率 2e-5、3e-5、5e-5 和其他不同的學習率 2e-4、1e-5、4e-5 和 2e-6；經過測試後，在 1e-5 的學習率下，嘗試 2~5 個的 Epochs；而 Batch Size 則嘗試了 4 個和 8 個。

3.3.2 模型預測力之評估

本研究使用混淆矩陣 (confusion matrix) 評估模型的預測表現 (參見表 2)，衡量指標包括：Accuracy、Precision、Recall 和 F1-Score。

表 2 中，True Positive (TP) 代表模型預測語調與實際語調一致為正向語調；False Positive (FP) 代表實際語調係負向但被模型判斷為正向；False Negative (FN) 代表實際語調係正向但被模型判斷為負向；True Negative (TN) 代表模型預測語調與實際語調一致為負向語調。Accuracy 係衡量被正確地分類樣本 (TP + TN) 占有所有樣本比例，但該指標在不平

表 2 混淆矩陣

	實際為正	實際為負
預測為正	True Positive (TP)	False Positive (FP)
預測為負	False Negative (FN)	True Negative (TN)

<sup>17</sup> 在機器學習中，Early Stopping 是一種在使用梯度下降的疊代優化方法時，可對抗過擬合的正則化方法，在某個節點之前，訓練集使得模型在驗證集的數據上表現得更好；但在該節點之後，更多地訓練會增大誤差。在本研究中是以 validation loss 作為衡量基準。

衡資料情況下，容易造成準確率高估，如：模型全部猜標籤較多的一方，會比隨機猜準確率高，但是模型實際上並未良好地學習。**Precision** 代表預測正向語調，實際上確實是正向語調  $[TP / (TP + FP)]$ 。**Recall** 代表實際為正向語調，模型成功預測為正向語調之比例  $(TP / (TP + FN))$ 。**F1-Score** 則為綜合考量 **Precision** 及 **Recall** 之衡量指標  $[2 \times Precision \times Recall / (Precision + Recall)]$ 。由於本研究係多類別分類問題，因此將 **Precision** 表示為多類別形式，給定一個  $k$  類別的語調，分母為模型根據文字預測語調為  $k$  的數量，分子為實際語調與預測語調同樣為  $k$  的數量，相除後可得模型預測  $k$  類別正確的比例，如等式 (2)：

$$\sum_{i=1}^N \frac{y_{i,k} \cap h(x_i)}{[h(x_i) = k]}, \quad h(x_i) \text{ 是預測值, } y_{i,k} \text{ 是實際為第 } k \text{ 類別} \quad (2)$$

**Recall** 的多類別形式是給定一個  $k$  類別的語調，分母為實際語調為  $k$  的數量，分子為實際語調與預測語調同樣為  $k$  的數量，相除後可得實際為  $k$  類別數量，模型能正確「召回」多少實際為  $k$  類別的數量，如等式 (3)：

$$\sum_{i=1}^N \frac{y_{i,k} \cap h(x_i)}{y_{i,k}}, \quad h(x_i) \text{ 是預測值, } y_{i,k} \text{ 是實際為第 } k \text{ 類別} \quad (3)$$

### 3.3.3 交叉驗證

當透過驗證資料集選出表現最好的超參數後，需要評估該模型在樣本外的測試。為避免拆分訓練集和測試集產生之偏誤，會做 5 折交叉驗證<sup>18</sup>（見圖 3）。首先將驗證集資料外的樣本數隨機拆分成五組資料，選其中一組作為測試資料集，其他四組則作為訓練資料集，因此最終將產生五組模型評估損失和衡量指標，如準確率等，最後再將 5 次的模型評估損失及衡量指標平均，作為該模型表現分數。

### 3.3.4 字典法與 TF-IDF

為了檢驗 BERT 是否優於前文所提及的傳統文字探勘方法，本文利用中研院中文斷詞系統配合字典法衡量句子的情緒，<sup>19</sup> 作為 BERT 交叉驗證的比較基準。首先，樣本與 BERT 樣本皆為同期間，接著透過中文斷詞系統斷詞，並輔以人工方式標記各個詞彙的情緒，

<sup>18</sup> 為什麼使用 5 折而非更多折的原因是模型的訓練速度，當使用 5 折交叉驗證時，需要訓練五次模型，然而基於 BERT 模型非常龐大，導致訓練一次模型費時 50~60 分鐘，若使用更多折數的交叉驗證會使效率大打折扣，因此最後本研究決定使用 5 折交叉驗證。

<sup>19</sup> <http://ckipsvr.iis.sinica.edu.tw/>

正向為 +1、中立為 0 及負向為 -1，作成一情緒辭典。最後，分別使用字典法以及 TF-IDF (Term Frequency-Inverse Document Frequency) 和 BERT 比較情緒分析的準確率。以下介紹本研究如何應用字典法與 TF-IDF：

### 3.3.4.1 字典法

本研究參考 Henry (2008) 的作法，該篇論文計算情緒分數的方式為假設每個詞彙權重皆為 1，並將句子正向詞彙數量減除負向詞彙數量，並除以正向及負向詞彙的總數，即是該句子的情緒分數。由於本研究包含中立的情緒，因此分母部分是除以正向、中立和負向的總數。本研究將正向情緒標記為 +1，中立情緒標記為 0 以及負向情緒標記為 -1，因此將所有辭彙情緒相加即可。計算每行句子的情緒分數後，本研究將句子情緒分數取四分位數，並依照驗證資料集的表現，選取適合的分位數作為區分正向、中立及負向的門檻。本研究最後選擇中位數和第一四分位數作為拆分門檻。

### 3.3.4.2 TF-IDF

本研究參考 Loughran and McDonald (2011) 的作法，該篇論文指出單純計算某一情緒詞彙出現的數量，容易受到句子長度的影響，導致使用計算詞彙數量的方式衡量文字資訊的內容存在缺陷。因此，作者說明考慮權重的作法以及權重帶來的優勢，第一是衡量詞彙的重要性，通常以詞彙出現的比率衡量；第二是對句子長度去規模化，擺脫以往計算數量時，容易受到句子的長度而影響；第三是衡量詞彙在整個文本中的獨特性，例如：「公司」經常出現於不同家公司的致股東報告書，然而並非每間公司營運狀況皆會「上升」，因此「上升」被提及的次數相比於「公司」被提及的次數少，因此 TF-IDF 將判斷「上升」比「公司」更重要。本研究使用 TF-IDF 的定義為考量某一給定詞彙 (i) 在某個句子 (j) 出現的頻率 ( $tf_{i,j}$ )，和包含該詞彙的句子數量 ( $df_i$ )，計算出每個詞彙在各個句子相對應的權重 ( $w_{i,j}$ )，如權重公式 (4) 所示：

$$w_{i,j} = tf_{i,j} \times idf_i \quad (4)$$

權重要素為：

- (1)  $tf_{i,j}$ ：一給定詞彙 (i) 在某個句子 (j) 出現的頻率，代表該詞彙的重要性。
- (2)  $idf_i$ ： $\log\left(\frac{N}{df_i}\right)$ ，N 為句子的數量， $df_i$  為包含該詞彙 i 的句子數量。

本研究使用字典法建置的情緒字典結合 TF-IDF 權重後，將每個句子 (j) 的情緒分數計

算如公式 (5) 所示：

$$\frac{\sum w_{i,j} \times [(i = \text{正向詞彙})] - \sum w_{i,j} \times [(i = \text{負向詞彙})]}{w_{i,j}} \quad (5)$$

最後，與字典法相同，取情緒分數的四分位數，並同樣依據驗證資料集的表現，選取中位數和第一四分位數作為區分正向、中立及負向的門檻。

### 3.3.5 語境測試

本文參考 Siano and Wysocki (2021) 的作法，將測試資料集文字順序打亂，以驗證 BERT 有將語言結構及語境的問題納入考量，這樣的作法預期會大幅降低模型的準確度。

### 3.3.6 BERT 視覺化

由於 BERT 的模型非常複雜，其所學習到的權重很難像一般的線性迴歸或羅吉斯迴歸去解讀在不同特徵下其權重的意義。因此，資料科學家開發視覺化工具來幫助理解深度學習的模型。BERT 最重要的機制為 Attention，當一特定文字產生下一個特徵表示 (representation) 時，會關注哪些文字，當有某一字對該特定文字特別重要時，其 Attention 權重就會比其他文字高，本研究透過線條的深淺代表 Attention 權重的大小。附錄一詳細介紹 BERTViz 的視覺化工具及 Attention 背後原理。最後將於實證結果說明 BERT 視覺化的結果。本研究預期 BERTViz 能夠透過視覺化方式，捕捉到否定詞與其修飾的詞彙，以解決字典法的否定詞解讀問題，例如：「營收未如預期成長」能夠捕捉到「未」是在形容成長，而導致該句話為負向語調；「原物料成本上漲」能夠捕捉到上漲是在形容原物料成本或是「虧損金額減少」能知道虧損減少是正向，而非因為「虧損」、「減少」這兩個負向詞彙，而判斷為負向。

## 3.4 盈餘預測

本研究首先利用驗證資料集所選擇的超參數結合 2011 年至 2017 年致股東報告書句子，重新訓練一個模型，進一步去預測 2011 年至 2018 年的致股東報告書情緒。若模型判斷第 k 個句子語調為中立，則標記為 0，若為正向則標記為 +1；若為負向則標記為 -1。最後將第 i 個公司致股東報告書情緒定義為該公司的所有句子語調加總平均。

$$\text{語調}_i = \frac{1}{K} \sum_{k=1}^K \text{語調}_{i,k}, K = \text{該公司總句子數}$$

接著參考 Li (2010a) 建立盈餘預測模型，<sup>20</sup> 惟本研究與 Li (2010a) 差別有二，一是分類語調的模型，本研究使用 BERT 模型，而 Li (2010a) 是使用樸素貝氏演算法。二是 Li (2010a) 使用的文本是包含於 MD&A 的 FLS，該文本有出現於 10-Q 和 10-K 的財報，因此其預測未來盈餘及盈餘變動是以季度為單位。但臺灣致股東報告書只有出現於年度財務報告中，故本研究預測未來盈餘是以年度為單位。本研究依變數分別為下一年盈餘 ( $t + 1$  年) 及盈餘變動，自變數包含語調及其他控制變數，如：盈餘 ( $t$  年)、股價年報酬率、應計項目、公司市值、股價淨值比、股價報酬率波動、盈餘波動、營運項目數量和成立年數。本研究預期致股東報告書語調與下一年盈餘及盈餘變動呈正相關，代表當管理階層對於公司營運保持樂觀，未來盈餘及盈餘變動較可能會上升；根據盈餘持續性，當前盈餘具有持續影響未來且有正確預測未來盈餘之可能性，因此預期當年度盈餘與下一年盈餘呈正相關。

### 3.5 樣本選取

#### 3.5.1 樣本期間

本研究以臺灣半導體產業為研究樣本，主要是因半導體產業多年來一直由政府有計劃地輔導、推動加上業界自身穩健地經營，產業鏈相當完整。半導體產業鏈包含上游的晶圓和 IC 設計業，中游的 IC 製造業及下游的 IC 封裝業和 IC 測試業，其中 IC 製造業主要以晶圓代工與 DRAM 製造為主。2020 年臺灣半導體產業的總產值占我國 GDP 超過 15% 且占全球半導體產值近 20%，<sup>21</sup> 為臺灣經濟的支柱。半導體業的景氣常隨著產品需求變化而產生興衰，因此若能從管理階層的角度瞭解目前的公司發展趨勢或是未來的需求變化，則可以協助預測未來公司的表現。

表 3 顯示 2011 年至 2019 年半導體業家數雖逐年下降，IC 產業產值卻逐年成長，惟成長率有起伏變化，<sup>22</sup> 2010 年智慧型手機持續帶動晶片市場需求，DRAM 產業也因 PC 的快速成長而需求大增，但 2011 年歷經了日本地震、美國舉債危機導致全球半導體設備與上游原材料供應不穩定，以及取代桌機的趨勢導致 PC 成長力道逐漸疲軟，需求多轉往智慧型手機及平板電腦，使許多公司處於一個青黃不接的時期。而 2012 年至 2014 年智慧型手機數

<sup>20</sup> Li (2010a) 模型中有些變數因未包含於經濟新報資料庫或國內不適用，故未納入本研究實證測試，例如 Fog 指數、未遺失項目數量 (number of non-missing items)、不同部門數量、不同地區數量 (number of geographic segments)、特別項目 (special items) 和公司是否位於德拉瓦州等。

<sup>21</sup> 資料來源：工業技術研究院產業經濟與趨勢研究中心編寫的半導體產業與應用年鑑 <https://money.udn.com/money/story/5612/5082831>

<sup>22</sup> 資料來源：工業技術研究院產業科技國際策略發展所編寫的電子月刊 <https://ieknet.iek.org.tw/book/BookListFree.aspx>



表 3 半導體產業家數、產值及成長率

年份	2011	2012	2013	2014	2015	2016	2017	2018	2019
家數	503	488	478	476	473	448	439	427	422
IC 產業產值 ( 兆 / 新臺幣 )	1.56	1.63	1.89	2.20	2.26	2.45	2.46	2.62	2.64
成長率 (%)	-11.7	4.6	15.6	16.7	2.8	8.2	0.5	6.4	0.9

資料來源：財政部 - 財政統計月報、工業技術研究院 (ISTI)、經濟部工業局智慧電子產業計畫推動辦公室 (SIPO)。

量大幅增長、雲端商機崛起及物聯網應用使經濟回溫。

綜上所述，本研究將樣本期間鎖定於 2011 年至 2019 年，因該期間內有一波景氣循環，能夠讓文本內所傳達情緒較為豐富，且本研究旨在分析 2011 年至 2017 年的致股東報告書之文字資訊，預測 2011 年至 2018 年的致股東報告書情緒，再結合各年的財務資料，測試是否能幫助投資人預估隔年 (2012-2019 年) 公司營運情形。本文為避免 COVID-19 造成的不確定性，導致致股東報告書與實際情形相差太多，樣本並未包含 2020 年的資料。

### 3.5.2 樣本篩選

首先，使用 Python 至公開資訊觀測站下載上市 (櫃) 半導體公司 2011 年至 2018 年年報，總共有 1,203 筆，排除公開資訊觀測站無年報資料、非 PDF 檔案和 Python 爬蟲失敗的年報 336 筆，最後樣本剩 867 筆。其次，閱讀致股東報告書之文字，但不包含表格，一行是以一個句點或者是一個分號作為切割。排除表格之原因，是由於各家公司之致股東報告書表格所揭露之事項不同，有些表格所包含的資訊對於投資人並不是那麼有價值，例如：發展新產品的編號，且表格的文字資訊含量少或是已經揭露於前後文中。<sup>23</sup> 但本研究並未排除句子中之數字，因過去研究發現在自然語言分析中，夾雜在句子裡的數字是具有資訊價值的 (Siano and Wysocki, 2019)。

### 3.5.3 樣本標記

監督式學習為了達到良好的準確率，需要大量資料，即需要大量的人工標記。除了

<sup>23</sup> 本文在閱讀致股東報告書過程中，發現內容包含許多關於產品代號、專有名詞英文縮寫或是日常用語的英文，由於本研究所使用的是 BERT 中文預訓練模型，因此不適合處理太多英文及無意義的產品代碼。首先使用 regular expression 篩選一部分英文字以人工的方式建立字典進行轉換，例如將「DRAM」轉換為「記憶體」、「MCU」轉換為「微控制器」等。接著，透過交叉驗證的表現決定將長度大於 4 且不在字典中的辭彙轉換為「產品」，藉此幫助模型能夠更容易理解句子，並有效降低模型評估損失，提升各項衡量指標。因此於測試時，在輸入的部分嘗試原始文字版本以及經過英文字轉換過後的版本。

作者外，本研究徵求三名會計系研究所學生協助閱讀致股東報告書，以判別語調係屬於正向、負向或中立。<sup>24</sup> 附錄二說明樣本標示之判斷標準釋例。

本研究經前述之資料蒐集、處理和標記後，將這些人工標記句子拆分為訓練、驗證和測試資料集。表 4 顯示，標記後之 2011-2018 年上市 (櫃) 公司致股東報告書共 17,976 句，為完整預測 2018 年之致股東報告書，避免其分散於各資料集，本研究建立資料集時排除 2018 年的 3,140 句。2011-2017 年共 14,836 句，拆分後訓練資料集 (80%) 有 11,868 句，驗證資料集 (20%) 有 2,968 句。本研究再將訓練資料集分為子訓練資料集 (80%) 及測試資料集

表 4 致股東報告書句子拆分為訓練、驗證和測試資料集

	句子數
半導體上市 (櫃) 公司致股東報告書 (2011-2018) (公司年)	17,976
減：2018 年致股東報告書	3,140
半導體上市 (櫃) 公司致股東報告書 (2011-2017)	14,836
減：驗證資料集 (validation) (20%)	2,968
訓練資料集 (training) (80%)	11,868
子訓練資料集 (subtraining) (占訓練資料集 80%)	9,494
－正向	46%
－負向	11%
－中立	43%
測試資料集 (test) (占訓練資料集 20%)	2,374
－正向	47%
－負向	10%
－中立	43%
驗證資料集 (validation) (占總資料集 20%)	2,968
－正向	47%
－負向	10%
－中立	43%

<sup>24</sup> 本研究以 Python PDFplumber 套件將下載之致股東報告書逐句擷取為待標記之文本，每名研究生標記人員分配到已彙整完畢之文本電子檔，然後閱讀完整之致股東報告書，並將每一句子標記正向、負向或中立情緒。為求標記之標準一致，開始標記前標記人員須參加說明會，並熟讀標記規則與注意事項。標記過程作者提供樣本期間已被標記之兩家公司範例供參考，並請標記人員彙整不確定情緒之句子，再與作者討論。最後，作者對所有已標記之樣本進行覆核，以確保樣本標記之品質與可靠性。

(20%)。其中，子訓練資料集被標示為正向、負向及中立之比例分別為 46%、11% 及 43%。測試資料集被標示為正向、負向及中立之比例分別為 47%、10% 及 43%。驗證資料集被標示為正向、負向及中立之比例同測試資料集，分別為 47%、10% 及 43%。

## 4. 實證結果

### 4.1 模型分類結果

#### 4.1.1 模型選擇

表 5 彙整 BERT 模型在驗證資料集的評估結果，表 5 上半部顯示，經過 3 個 epochs 及  $1e-5$  的學習率訓練，Evaluation loss<sup>25</sup> 為 0.37，而 Accuracy 為 0.86，正向的 Precision、Recall、F1-Score 分別為 0.86、0.90、0.88；負向的 Precision、Recall、F1-Score 分別為 0.81、0.87、0.84；而中立的 Precision、Recall、F1-Score 分別為 0.88、0.81、0.85。此結果指出，負向的 Precision 指標 (0.81) 皆低於正向 (0.86) 及中立 (0.88) 的 Precision 指標，而負向的 Recall 指標 (0.87) 低於正向 (0.90) 的 Recall 指標但高於中立 (0.81) 的 Recall 指標。本研究觀察發現，模型較易誤判中立或負向語句的原因可能是未考慮上下文，而將實際為中立或負向語句，做錯誤之判斷。另一可能原因是連接詞語句不易判斷，尤其是轉折語氣不強烈或前半段為負向語句，後半段為正向或中立語句，導致模型預測不精確，使負向 Precision 較低。<sup>26</sup>

表 5 下半部是將致股東報告書原始資料作預處理（如英文字作轉換）後之驗證資料集的表現。在經過 2 個 epochs 及  $1e-5$  的學習率訓練，Evaluation loss 下降至 0.34，而 Accuracy 上升至 0.88。正向的 Precision、Recall、F1-Score 分別為 0.88、0.89、0.89；負向的 Precision、Recall、F1-Score 分別為 0.82、0.91、0.86；中立的 Precision、Recall、F1-Score 分別為 0.89、0.85、0.87。根據驗證集結果顯示，最好的輸入是經過英文字轉換過後的文本，而學習率是  $1e-5$ 、Epochs 為 2 個、Batch Size 為 8 個，並發現當 Epochs 大於二個

<sup>25</sup> Evaluation loss 係指真實值與模型之預測值不一致，透過函數之轉換產生之損失。該差額可以透過各式函數進行轉換。本研究所使用之損失函數係 cross-entropy。由於類神經網路目的係透過最小化損失，以達到良好的訓練，因此較高損失代表其模型分類準確率較低。

<sup>26</sup> 抽取幾個實際為中立，模型卻將其預測為負向的句子。例如，某公司「104 年度營收淨額為新臺幣 12 億 7 千 7 百萬元，稅後淨利約為 8 千 8 百萬元，相較於 103 年度，營收增加近 0.58%，稅後淨利則減少 8 百 8 拾萬元，稅後 EPS 為 2.99 元。」該句話標記實際為中立，因為稅後淨利減少與營收增加相抵，但模型卻將其判斷為負向語調。又另一公司「2016 年通訊市場的成長與 PC 市場的衰退仍舊形成一種互相抵消的狀態，無線通訊營收在智慧型手機與記憶體帶動下，成長 9.6%，可是 PC 與平板電腦的半導體市場卻衰退 8.3%」。該句話標記實際為中立，因為通訊市場的成長與 PC 與平板電腦衰退相抵，但模型卻將此句判斷為負向語調。觀察這些模型判斷錯誤的例子，本研究發現模型容易在轉折語氣不強烈的語句下判斷錯誤。

表 5 BERT 模型於驗證資料集評估結果

驗證資料集結果 (n = 2,968)	Evaluation Loss	Accuracy	Precision			Recall			F1-Score		
			正向	中立	負向	正向	中立	負向	正向	中立	負向
輸入：致股東報告 書原始文字	0.37	0.86	0.86	0.88	0.81	0.90	0.81	0.87	0.88	0.85	0.84
BERT 超參數 -BERT_base_chinese 1e-5 學習率 3 個 epochs											
輸入：經前處理之 致股東報告書文字	0.34	0.88	0.88	0.89	0.82	0.89	0.85	0.91	0.89	0.87	0.86
BERT 超參數 -BERT_base_chinese 1e-5 學習率 2 個 epochs											

或學習率大於 1e-5，模型便開始容易產生過擬合的情形，因此最後選擇讓驗證資料集損失最低的模型超參數進行交叉驗證。

#### 4.1.2 交叉驗證

本研究將測試集資料 (N = 2,374) 經過 5 折交叉驗證後，其平均結果及與其他分析方法之比較如表 6 所示。表 6 顯示，BERT 的 Evaluation loss 為 0.38，Accuracy 為 0.86、

表 6 測試集 5-Fold 交叉驗證平均結果及與其他分析法結果之比較

5-Fold CV (n = 2,374)	Evaluation Loss	Accuracy	Precision			Recall			F1-Score		
			正向	中立	負向	正向	中立	負向	正向	中立	負向
輸入：經前處理之 致股東報告書文字	0.38	0.86	0.82	0.90	0.88	0.93	0.79	0.82	0.86	0.83	0.83
BERT 超參數 - 基於 Validation 所 選擇之最佳超參數											
字典法情緒分析		0.55	0.56	0.52	0.59	0.62	0.46	0.63	0.59	0.49	0.61
TF-IDF 情緒分析		0.54	0.56	0.53	0.53	0.62	0.42	0.75	0.59	0.47	0.62
隨機打亂的致股東 報告書 (語境測試)	0.91	0.60	0.85	0.52	0.76	0.41	0.94	0.12	0.55	0.67	0.21

Precision、Recall、F1-Score 表現均優於其他方法。字典法的結果準確率僅達 0.55，其正向與負向語句的衡量指標僅落於 0.56~0.63 的範圍。TF-IDF 的結果與字典法差異不大，唯一值得關注的是，負向情緒的 Recall (0.75) 顯著較高，此說明 TF-IDF 能夠更準確地辨認負向語句。一般而言，預測負向和中立語句涉及較多人為主觀判斷，因此導致模型預測不準確，故三個方法的正向 F1-Score 指標高於中立與負向的 F1-Score 指標。整體來說，BERT 模型準確率高達 0.86，表現優異。

本研究發現，BERT 模型預測正向語調的 Precision 低於負向語調的 Precision 指標 (表 6)，與 Loughran and McDonald (2016) 及 Siano and Wysocki (2021) 的實證結果相似，但形成原因不同。本研究發現模型經常將中立語句預測為正向語句，與模型選擇情形相同，連接詞語句容易使模型判斷錯誤，亦可能如附錄二提及未來展望情形易導致模型預測錯誤。<sup>27</sup> Siano and Wysocki (2021) 說明 BERT 更容易預測低於所有樣本營收變化率的中位數之公司出具的 MD&A，本研究與其差異可能來自於下列兩個原因：

語調類別數目：Siano and Wysocki (2021) 將句子語調區分為正向語調和負向語調，但在閱讀致股東報告書中，可以發現許多句子其實係屬於中立語調，並不適用於二分法。

標記類別的方法：前述研究設計中，有特別提到 Siano and Wysocki (2021) 之做法可能導致標記與句子本身的意義脫離，進而導致預測不準確的結果。

#### 4.1.3 語境測試

表 6 最後一行顯示語境測試結果。本研究將致股東報告書句子隨機打亂後，進行交叉驗證，得到平均 Evaluation Loss 明顯上升高達 0.91，Accuracy 降為 0.60，中立的 Precision、Recall、F1-Score 分別為 0.52、0.94、0.67；而正向的 Precision、Recall、F1-Score 分別為 0.85、0.41、0.55；負向的 Precision、Recall、F1-Score 分別為 0.76、0.12、0.21，以上顯示 BERT 的預測衡量指標下降許多，此亦顯示，若 BERT 模型如樸素貝氏演算法等未考慮文字順序，則此項測試之各項指標不應隨著文字順序打亂而大幅下降，因此可知，BERT 模型的確有考慮語言結構及語境問題。

<sup>27</sup> 抽取幾個實際為中立，模型預測為正向且屬於未來展望的句子。例一：「後段封測廠之整合，加強與策略代工廠之配合，繼續成本下降之目標，並使用更高階的封裝技術，提高產品價值，創造更高的毛利。」該句話實際標記為中立，因為許多公司每年皆會提出整合工廠、成本下降等喊話但不一定實際執行，但模型卻將其判斷為正向。例二：「(四) 未來公司發展策略：1. 結合客戶的應用需求，開拓產品線之廣度及深度，無論是 3C 通用型產品或車用，工具機等高工作電壓電流應用產品，同時運行開發，使產品多元化以滿足客戶之全方位解決 (TotalSolution) 需求。」該句話實際標記為中立，公司常提到將產品多元化，開發產品線的深度與廣度，但尚未付諸行動或無明確執行，因此將其標記為中立，但模型卻將其判斷為正向。



## 4.2 視覺化及與其他分析方法之比較

### 4.2.1 BERT 視覺化釋例

本節將模型預測正確的語句取出，透過 BERTViz 的套件進行視覺化分析，圖 4 中句子的線條代表 head 在更新其（左側）詞向量時，<sup>28</sup> 關注其他詞彙（右側）的注意力程度。由圖之句子：「雖經濟景氣信號未有明顯起色，經濟環境及競爭態勢仍然險峻。」可以發現，BERT 能辨認「未有明顯起色」，知道該否定詞「未」與「明顯起色」相關聯，能夠將否定詞與其形容的正向形容詞組合一起，理解該句是負向語句，解決字典法中無法辨識否定詞在形容哪個詞彙的難題。

根據圖 5 之句子：「全球半導體市場在 2014 年大幅躍升 9.9% 後，2015 年轉為衰退」，

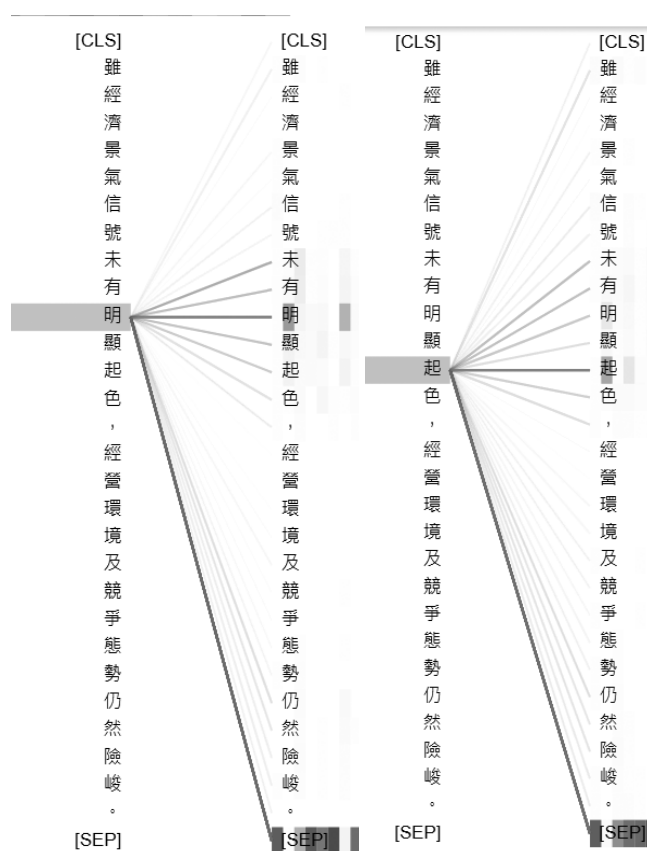


圖 4 BERT 視覺化 (1)

<sup>28</sup> 參考附錄一。

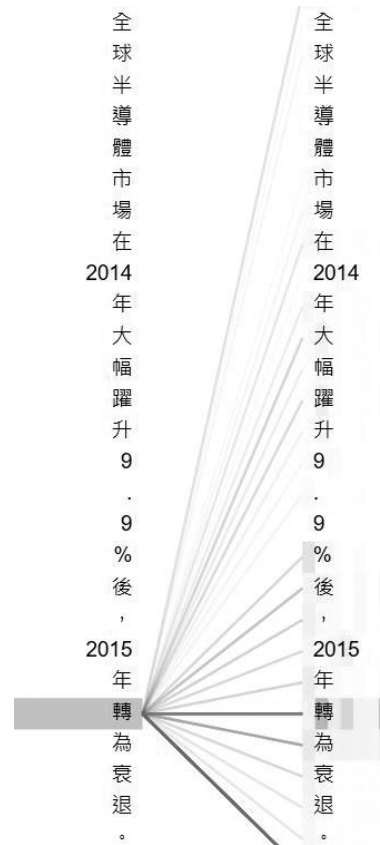


圖 5 BERT 視覺化 (2)

這句話中的「轉」是扮演轉折詞的效果，是形容「大幅躍升」轉為「衰退」，而該句牽涉上下文的結構，如「大幅躍升」及「衰退」對調位置的話，語句情緒便完全相反，而字典法及樸素貝氏演算法是無法理解這樣的上下文轉折的情境，但 BERT 卻能理解這樣複雜的語言結構。

#### 4.2.2 字典法、TF-IDF 與 BERT 之比較

由表 7 範例一可以發現，詞彙之間的修飾能夠成功被 BERT 捕捉，然而字典法卻只依照詞彙意義分辨情緒，容易曲解本身句子傳達的內容，TF-IDF 則更專注於負向的詞彙上，導致其無法正確地判斷「脫離」與「陰霾」的關係。表 7 範例二則顯示，BERT 有捕捉到「並未」這個否定詞是在形容「緊縮」，而「減弱」這個動詞是在形容「不穩定性」，這兩種情況說明句子中時常存在負負得正的情形，而 BERT 能夠成功將詞彙之間的關係串聯，字典法與 TF-IDF 卻需要一些更深入的研究。

表 7 字典法、TF-IDF 與 BERT 之比較 - 範例說明

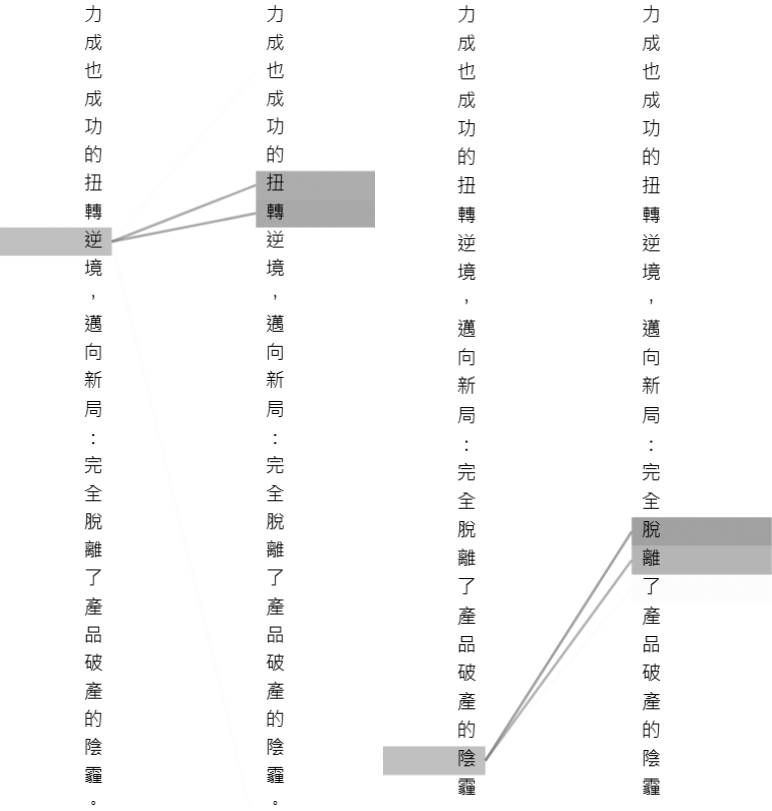
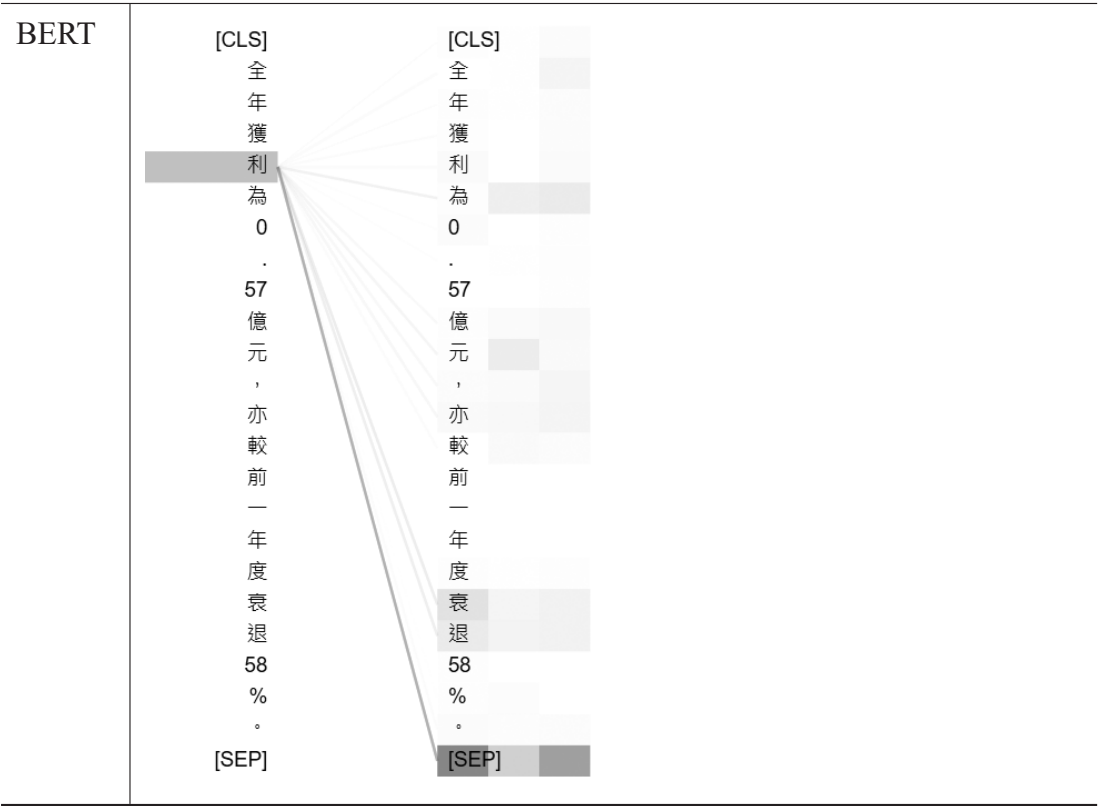
範例一	而 XX 也成功的扭轉逆境，邁向新局：完全脫離了 Elpida 破產的陰霾。
字典法	將「成功」、「扭轉」、「邁向」視為正向詞彙，而「逆境」、「脫離」、「破產」、「陰霾」視為負向詞彙，剩下詞彙皆為中立。由於負向詞彙多於正向詞彙導致情緒分數小於 0.000，因此字典法將其判斷為負向語句。
TF-IDF	由於負向的數量和權重皆大於正向，因此 TF-IDF 將其判斷為負向。
BERT	
範例二	主要國家貨幣政策並未加速緊縮，以及已開發中國家及新興市場國家的政策不穩定性皆有所減弱，是支撐 2017 年全球經濟得以穩健復甦的重要關鍵。
字典法	將「支撐」、「復甦」視為正向詞彙，而「緊縮」、「不穩定性」、「減弱」視為負向詞彙，剩下詞彙皆為中立。此句同樣是源於負向詞彙多於正向詞彙導致字典法將其判斷為負向語句。
TF-IDF	由於負向詞彙的權重大於正向詞彙的權重，因此 TF-IDF 將其判斷為負向。

表 7 字典法、TF-IDF 與 BERT 之比較 - 範例說明 (續)

BERT	
範例三	全年獲利為 0.57 億元，亦較前一年度衰退 58%。
字典法	將「獲利」視為正向詞彙，而「衰退」詞彙視為負向，一正一負，因此字典法判斷為中立。
TF-IDF	由於「衰退」的權重大於「獲利」，因此 TF-IDF 將其判斷為負向。
BERT	

表 7 字典法、TF-IDF 與 BERT 之比較 - 範例說明 (續)



範例三顯示，BERT 在關注「較」此字時，亦注意「全年獲利」和「衰退」，而在關注「獲」與「利」字時，關注「衰退」，顯示 BERT 能辨認「衰退」的主詞是「獲利」，因而判斷為負向；字典法則是因為一正一負的詞彙而判斷為中立；TF-IDF 則是負向詞彙權重大於正向詞彙權重，因此在數量相等情況下，判斷該句為負向。

根據上述的例子，可以發現 BERT 透過複雜的運算邏輯和龐大的參數，能夠將否定詞與其形容的形容詞綁在一起，或是將轉折詞與上下文連結，這些步驟是以往字典法或樸素貝氏演算法做不到的。本研究瀏覽近千份的致股東報告書，觀察到大部分公司前後幾年的致股東報告書表達方式非常相近，甚至有些是相同的句子，該性質有助於模型能夠更精確地學習判斷語句。表 6 中 TF-IDF 造成負向 Recall 上升 (0.75) 可能是致股東報告書負向詞彙數量較少，導致負面詞彙的權重大於正向詞彙的權重。<sup>29</sup> 相較之下，BERT 能夠理解語

<sup>29</sup> 本研究計算正向詞彙權重平均值 (中位數) 為 8.30 (8.62)，而負向詞彙的平均值 (中位數) 則為 8.52 (8.91)，即負向詞彙的權重大於正向詞彙的權重，因此當一句話正向詞彙與負向詞彙數量相差不多時，TF-IDF 將認為負向詞彙較正向詞彙更為重要，進而將其判斷為負向語句。



意結構，判斷否定詞與詞彙的關係，因此能夠更準確地分辨語句的情緒。本研究的交叉驗證結果亦證明 BERT 情緒分析的表現是優於其他兩個方法。

## 4.3 盈餘預測

### 4.3.1 敘述性統計

如前所述，致股東報告書情緒（語調）係指報告書所有句子語調的加總平均，其值介於 +1 於 -1 間。若語調為 +1 (-1)，代表該致股東報告書所有句子皆為正向（反向），亦即，語調愈高，致股東報告書愈正向。

表 8 顯示 2011 年至 2018 年各變數的平均數和分位數。整體而言，這段期間的致股東報告書較樂觀，語調平均數與中位數均為 0.28，且平均數顯著異於 0。股價年報酬率平均為 0.13，中位數為 0.02，顯示樣本期間半導體業股價呈上漲趨勢。

表 8 敘述性統計——盈餘預測之相關變數

變數	平均數	Q1	中位數	Q3	標準差
下一年盈餘	0.04	-0.00	0.06	0.10	0.12
下一年盈餘變動	0.01	-0.02	0.01	0.04	0.10
語調	0.28	0.08	0.28	0.50	0.28
盈餘	0.06	0.03	0.04	0.07	0.06
年報酬率	0.13	-0.21	0.02	0.30	0.58
市值	15.27	14.07	15.19	16.16	1.67
股價波動	0.12	0.08	0.11	0.15	0.07
盈餘波動	0.07	0.03	0.04	0.07	0.07
應計項目	-0.05	-0.10	-0.05	-0.00	0.11
股價淨值比	1.86	1.00	1.35	1.98	1.78
營運項目	1.05	0.69	1.10	1.39	0.39
成立年數	20.08	14.00	19.00	25.00	8.58

註：本表 N = 840。變數定義如下：

語調：2011 年至 2018 年的致股東報告書平均語調。下一年盈餘：下一年 (t + 1) 合併總損益除以平均資產總額。下一年盈餘變動：下一年 (t + 1) 合併總損益減當年合併總損益 (t) 再除以 (t + 1) 年平均資產總額。盈餘：t 年合併總損益除以平均資產總額。股價年報酬率：年報酬率 (%) 除以 100。市值：年底普通股市值做自然對數的轉換。股價波動：樣本期間各年 12 個月的股價月報酬率標準差。盈餘波動：以五年為一窗期，將各年合併總損益除以各自平均資產總額，並取標準差。應計項目：合併總損益減來自營運活動的現金流量，並除以平均資產總額。股價淨值比：每年年底普通股市值加上負債總額並除以平均資產總額計算而得。營運項目：營運項目 (財報) 並加上 1 取 log。成立年數：公司成立年數。前述變數之資料均取自 TEJ 資料庫。

4.3.2 迴歸結果

依據 Pearson correlation 分析，BERT 預測之語調和下一年的盈餘呈現顯著正相關 ( 相關係數 0.3042，P 值為 0.000)。表 9 迴歸結果顯示，致股東報告書語調與下一年盈餘呈負向關係而與下一年盈餘變動有正向關係 ( 係數分別為 -0.0025 與 0.0002)，但不顯著，與 Li (2010a) 結果不同。至於其他控制變數如盈餘、市值和應計項目之結果則與 Li (2010a) 研究結果類似。

本研究並未發現致股東報告書語調與下一年盈餘及盈餘變動呈顯著正關係，分析可能

表 9 語調預測下一年盈餘及盈餘變動之線性迴歸結果——BERT

依變數：下一年盈餘			依變數：下一年盈餘變動	
變數	係數	p 值	係數	p 值
語調	-0.0025	0.836	0.0002	0.985
盈餘	0.5760***	0.000	-0.4557***	0.000
年報酬率	0.0107	0.124	0.0150**	0.031
市值	0.0072***	0.001	0.0082***	0.000
股價波動	0.0642	0.192	0.0136	0.780
盈餘之波動	-0.1396***	0.002	0.0412	0.365
應計項目	-0.0585	0.051	-0.1376***	0.000
股價淨值比	0.0094***	0.000	0.0115***	0.000
營運項目	-0.0086	0.318	-0.0012	0.889
成立年數	-0.0003	0.489	-0.0009**	0.018
2012	0.0309**	0.017	0.0410***	0.002
2013	0.0351***	0.009	0.0384***	0.004
2014	0.0015	0.911	0.0047	0.728
2015	0.0304**	0.015	0.0339***	0.006
2016	0.0254	0.054	0.0281**	0.031
2017	0.0131	0.343	0.0183	0.183
2018	0.0161	0.203	0.0244	0.053
	觀察值	840	觀察值	840
	R-squared	0.55	R-squared	0.35
	Adj. R-squared	0.54	Adj. R-squared	0.34
	F 值	59.02	F 值	26.08

註：樣本從 867 筆降為 840 筆，主要是年報酬率有 21 筆和股價波動有 6 筆資料遺失。依變數下一年盈餘為下一年 (t + 1) 合併總損益除以平均資產總額。依變數下一年盈餘變動為下一年 (t + 1) 合併總損益減當年合併總損益 (t) 再除以 (t + 1) 年平均資產總額。

\*\*\*  $p < 0.01$ ; \*\*  $p < 0.05$ 。

原因如下。本研究首先檢視不顯著原因是否導因於 BERT 預測致股東報告書的語調較不準確。但經比較 BERT 模型與人工判讀 2011-2018 年的致股東報告書的結果，可以發現 BERT 模型準確率高達 0.91。本研究進一步使用人工判讀的語調重新測試，結果如表 10。該結果顯示，語調與下一年盈餘及盈餘變動同樣沒有呈顯著相關。據此，可以排除是因為 BERT 模型的準確率問題導致語調與下一年盈餘及盈餘變動不顯著。本研究推測較可能的原因是，臺灣致股東報告書及美國 MD&A 之資訊含量本質上存有差異（例如投資人結構或資訊透明度等因素），導致語調和下一年盈餘及盈餘變動皆未呈顯著關聯，這方面有待未來研究者進一步深入分析。

表 10 語調預測下一年盈餘及盈餘變動之線性迴歸結果——人工判讀

變數	依變數：下一年盈餘		依變數：下一年盈餘變動	
	係數	p 值	係數	p 值
語調	-0.0032	0.793	-0.0011	0.928
盈餘	0.5762***	0.000	-0.4553***	0.000
年報酬率	0.0108	0.123	0.0150**	0.031
市值	0.0072***	0.001	0.0083***	0.000
股價波動	0.0644	0.191	0.0138	0.777
盈餘之波動	-0.1398***	0.002	0.0409	0.368
應計項目	-0.0583	0.052	-0.1375***	0.000
股價淨值比	0.0094***	0.000	0.0116***	0.000
營運項目	-0.0087	0.317	-0.0012	0.890
成立年數	-0.0003	0.486	-0.0009**	0.018
2012	0.0309**	0.017	0.0410***	0.002
2013	0.0351***	0.009	0.0385***	0.004
2014	0.0017	0.902	0.0049	0.716
2015	0.0304**	0.014	0.0340***	0.006
2016	0.0255	0.053	0.0283**	0.031
2017	0.0132	0.340	0.0184	0.180
2018	0.0162	0.201	0.0245	0.052
	觀察值	840	觀察值	840
	R-squared	0.55	R-squared	0.35
	Adj. R-squared	0.54	Adj. R-squared	0.34
	F 值	59.02	F 值	26.08

註：樣本從 867 筆降為 840 筆，主要是年報酬率有 21 筆和股價波動有 6 筆資料遺失。依變數下一年盈餘為下一年 (t + 1) 合併總損益除以平均資產總額。依變數下一年盈餘變動為下一年 (t + 1) 合併總損益減除當年合併總損益 (t) 再除以 (t + 1) 年平均資產總額。

\*\*\* p < 0.01; \*\* p < 0.05。

## 5. 研究結論與限制

過往會計領域的文字探勘研究常以人工判讀辨別文章的情緒，此蒐集方式耗時且樣本數量少。隨著電腦運算資源增加，文字探勘方法開始出現，例如字典法、樸素貝氏演算法和支撐向量機，但這些文字探勘方法通常限制很多，例如：字典法僅考量相關情緒詞彙出現的頻率，因而難以辨別否定詞是在修飾哪個詞彙；樸素貝氏演算法假設詞彙之間具獨立性，未考慮詞彙的順序性，亦忽略語言的結構性。近年隨科技進步，自然語言處理開始嶄露頭角，例如，2019 年 BERT 問世，於會計資訊的應用上，能透過少量文本資料，結合原本已經訓練完畢的大數據，訓練出專屬會計的大數據模型。

本研究利用 BERT 模型對臺灣上市（櫃）公司致股東報告書進行情緒分析，並依致股東報告書的狀況做微調，最後，在交叉驗證下，本研究 BERT 模型準確率高達 0.86，優於字典法與 TF-IDF；進一步預測 2011 年至 2018 年的致股東報告書，準確率也同樣高達 0.91。本研究透過視覺化和語境測試，檢驗 BERT 是否可成功地克服過往文字探勘法之缺點。結果顯示，BERT 不但成功捕捉否定詞扮演的角色，且同時能夠捕捉轉折詞的效果；BERT 同樣有考慮語言的結構，當文字順序隨機打亂以後，BERT 模型的預測能力大幅下滑。

最後，本研究利用 2011 年至 2017 年資料訓練的 BERT 模型，去預測 2011 年至 2018 年的致股東報告書，進而去預測下一年 ( $t + 1$ ) 的盈餘及盈餘變動，作為應用 BERT 模型的範例。雖然實證結果顯示，語調與下一年盈餘及盈餘變動皆未呈現顯著正相關，分析其因並非來自 BERT 模型預測的不準確，而可能是臺灣致股東報告書與美國 MD&A 相較，對未來盈餘之資訊內涵較不顯著。本文認為未來會計研究可朝著兩個方向更深入鑽研，一是模型訓練的方式，由於財報出現的詞彙較為限縮，若使用詞彙的方式訓練 BERT 是否取得更好的成果，或是將標記方式改變以訓練 BERT，觀察此更改是否改善預測結果；二是將 BERT 應用於不同的會計主題，例如：偵測財報舞弊、分析查核意見是否受新聞媒體影響等等，觀察其在不同主題上表現是否有差異。

本研究在情緒分析上有幾個限制。一是依賴人工標記 2011 年至 2017 年致股東報告書的語調情緒，此仰賴專業判斷，惟也涉及主觀。本文附註 24 及附錄二說明本研究如何降低標記判斷之不一致。本文亦隨機抽樣已標記之致股東報告書大約 200（占總句數約 2%）句，交給其他研究生（未受過標記訓練者）進行複核，複核結果只有 4 句話（占抽樣筆數的 2%）有疑慮。

限制二是硬體資源不足，BERT 有分為 12 層及 24 層的不同模型結構，越高的層數代表模型能夠達到越準確特徵表示，然而此代表其所消耗的運算資源越大。由於本研究使用 Google Colab 提供的 GPU 資源，無法成功嘗試較大的 BERT 結構，因此僅選擇使用 12 層結構。

最後是產業的限制，本研究僅分析半導體產業，因此關於致股東報告書語調的研究結果僅適用於半導體產業，是否適用其他產業，有待進一步之研究。

## 參考文獻

- 石慧好，2009，**應用企業年報中文字資訊於盈餘預測之研究**，國立臺灣大學會計學研究所碩士論文，臺北，臺灣。取自華藝線上圖書館 <https://www.airitilibrary.com/Publication/alDetailedMesh1?DocID=U0001-1106200914214200>
- 黃娟娟，2012，**公司年報文字探勘與財務預警資訊內涵**，私立逢甲大學商學研究所博士論文，臺中，臺灣。取自臺灣博碩士論文系統 <https://hdl.handle.net/11296/2chtfp>
- 劉妍伶，2011，**年報文字資訊訊息框架效果之研究**，國立臺灣大學會計學研究所碩士論文，臺北，臺灣。取自臺灣博碩士論文系統 <https://hdl.handle.net/11296/6ntke5>
- Antweiler, W., and Frank, M. 2004 . “Is all that talk just noise? The information content of internet stock message boards.” *The Journal of Finance* 59 (3): 1259-1294.
- Bochkay, K., & Levine, C. B. 2019 . “Using MD&A to improve earnings forecasts.” *Journal of Accounting, Auditing & Finance* 34 (3): 458-482. doi:10.1177/0148558X17722919.
- Bryan, S. H. 1997. “Incremental information content of required disclosures contained in management discussion and analysis.” *The Accounting Review* 72 (2): 285-301.
- Cole, C. J., and Jones, C. L. 2004 . “The usefulness of MD&A disclosures in the retail industry.” *Journal of Accounting, Auditing & Finance* 19 (4): 361-388.
- Delvin, J., Chang, M. W., Lee, K., and Toutanova, K. 2019. “BERT: Pre-training of deep bidirectional transformer for language understanding.” arXiv:1810.04805.
- Elwany, E., Moore, D., and Oberoi. G. 2019 . “BERT goes to law school: Quantifying the competitive advantage of access to large legal corpora in contract understanding.” In workshop on document intelligence at NeurIPS 2019, Vancouver, Canada.
- Henry, E. 2008. “Are investors influenced by how earnings press releases are written?” *The Journal of Business Communication* 45 (4): 363-407. doi:10.1177/0021943608319388
- Hiew, J. Z. G., Huang, X., Mou, H., Li, D., Wu, Q., and Xu, Y. 2019. “Bert-based financial sentiment index and LSTM-based stock return predictability.” Submitted to NeurIPS 2019, under review. arXiv:1906.09024.
- Hoberg, G., and Phillips, G. 2016. “Text-based network industries and endogenous product differentiation.” *Journal of Political Economy* 124 (5): 1423-1465.

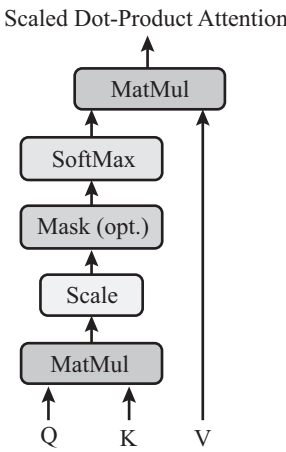


- Huang, A., Zang, A. & Zheng, R. 2014. "Evidence on the information content of text in analyst reports." *The Accounting Review* 89: 2151-2180. doi:10.2308/accr-50833.
- Li, F. 2008. "Annual report readability, current earnings, and earnings persistence." *Journal of Accounting and Economics* 45 (2-3): 221-247.
- Li, F. 2010a. "The information content of forward-looking statements in corporate filings—A naive Bayesian machine learning approach." *Journal of Accounting Research* 48: 1049-1102.
- Li, F. (2010b). "Textual analysis of corporate disclosures: A survey of the literature." *Journal of Accounting Literature* 29: 143-165.
- Li, F., Lundholm, R. J., and Minnis, M. (2013). "A measure of competition based on 10-K filings." *Journal of Accounting Research* 51 (2): 399-436.
- Li, M., Li, W., Wang, F., Jia, X, and Rui, G. 2020. "Applying BERT to analyze investor sentiment in stock market." *Neural Computing & Applications*. doi:10.1007/s00521-020-05411-7
- Loughran, T., and McDonald, B. 2011. "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks." *Journal of Finance* 66 (1): 35-65.
- Loughran, T., and McDonald, B. 2016. "Textual analysis in accounting and finance: A survey." *Journal of Accounting Research* 54 (4): 1187-1230.
- Miller, B. P. 2010. "The effects of reporting complexity on small and large investor trading." *The Accounting Review* 85: 2107-2143.
- Pan, S. J. and Yang, Q. 2010. "A survey on transfer learning." *IEEE Transactions on Knowledge and Data Engineering* 22 (10): 1345-1359. doi: 10.1109/TKDE.2009.191.
- Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., and Zettlemoyer, L. 2018. "Deep contextualized word representations." *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* 1: 2227-2237, New Orleans.
- Price, S. M., Doran, J. S., Peterson, D. R., and Bliss, B. A. 2012. "Earnings conference calls and stock returns: The incremental informativeness of textual tone." *Journal of Banking & Finance* 36 (4): 992-1011.
- Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. 2018. "Improving language understanding by generative pre-training." Retrieved from <https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf>
- Rogers, L. J., Buskirk A. V., and Zechman, S. L. C. 2011. "Disclosure tone and shareholder litigation." *The Accounting Review* 86 (6): 2155-2183.

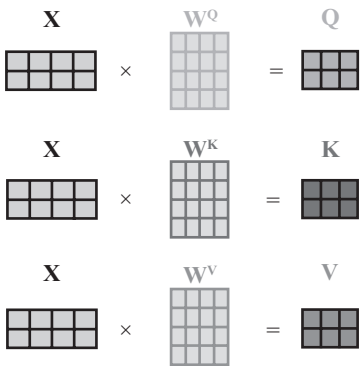
- Securities and Exchange Commission (SEC). 1987. "Concept release on management's discussion and analysis of financial condition and results of operations." Securities Act Release No. 6711. Washington, D.C.: SEC.
- Securities and Exchange Commission (SEC). 2003. "Interpretation: commission guidance regarding management's discussion and analysis of financial condition and results of operations." Securities Act Release No. 8350. Washington, D.C.: SEC.
- Siano, F., and Wysocki, P. 2019. "The primacy of numbers in financial and accounting disclosures: Implications for textual analysis research." Working paper, Boston University.
- Siano, F., and Wysocki, P. 2021. "Transfer learning and textual analysis of accounting disclosures: Applying big data methods to small(er) data sets." *Accounting Horizons* 35 (3): 217-244. doi: 10.2139/ssrn.3560355
- Sun, Y. 2010. "Do MD&A disclosures help users interpret disproportionate inventory increase?" *The Accounting Review* 85 (4): 1411-1440.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. 2017. "Attention is all you need." *In Advances in Neural Information Processing Systems*, 6000-6010.

# 附錄一 BERTViz

附錄一說明 BERTViz 運作的機制，首先要先介紹的是 BERT 模型使用的 Self-Attention，從 Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser, and Polosukhin (2017) 可以看到其結構如下圖。



其中 Q 代表 Query；K 代表 Key；V 代表 Value，這三個矩陣皆與輸入相對應，而 MatMul 代表矩陣相乘，Mask 代表隨機遮蔽，SoftMax 為歸一化的指數函式。下圖<sup>30</sup>將說明 BERT 是如何計算其 Attention。



首先，BERT 會將文字 X 透過不同的權重 W 對應到矩陣 Q、K 和 V，接著計算矩陣 Q 及矩陣 K 的內積，並除以  $\sqrt{d_k}$ ， $d_k$  為 K 矩陣的維度，目的是讓 gradient 更加穩定。接著再

<sup>30</sup> 資料來源：<https://jalammar.github.io/illustrated-transformer/>

利用 SoftMax 將計算結果歸一化，最後乘以各自的矩陣 V 得到加權向量並加總起來。其數學式表示如下：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right) V$$

下表<sup>31</sup>將上述的步驟彙整

Attention 計算步驟

輸入	Thinking	Machine
詞向量	$x_1$	$x_2$
Q	$q_1$	$q_2$
K	$k_1$	$k_2$
V	$v_1$	$v_2$
分數	$q_1 \cdot k_1 = 112$	$q_1 \cdot k_2 = 96$
除以 $8(\sqrt{d_k})$	14	12
SoftMax	0.88	0.12
SoftMax*V	$v_1^*$	$v_2^*$
加總	$z_1 = v_1^* + v_2^*$	

上述都是一個 Attention head 的情形，而 Vaswani et al. (2017) 論文中有使用多個 Self-Attention 的表示，稱為 Multi-Head Attention，其方式為連接多個 Attentions，其數學表示如下：

$$\begin{aligned} \text{MultiHead}(Q, K, V) &= \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) W^O, \text{ where } \text{head}_i \\ &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned}$$

Multi-Head Attention 意思為有多個 Self-Attention，因此對應到不同的  $Q_i$ 、 $K_i$  及  $V_i$ , for  $i = 1, 2, \dots, 12$ ，最後產出不同的  $Z_i$ , for  $i = 1, 2, \dots, 12$ ，並進一步將其連接起來乘以權重，得到最後的  $Z^*$  矩陣。而 BERTViz 便是將 Multi-head Attention 視覺化出來，線條代表 head 在更新詞向量（左側）時，關注其他詞彙的注意力程度（右側），因此連接的線條越粗代表在更新該詞向量時，對應的詞彙對該詞向量影響非常大。12 個 Multi-Head 分別對應到不同顏色的線條。

<sup>31</sup> 參考 <https://jalamar.github.io/illustrated-transformer/> 的表示

## 附錄二 致股東報告書標記之判斷標準釋例

以下分五種類別舉例說明之：

### 1. 當年度營業狀況

營業狀況表達大致上分成兩種，一種係「本集團民國 XXX 年度全年營業收入淨額為新臺幣 X,XXX,XXX 元，較前一年的 X,XXX,XXX 元增加 XX%；」，屬於有前後年比較的情況，這種句子會給予正向或負向之語調；而另一種為「本公司 XXX 年之營業收入為 X,XXX,XXX 元，稅後淨利為 X,XXX,XXX 元，稅後每股盈餘為 X.XX 元。」，沒有前後年之比較，只有單純敘述者，會給予中立之語調。主要原因係不知道該公司是進步、退步還是持平，故給予其中立之標籤。

### 2. 連接詞

連接詞代表轉折關係，前面負向語句連接後面正向語句，反之亦然。在某些情形下，後半段比前半段更具有強調意味，在此情形下，情緒標籤將以後半段語句為主。倘若後半段語句較空泛，則會給予中立之語調。例如：「雖然美中貿易與關稅的紛爭，造成政治及經濟上諸多不確定因素的影響，使得營運倍受挑戰，然而 XXX 自下半年庫存開始去化的驅動下，營收與毛利率皆較 XXX 年明顯成長與提升」，此句標記正向。第二例：「展望 XXX 年的半導體銷售的成長預估 XX% 達到 X,XXX 億美元的規模，可望重新回到成長的周期，惟 XXXX 年全球景氣依舊籠罩在中美貿易的不確定性，以及新冠肺炎疫情擴大，對電子業和半導體業供應鏈帶來挑戰，供貨和需求皆同步受到影響，也為 XXXX 年的復甦埋下不確定的因素」，此句標為負向。第三例：「XXXX 年度是競爭激烈的一年，近年來國內封裝大廠持續擴廠，進而影響中小型封裝業之業績成長但經營團隊仍努力維持既有產品及彈性的製造排程，以市場區隔的方式，不斷創新研發新產品」，此句標為中立。

### 3. 研究發展

有關研究發展，公司時常揭露「持續研發產品先進製程技術，增加產品競爭力」或是「持續降低生產成本」類型的空泛句子，則標記為中立語調，原因係觀察這幾年的致股東報告書，發現有許多類似句子，這樣的語句對於投資者沒有實質的參考價值。本研究資料集排除「開發 XXX 產品」等類似語句，主要考量有些公司產品名稱太過專業，投資者難以瞭解該產品是否有市場或效益；其次，通常這類句子常常列點出現，導致致股東報告書有四、五行是這種句子，而這類句子同樣對未來公司營運預測沒有太大的實質幫助，因此盡量將這類型的句子排除在外。



#### 4. 未來展望

有關未來展望，有些公司會於未來營業計劃或未來展望述說公司的規劃，然而如研究發展狀況一樣，不免會出現「拓展營業接單策略，積極開發新客戶」、「持續創新產品開發，提供多元式產品封裝製程，滿足客戶需求」等較為空泛的精神喊話，本研究對這類句子處理方式與研究發展一樣，皆標記中立語調。

#### 5. 無關

致股東報告書有時會出現「成功者找方法」、「只為成功找方法，不為失敗找藉口」等董事長信心喊話，抑或是「由衷感謝股東的長期愛護肯定、客戶的認同信任及供應商的全力支持」等感謝的話，本研究將這類型的句子標記為中立。

# Application of Transfer Learning and Text Mining on Reports to Shareholders

Yu-Te Chen<sup>1</sup> Chan-Jane Lin<sup>2</sup>

<sup>1</sup> Shopee

<sup>2</sup> Department of Accounting, National Taiwan University

Corresponding author: Chan-Jane Lin

Address: No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan (R.O.C.)

E-mail: cjlin@ntu.edu.tw

Received: July 19, 2021; After 2 rounds of review, Accepted: March 28, 2021

## Abstract

First, this study applies BERT (Bidirectional Encoder Representations from Transformers) on Reports to Shareholders (RTS) of the semi-conductor industry in Taiwan. Next, we discuss whether BERT can overcome some weaknesses of traditional text mining tools. Finally, we try to assess the association between the tone in RTS with company's future financial performance by using sentiment analysis. The empirical result shows that BERT classification accuracy reaches as high as 0.86, which outperforms other techniques. Moreover, by visualizing the operation in BERT, we find that BERT can capture the word association successfully. However, the empirical result fails to show that the sentiment in RTS has significant and positive association with next year's earnings and change in earnings, which is inconsistent with previous findings. We conjecture that it may be caused by the capital market attributes such as investor's structure or information transparency in Taiwan, resulting in differences in information content provided by RTS and MD&A.

**Keywords:** Transfer Learning, Sentiment Analysis, Earnings Prediction, Reports to Shareholders

---

The authors acknowledge the helpful comments of the field editor and two anonymous reviewers, and take sole responsibility for their views.

Data availability: Data used in this study are available from public sources identified in the study.



東華書局  
Tung Hua Book Co., Ltd.

## 1. Research Issues

Recently, Siano and Wysocki (2021) introduce and apply the state-of-the-art transfer learning model BERT (Bidirectional Encoder Representations from Transformers) to analyze corporate disclosures. BERT is designed to mitigate the limitations of prior textual analysis research, such as small sample size, the inability to deal with negations of stand-alone words, and the long-range associations of words. Siano and Wysocki (2021) find a higher classification accuracy in a sentiment analysis of Management Discussion and Analysis (MD&A) than in traditional text-mining techniques, such as the word-list approach and the Naïve Bayes Classifier.

Motivated by Siano and Wysocki (2021), this paper attempts to apply BERT to conduct a Chinese sentiment analysis of Reports to Shareholders (RTS) in the annual report of listed companies in Taiwan, which is similar to an MD&A for US companies. According to regulations governing information to be published in annual reports of public companies in Taiwan, RTS shall include the operating results of the preceding year, a summary of the business plan for the current year, the company's future development strategy, and the impacts of external competition, the legal environment, and the overall business environment. The disclosures in RTS, such as product or technology development and business prospects, are regarded as important information for managers to communicate to investors. We first define and explain the concept of transfer learning and machine learning. We then train the BERT language model with RTS, visualize how the BERT language model works, and compare BERT with other traditional text-mining techniques. Lastly, following Li (2010a), we investigate whether there exists a positive association between the tone in RTS and the company's future financial performance.

## 2. Research Hypothesis

This paper explores whether the BERT model performs better in Chinese textual analysis as well. Analyzing RTS in Taiwan and comparing BERT with dictionary and TF-IDF (Term Frequency-Inverse Document Frequency) methods, we expect BERT to outperform other text-mining techniques in terms of accuracy classification. Moreover, by learning more complicated word representations, we expect that BERT can show the relation between a word and its negation or show associations between adjectives and nouns.

This paper also examines whether the tone in RTS contains information about future profitability. Since RTS contains information not only for prior year operating results but also for future development strategies, it is expected to provide information content for a company's future operating performance. As suggested by Li (2010a), the empirical tests in this paper are indeed a joint test of the machine learning tool's ability to capture tone in RTS and to discern whether management follows regulations and provides forward-looking information, and

whether management's disclosure is truthful. If BERT is able to capture management tone, then the evidence that RTS tone predicts future earnings is consistent with the hypothesis that management is truthfully disclosing information in RTS. The results from Li (2010a) show that the average tone of forward-looking statements in a firm's MD&A is positively correlated with its future earnings. If the results fail to find the association between the tone in RTS and future profitability, we cannot reject the hypothesis that RTS have no information content because the result could be due to the low-power BERT or other factors.

### 3. Research Methods

We start with a sample of publicly traded companies in the semiconductor industry in Taiwan from 2011 to 2019. All financial and textual data used in this study are obtained from the Taiwan Economic Journal (TEJ) database and the Taiwan Stock Exchange (TWSE) database, respectively. We choose the semiconductor industry as our sample because Taiwan owns the most complete and leading semiconductor industry chains in the world. Furthermore, the market value of this industry accounts for 40% of total market value on the TWSE and the Taipei Exchange. After excluding missing reports and image files, we obtain 867 firm-year observations of RTS from 2011–2018 and select all the sentences in the RTS for our analyses. Finally, we collect 14,836 sentences, excluding 3,140 sentences in 2018, and split the sentences into a sub-training dataset (9,494 observations), a validation dataset (2,968 observations), and a test dataset (2,374 observations) to be used for initial model training.

After training the model, we use the validation dataset to choose the best parameters and assess the classification accuracy by using an out-of-sample set of testing observations. To alleviate the concern that test results are driven by the random choice of training and test sample, we apply a cross-validation technique on the evaluation process.

With the aim of measuring BERT comprehension, we shuffle the words in the sentences so that if word order plays an important role in BERT, then shuffling the words should impair the accuracy. Furthermore, we visualize how BERT works in training, which we expect to find the relation between the word and its adjective. Finally, we use our final model to predict the tone and combine the tone with other financial data, i.e., accruals, return on stock, market-to-book, etc., to predict earnings and the change in next-year earnings, i.e., the year 2019.

### 4. Results

Examining a sample of companies in the semiconductor industry in Taiwan, we find that, consistent with our expectations, BERT's classification accuracy reaches as high as 86%, which outperforms the result of dictionary (55%) and TF-IDF (54%) methods. Moreover, this paper

shows that the classification accuracy significantly reduces to 60%, after shuffling the word order, indicating that BERT is able to consider word order and context in the training process. Through visualizing the operations of BERT, we find that BERT understands the relation between a word and its negation and also knows the role of conjunctions, which cannot be understood by traditional text-mining techniques. However, the empirical result of earnings prediction fails to show that the tone in RTS has a significant and positive relation with earnings and earnings change one year ahead, which is inconsistent with the previous findings using US companies' MD&A. We conjecture that it may be caused by the different capital market attributes, such as investor's structure or information transparency in Taiwan, resulting in differences in information content provided by RTS and the MD&A.

Although the earnings prediction is inconsistent with our expectation, this paper contributes to accounting research by first applying a deep-learning model on Taiwan's RTS that proved able to overcome the limitations of traditional text-mining techniques in prior research. Additionally, we have shown how BERT works inside the black box, so future research can study word associations more deeply. Importantly, we create an example for future accounting research by combining a transfer learning model with the earnings prediction model to evaluate whether RTS contain information content.

