金融科技與機器學習

財務文字資料分析

作者：王釧茹

作者簡介：王釧茹是中央研究院資訊科技創新研究中心的副研究員，研究領域是計算金融與資料分析。除研究工作外，她還致力開發可用於檢驗擬議方法的實用性和通過數據可視化技術呈現結果的互動式系統。

（rawpixel）

近年來，由於財務相關資料的大量累積，如何有效率地從中發現有用的訊息並用以進行更精準且有效的財務決策，遂成爲一個重要的研究議題。這些訊息不僅可協助相關從業人員（如：會計師、分析師）掌握投資機會、最大程度地降低風險或控制成本，更可以運用至各種不同的金融場景。

本文將介紹文字分析（text mining）及自然語言處理（natural language processing）技術於非結構化財務文字資料上的相關技術及應用。全文分成若干子題，前半部介紹不同類型之財務資料及財務文字分析目前在文獻上的主要做法，後半部則帶入特定應用及其機器學習模型之相關探討，最後將討論此跨領域研究之挑戰與未來可能之研究方向。

財務資料：硬訊息（hard information）及軟訊息（soft information）

由於資料分析的盛行，近年來，在金融和計算機科學領域進行了諸多有關財務資料分析及預測的研究。在財務領域中，一般將資訊分爲硬訊息（hard

information）及軟訊息（soft information）兩大類[1]，前者通常指的是數字，如：財務指標和歷史價格；後者通常指文字訊息，如：財經新聞及財務報告之文字資訊、市場評論等。過往文獻——尤其是在傳統財務或會計領域——大部分針對硬訊息進行建模或分析，其主要原因有三：

1. 文字資料未被系統化收集，導致資料量不足；
2. 非結構化文字處理較爲困難且需要更多運算資源；
3. 相關技術和硬體尚未發展成熟。

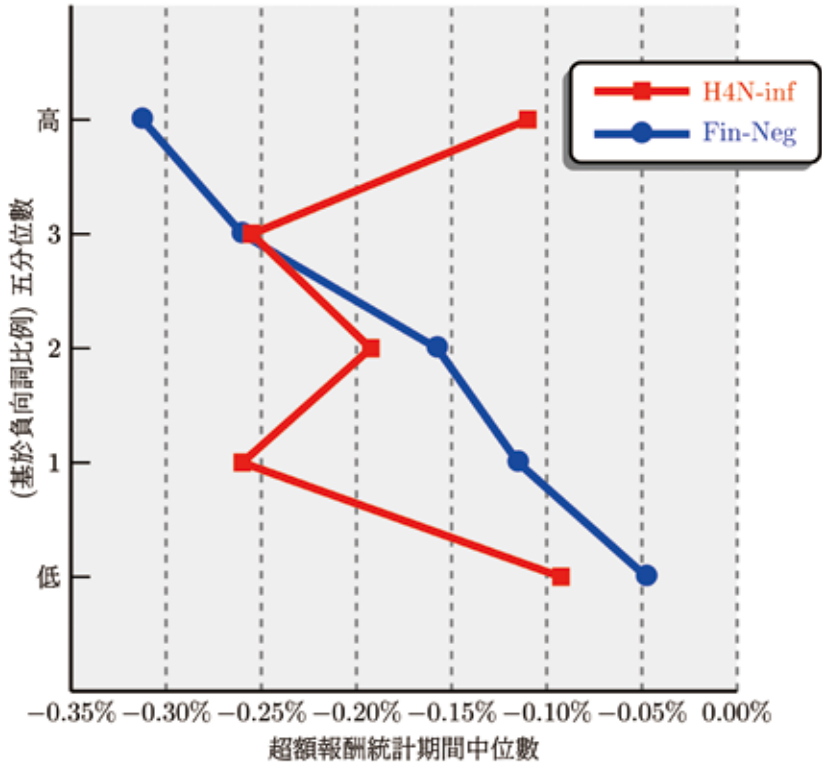
然而，由於近年來機器學習領域的蓬勃發展以及電腦效能的大幅提升，許多研究開始針對大量累積下來的財務文字資料進行分析及建模，不論學術界或產業界皆發展出許多相關模型及其對應之應用。

財務文獻上之財務文本分析

傳統上，財務相關文獻中關於軟訊息的分析方法一般分爲兩個部分[2]：第一部分爲提供一份人工產生之單詞列表或由某種演算法產生的詞典，其中每個單詞會被分爲正向或負向（或看漲或看跌）兩種類別之一。包括泰特洛克（Paul Tetlock）（2007）[3]以及泰特洛克、薩爾策錢斯基（Maytal Saar-Tsechansky）和馬可斯卡西（Sofus Macskassy）（2008）[4]在內的許多早期文獻都使用《哈佛心理社會學詞典》（*Harvard Psychosociological Dictionary*）的分類將單詞分爲正向或負向字。然而，上述的一般化情緒詞典往往不適用於特定專業領域。勞克蘭（Tim Loughran）和麥當勞（Bill McDonald）（2011）[5]在財務頂尖期刊《金融期刊》（*Journal of Finance*）中的文章表示，《哈佛心理社會學詞典》經常將財務文本中的常用詞錯誤分類，例如：vice 這個字被分類在該字典中的負面字詞，然而在財務的前後文中，vice 常與 president 一同出現並爲中性字詞。有鑑於此，勞克蘭及麥當勞於 2011 年運用簡易的資料分析技術並配合人工篩選的方式，針對美國證券交易委員會（U.S. Securities and Exchange Commission，SEC）所要求之財務年報 Form 10-K 中的文字資料建立了六部財務列表（後稱 LM 字典），該字典已成爲財務文字分析中最具代表性的一項資源。以下簡述此六部單詞列表相對應的字詞：

1. 負向字（Fin-Neg）：負向商務字詞，如：deficit、default。
2. 正向字（Fin-Pos）：正向商務字詞，如：achieve、profit。
3. 不確定字（Fin-Unc）：與不確定性相關的字詞（側重於不精確的一般概念，而非僅關注風險），如：appear、doubt。
4. 法律字（Fin-Lit）：反映法律訴訟傾向的字詞，如：amend、forbear。
5. 強情態詞（MW-Strong）：表達較強情態的動詞，如：always、must。
6. 弱情態詞（MW-Weak）：表達較弱情態的動詞，如：could、might。

測試字典品質的方法之一是檢查財務年報 Form 10-K 發佈時市場的反應與字詞之相關性。勞克蘭及麥當勞（2011）[5]表示如果年報中的情緒語氣很重要，包含較高比例負向字詞的財報所對應之公司平均而言應在發佈日期前後經歷負的



圖一：不同詞典與超額報酬之關係圖。圖中對於兩個單詞列表：H4N-Inf 及 Fin-Neg，根據負向單詞的比例，將 50,115 個財務年報 Form 10-K 的樣本分為五組。超額報酬係運用財報發佈日以降 3 天公司股票報酬來計算。

超額報酬（excess return）[1]。圖一爲發佈日期後公司股票價格之超額報酬中位數與《哈佛心理社會學詞典》中的「負向字詞（H4N-Inf）」、「負向財務單詞列表（Fin-Neg）」於財報中出現頻率比例的對應圖。圖一顯示，H4N-Inf 列表中的負向字詞所佔的比例與超額報酬中位數並無一致關係；然而，Fin-Neg 列表產生的單調遞減模式則是我們希望單詞列表能夠捕捉到有用訊息的方式 —— 包含較低負向字頻率的公司在財務年報 Form 10-K 發佈日期以降 3 天內的收益率略爲負值，而五分位數中包含較高負面詞頻的公司之中值收益率則急劇下降。

財務相關文獻中關於軟訊息的分析方法之第二部分爲如何對詞典中的每個單詞進行加權，並運用此加權將每篇文章之文字資訊對應到一個量化的數值。許多傳統財務文獻採用比例加權法，該方法透過文章中負向詞或正向詞與文檔中總詞數之比例來衡量此篇文章的語氣或情緒；然而，該方法也隱式地假定該類別中的所有單詞都同等重要。後續有相關文獻引入傳統文字分析常用的詞頻（term frequency，tf）、逆向文件頻率（inverse document frequency，idf）、兩者的組合及變形進行詞典中的每個單詞之加權 [6]。另一方面，傑格迪什（Narasimhan Jegadeesh）與吳迪（Di Wu）（2013）[2] 兩位學者於 2013 年首度使用迴歸模型計算字典中單詞與超額報酬的關係，運用迴歸模型中每個單詞的係數進行單詞加權，並計算每篇文本的語氣指標。表一比較上述超額報酬線性迴歸模型（word power，WP）與逆向文件頻率（idf）所計算出之字詞重要程度排序。由表一可得知，在負向詞中，根據 idf 被評爲影響最大的某些詞（如：disgrace）在 WP 排名中反被列爲影響最小的詞之一，此結果進一步凸顯了不同的加權方法所導致的明顯差異。

[1] 註：超額報酬係指公司的普通股與證券價格研究中心（The Center for Research in Security Prices，CRSP）價值加權市場指數購買和持有的收益（buy-and-hold return）之差值。

字詞的影響力			
		WP 排序	idf 排序
正向字			
ingenuity	獨創性	1	14
acclaimed	廣受好評的	2	7
influential	引領風潮的	3	26
⋮		⋮	⋮
worthy	有價值的	121	22
tremendous	巨大的	122	35
lucrative	暴利的	123	13
負向字			
imperil	威脅	1	18
disavow	否認	2	22
insubordination	不服從	3	20
⋮		⋮	⋮
disgrace	恥辱	710	1
⋮		⋮	⋮
dispossess	處置	718	8

表一：上述超額報酬迴歸模型（WP rank）與逆向文件頻率（idf rank）所計算出之字詞重要程度排序。

引入機器學習技術

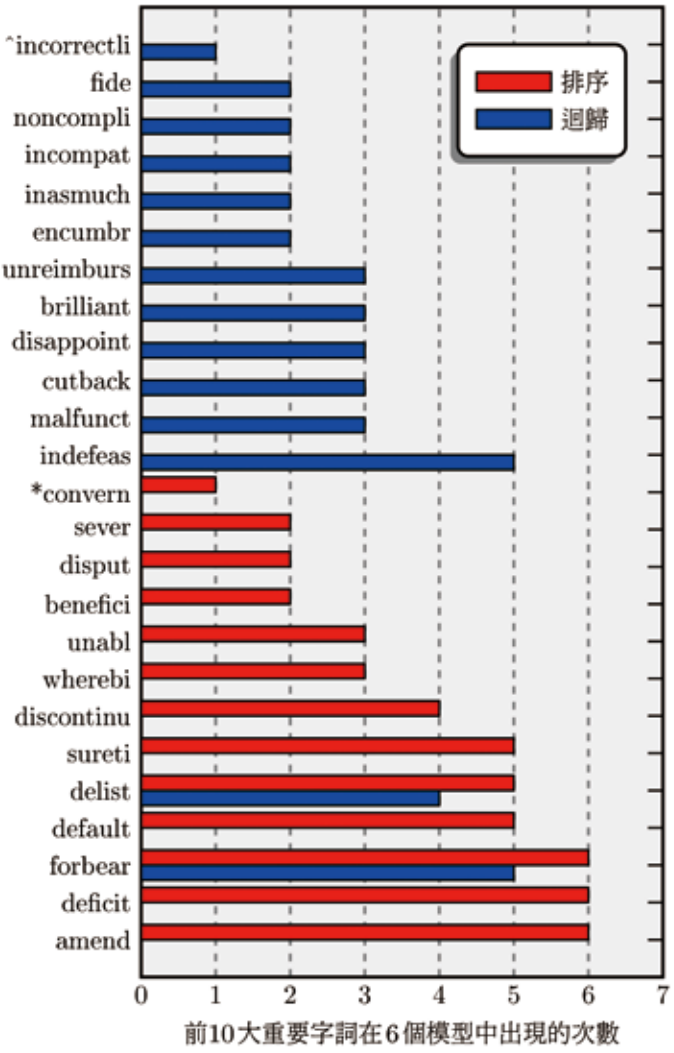
近年來，由於資料分析領域的蓬勃發展與資料的大量累積，一些學者開始利用文字探勘、自然語言處理、機器學習模型等相關技術針對財務文字資料進行研究。與傳統財務文獻研究重視分析現象的特性不同的是，機器學習模型更重視其模型針對目標的預測精準度。

柯岡（Shimon Kogan）等學者於 2009 年利用文字分析中常用的詞袋模型（bag-of-words models）及進一步之二元模型（bag-of-bigrams），針對財務中常用的風險指標「股票報酬波動度」（stock return volatility）進行迴歸模型之建立 [7]。其中，股票報酬波動度定義爲一段時間每日股票報酬率的標準差，而使用之迴歸模型則爲一般在機器學習領域中常用的支持向量機迴歸模型（support vector regression，SVR）（扎克〔Harris Drucker〕等，1997）。與傳統財務文獻研究中線性迴歸模型不同的是，SVR 可以更有效地處理詞袋模型中大量且稀疏的特徵值。精準地說，一個文本若使用詞袋模型來表示，其特徵將被顯示爲一個十分稀疏的高維度向量，每一個維度中的數值爲一種字詞的特徵，其可爲該字詞在此篇文章中是否出現（0 或 1）、出現的次數（詞頻）或詞頻與逆向文件頻率的乘積（term frequency-inverse document frequency，tf-idf）等不同形式。假設給定財務報告集合 $D = \{d_1, d_2, \cdots, d_n\}$，其中每個 $d_i \in \mathbb{R}^p$（即每個文檔都用一個 p 維度的向量表示），每篇財報 d_i 對應一家公司 c_i，作者試圖建立一個模型用以預測該公司的未來風險（在此用股票報酬波動度 v_i 爲代表）。這種預測可以通過參數化函數 f 來定義：$\hat{v}_i = f(d_i; w)$。上述目標在於給定訓練資料 $T = \{(d_i, v_i) | d_i \in \mathbb{R}^p, v_i \in \mathbb{R}\}$ 的情況下學習 p 維度之模型參數 w。此篇文獻的實驗結果顯示：

1. 僅使用文字訊息來預測波動率的模型在某些年份中非常接近運用歷史股票報酬波動度做爲特徵之模型；
2. 結合歷史股票報酬波動度及文字資訊可獲致具有更精準預測性能之模型；
3. 訓練資料對於時間因素是敏感的，在某些情況下使用更多的歷史財報增加訓練資料，並不一定能有更好的模型效能。

在此篇研究之後，陸續有許多學者利用機器學習技術針對財務風險進行文字分析之相關研究。舉例

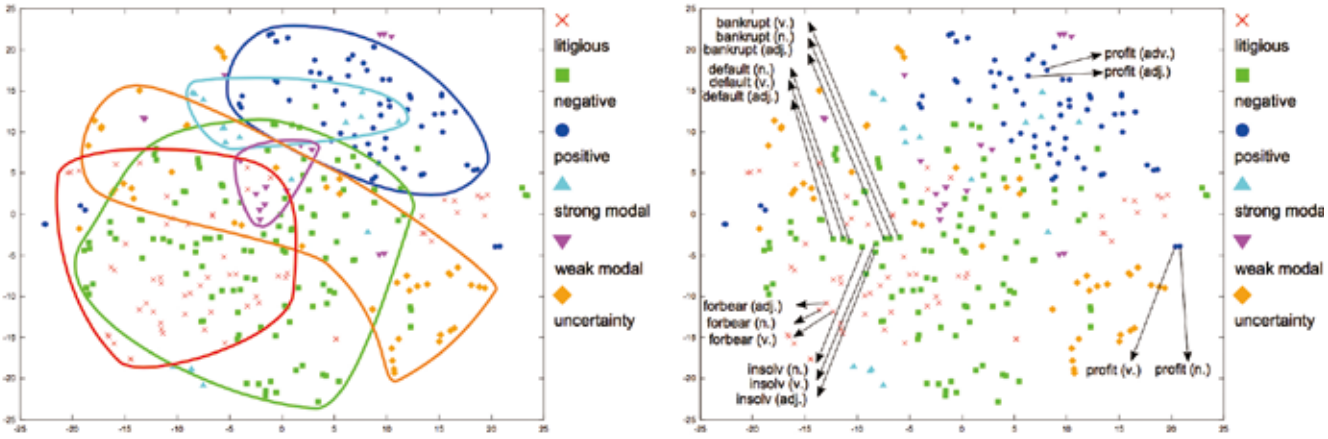
而言，蔡銘峰與作者在 2017 年的論文中引入機器學習中的學習排序（learning to rank）技術進行財報文字分析，該研究認爲直接利用單純文字資訊對實數值（如：股票報酬波動度）進行迴歸預測，可能會因爲文字訊息與數值資訊之本質差異太大，造成不易找出財務風險與字詞之間的關係 [8]。有鑑於此，該研究將預測問題簡化，透過根據公司之未來超額報酬對其進行風險等級分類，並利用排序模型找出公司未來風險等級及對應財務報告文字內容之關聯性。更精確地說，作者首先針對每年所有公司的股票報酬波動度切割爲不同的風險水平，其後按以下方式定義排序任務：給定財務報告 $D = \{d_1, d_2, \cdots, d_n\}$ 的集合，目標爲建立一個排序模型 $f : \mathbb{R}^p \to \mathbb{R}$，使得針對 $c_i \succ c_j$，$f(d_i) > f(d_j)$，其中，$c_i \succ c_j$ 表示公司 c_i 的風險排名高於 c_j，亦即公司 c_i 的風險大於 c_j。圖二爲該文所顯示迴歸模型（SVR）與排序模型（SVMRank）中最重要之前 10 項字詞列表及其在 6 個模型中出現的次數，此部分的 6 個模型分別爲運用 1996-2000、1997-2001、1998-2002、1999-2003、2000-2004、2001-2005 等 6 個時間區段中之財務報告及其對應公司風險等級之迴歸及排序模型所學習出的模型。由圖二可以發現，從排序模型中學到的單詞比從迴歸模型學到的單詞更加一致，例如：在 6 個排序模型中都出現 amend、deficit、forbear，也有 7 個前 10 大重要字詞出現於 4 個以上的模型中，獲得了多數票；而迴歸模型中只有 3 個單詞出現了 4 次以上。此實驗結果指出，採用排序模型來分析財務風險和文本訊息之間的關係可能比使用迴歸模型更爲合理，並更容易找出與風險高度相關的字詞。



圖二：迴歸模型與排序模型中前 10 大重要之字詞列表及其在 6 個模型中出現的次數。

神經語言模型（neural language model）

神經語言模型（或連續空間語言模型）將每個單詞表示爲一個連續表示式（representation 或 embedding）。此類型模型背後有一個稱爲「分布假說」（distributional hypothesis）的理論基礎，而分布假說的核心概念是具有類似前後文（context）



圖三：根據學習的單詞表示形式，對六個財務情緒單詞列表進行二維可視化。（Figure 1 in[9]）

的詞，其詞意較容易相近。在此類型模型中，每個單詞被顯示爲一個低維度的實數值密集向量（a dense vector with real-valued elements），而這些稱爲「詞向量」（word embedding）的表示法，由於其在捕捉語言的句法和語境規律方面具有優勢，近年來在各種自然語言的任務中顯現出令人鼓舞之結果。

有鑑於此，部分學者亦利用此類型語言模型之優勢，針對財務問題進行了一些嘗試和相關研究，舉例而言，蔡銘峰等學者於 2016 年利用當時著名之 word2vec 模型中的連續詞袋模型（continuous bag-of-words model，CBOW），針對大量財務報告資訊進行語言模型訓練，亦在訓練過程當中考慮詞性不同所造成的影響，並利用訓練完成的詞向量進行 LM 財務字典的自動化擴充 [9]。圖三爲其圖像化 LM 中六部財務情緒字典之詞向量在二維空間的分佈，其中，詞向量爲利用 1996 至 2013 年共 40,708 篇財務報告（在提取詞幹後一共包含 125,370 個不同的詞彙）之文字資訊進行學習。如圖三中左圖所示，同一字詞列表中的單詞通常會聚合爲一組，正向詞和負向詞這兩組之間幾乎沒有重疊，而訴訟和不確定性字詞與負向詞有很大一部分重疊，這意味著在財務上訴訟和不確定性單詞通常與負向詞彙含義具有高度關聯性。此外，從圖三中右圖所示，作者發現根據學習到的詞向量，default、insolvent、bankruptcy 三個具有相似含義的單詞在空間中彼此接近。以上這些現象說明了此類型模型在財務報告中具有捕捉情境規律性的能力。因此，作者認爲透過找尋與原 LM 字典中在詞向量空間中相近的字，即可有效地進行財務情緒字典的自動化擴充，用以發現新的財務關鍵字。另一方面，作者亦發現（見圖三右），儘管帶有不同詞性標籤的同一單詞彼此在空間中有時非常接近（如：profit 的名詞和動詞），但有時卻相距甚遠（如：profit 的名詞和形容詞），這突顯了在擴展財務字典關鍵詞時考慮詞性標籤的必要性。

2017 年，雷克薩斯（Navid Rekabsaz）等學者更進一步結合 word2vec 模型根據 LM 情緒字典延伸出來的字詞，針對公司未來股票報酬波動度進行預測 [10]。其做法主要運用下列四種字詞的加權方法，針對 LM 字典中的正向、負向和不確定字詞（後稱財務關鍵詞）進行計算：

$$\text{TC} : \log(1 + \text{tc}_{d_i}(t))\,,$$

$$\text{TF} : \frac{\log(1 + \text{tc}_{d_i}(t))}{||d_i||}\,,$$

$$\text{TFIDF} : \frac{\log(1 + \text{tc}_{d_i}(t))}{||d_i||} \log(1 + \frac{d_i}{\text{df}(t)})\,,$$

$$\text{BM25} : \frac{(k+1)\overline{\text{tf}_{d_i}(t)}}{k + \overline{\text{tf}_{d_i}(t)}}\,,$$

$$\overline{\text{tf}_{d_i}(t)} = \frac{\text{tc}_{d_i}(t)}{(1-b) + b\frac{|d_i|}{\text{avgdl}}}\,。$$

在以上式子中，$\text{tc}_{d_i}(t)$ 是財務報告 d_i 中財務關鍵詞 t 的出現次數，$||d_i||$ 表示財務關鍵詞權重的歐幾里得距離，$|d_i|$ 是財報的長度（報告中的單詞數），avgdl 指平均財報長度，最後 k 和 b 是外部參數（作者根據 [11] 在文章中使用 $k = 1.2$ 及 $b = 0.65$ ）。另一方面，針對這些財務關鍵詞，作者利用 word2vec 模型找出其相似字詞，並利用當時最先進的資料檢索方法 [11]，針對上述加權方法進行修正：

$$\overline{\text{tc}_{d_i}}(t) = \text{tc}_{d_i}(t) + \sum_{t' \in R(t)} \text{sim}(t, t')\text{tc}_{d_i}(t')$$

其中，$\text{sim}(t, t')$ 爲單詞 t 與 t' 之 word2vec 向量表示法的餘弦相似度（cosine similarity），而 $R(t)$ 爲單詞 t 利用 word2vec 模型找出的相似詞之集合（此處作者參考 [11] 使用閾值爲 0.70 的餘弦相似度來選擇相似單詞）。作者於上述四種加權方式中置換 $\text{tc}_{d_i}(t)$ 爲 $\overline{\text{tc}_{d_i}}(t)$ ，用以加入 LM 字典以外之相似詞的影響，最後利用主成分分析（principle component analysis）對稀疏的向量表示式進行降維，並針對降維過後的財報向量進行 SVR 模型的訓練。此篇文章結果顯示，考慮延伸字詞的 BM25 在預測上可達到最佳效能。

超越字詞等級之文字分析

以上所提及之研究工作主要是針對單詞層次進行分析，然若要進行更精準的語言理解，光靠單詞層次的分析是不足的，以下列從財報中抽取出的一句話爲例：「A technological breakthrough or marketing or promotional success by one of our competitors could adversely affect our competitive position.」其中，breakthrough 和 success 爲正向單詞，而 adversely 爲負向單詞，如果單純以單詞頻率計算，此句話可能會被演算法視爲一個具有正向意涵的句子，但就文意而言，此句話應爲具負向意涵之句子。上述例子顯示，文本理解中一個重要的挑戰 —— 即文本中包含的語義絕非多個單詞含義的簡單組合，在某些情況下，常用的關鍵字比對技術或單詞層次的分析通常是不足甚或不可行的。有鑑於此，作者認爲後續如需進一步針對財務非結構化進行更細緻的分析或語義理解，引入超越單詞層次的方法（如：多字詞表示式層次〔multi-word expression level〕、句子層次〔sentence level〕抑或段落層次〔paragraph level〕）是十分重要且不可或缺的。針對此部分，諸多新穎的深度學習模型亦在自然語言處理上有快速的進展，例如：2013 至 2018 年許多基於連續空間語言模型之語句層次學習演算法紛紛被提出（如：skip-thought [12]、Siamese CBOW [13] 等）。但上述方法存在一個問題 —— 同樣的單詞在不同前後文中其實具有不同

的含義，但在前述方法中，每個單詞的向量表示式是相同的。爲了解決此一問題，彼得斯（Matthew Peters）等學者首先提出 ELMo 模型 [14]，係一個深度語境單詞表示式學習法（deep contextualized word representations），後續谷歌（Google）提出的 BERT[15] 和臉書（Facebook）所提出的 RoBERTa [16] 模型亦具有相同的特性，並已在多項自然語言處理任務上達到令人相當驚豔的效能。

財務文字分析對於自然語言處理技術之影響

以上文章內容均在討論不同文字分析或自然語言處理技術對於財務非結構化文字分析的做法，但從另一個面向而言，財務文字分析中的一些特性本身即對於自然語言處理研究具有相當的吸引力 [7]。首先，在財務文字分析中預測目標的實用性或存在性（以上述財務風險預測工作爲例，即爲股票報酬波動度）大多是沒有爭議性的，而在許多自然語言處理的工作中，評估困難和註釋者（annotator）之間的分歧通常是個亟待解決的問題；但如股票報酬波動度這樣的預測目標則是總結有關現實世界事實（股票價格）的一個統計量，不受人類專業、知識或直覺的主觀影響。因此，這樣的預測任務可爲任何類型的語言分析提供一個嶄新的、客觀的測試平台；其次，這樣的預測目標亦解決了標註資料的問題，許多自然語言工作依舊仰賴昂貴的標記資料資源（如：對齊的雙語語料庫），而上述的財務報告對於風險預測的工作，其使用的文本和歷史財務數據皆爲免費的（依法提供），且是美國經濟的副產品，所以任何人皆能以相對較少的費用獲得大量的數據。是以，這樣的財務文字分析對於測試自然語言模型語義理解之能力是一個十分優良的平臺。

結語

與許多跨領域研究相同，財務文字分析涉及財務與機器學習、自然語言處理之知識、技術的整合。如何透過財務領域專家與資訊科學家充分的討論與合作，找出重要的財務研究議題，並利用或開發相對應之機器學習與自然語言處理演算法，乃此類型研究之重要課題。此外，不同於傳統財務研究多使用硬訊息（數字訊息），目前研究和應用軟訊息於財務上仍相對較少，若能配合近年來自然語言模型在語義理解上的快速發展，更有效地利用並整合不同來源之文字資訊，對金融市場及財會相關研究應能有突破性的發展。

本文參考資料請見〈數理人文資料網頁〉
http://yaucenter.nctu.edu.tw/periodical.php

延伸閱讀

- ▶彭文志與黃思浩〈人工智慧在金融科技上的應用〉《科學發展》2019 年 3 月，555 期，20-27 頁。
- ▶國立臺灣大學計算機及資訊網路中心電子報，第 0031 期，專題報導「Text Mining」的相關文章。
  http://www.cc.ntu.edu.tw/chinese/epaper/0031/
- ▶謝邦昌《Text Mining 文本探勘》，中華資料採礦協會授權出版。介紹文本探勘的概念和資料分析的流程，以及如何使用多種軟體執行。
- ▶王釧茹〈注意！若財報出現這些字，未來財務風險高〉，《研之有物》，中央研究院。
  http://research.sinica.edu.tw/computational-finance-wang-chuan-ju/