國立政治大學資訊管理學系

碩士學位論文

指導教授: 姜國輝博士、季延平博士

以深度學習實作情感分析於台灣股價指數趨勢之研究—以光電類股為例

Applying Deep Learning to Sentiment Analysis for Prediction of Taiwan Optoelectronic Index Trend

研究生：黃彥霖

中華民國一〇七年七月

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

摘要

光電是一門結合光學、電子與電機的技術，以光電技術為核心，合成各類零組件、設備，並廣泛應用於通訊、生化、醫療、工業、能源等各種領域。近年來台灣的精密光學元件發展快速，受惠於智慧型手機興起，帶動了台灣光電產業發展。據過往研究指出，網路上的文本訊息會對群眾情緒造成影響，進而影響股價波動，因此若能將網路上大量財經文本快速分析，來推測投資大眾情緒，進而預測走勢，可以幫助投資者增加其獲利。為檢驗新聞情緒對於價格預測的重要性，本研究嘗試使用混和非監督式學習與監督式學習的方式進行實驗模型的建立。

在非監督式學習方面，本研究利用 LDA 主題模型，將光電類股之財經新聞文本，以進行主題分群。將各主題之文本，藉由比對情感詞集的方式進行情緒指數之計算，藉此找到各文本的情緒傾向。接著利用視覺化的方式，找出對後續分類模型有所助益的領先指標特性之主題。

在監督式學習方面，本研究分別建立 SVM 分類模型與 LSTM 神經網路分類模型，針對光電類股的財經新聞文本進行情感分析，並與單純使用技術指標之分類模型做比較。實驗結果發現，結合技術指標與情緒指數之準確率，SVM 模型約為 70%，神經網路模型約為 60%。在股價漲跌分類方面，SVM 模型優於 LSTM 分類模型，但不論是使用 SVM 模型或是 LSTM 模型，混和技術指標與情緒指數之模型在準確率上都高於單純只用技術指標之模型，前者較後者可提升多達 7%之準確率，顯示情感分析確實能有效提升光電類股指數趨勢模型之準確度。

關鍵詞：情感分析、LDA 主題模型、支援向量機、長短期記憶網路

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

Abstract

The optoelectronic is a technology that combines optics and electronics. In recent years, Taiwan's precision optical components have developed rapidly, benefiting from the rise of smart phones, which has driven the development of Taiwan's optoelectronic industry. According to past studies, text documents on the Internet will affect the mood of the investors, and then affect stock price indirectly. For investors, it is important to know how to analyze the potential emotion in text documents and use it to predict the stock trend. In order to test the importance of text emotions for price prediction, this research established a model that mixed unsupervised learning and supervised learning.

In the part of unsupervised learning, this research used the LDA model to assign documents to topics. Next, calculated the sentiment index to find the sentiment tendency of each documents. Then this research used the visualization to find the topic that was probably a leading indicator and helpful to classification model.

In the part of supervised learning, this research established the SVM model and the LSTM neural network model respectively. The result showed that, mixed with technical indicators and sentiment index, the accuracy of the SVM model was about 70%, and that of the neural network model was about 60%. In classification, the SVM model was better than the LSTM classification model, but in both of the SVM model and the LSTM model, the accuracy of the model that included the sentiment index and technical index was higher than that which included only technical index. The former could improve the accuracy by up to 7% compared with the latter, showing that sentiment analysis was able to improve the accuracy of the prediction of Optoelectronics stock index trend effectively.

Keywords：Sentimental Analysis, LDA, SVM, LSTM

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

目　次

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

表次

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

圖次

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

第一章　概論

第一節　研究背景

光電產業是一門結合光學、電子與電機的技術，以光電技術為核心，合成各類零組件、設備，並廣泛應用於通訊、生化、醫療、工業、能源等各種領域。(邱世芳，2008)。依據光電科技工業協進會（PIDA）對光電產業之分類，可將其分為六大類：光學元件、光電顯示器、光輸出入、光儲存、光通訊、雷射及其他應用，如表所示。

大分類	中分類	項目
光學元件	發光元件	雷射二極體、發光二極體
	受光元件	光二極體與光電晶體、電荷耦合元件、接觸式影像感測器、太陽電池
	光學元件	
光電顯示器	液晶顯示器（LCD）、發光二極體顯示幕(LED)、真空螢光顯示器、電漿顯示器、有機電激發光顯示器(OLED)、場發射顯示器、液晶投影機	
光輸出入	影像掃描器、條碼掃描器、雷射印表機、傳真機、影印機、數位相機	
光儲存	裝置	消費用途、資訊用唯讀型、資訊用可讀寫型
	謀體	唯讀型、可寫一次型、可讀寫型
光通訊	光通訊零組件	光纖、光纜、光主動元件、光被動元件

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

	光通訊設備	光纖區域網路設備、電信光傳輸設備、有線電視光傳輸設備、光通訊量測設備
雷射及其他光電應用	雷射本體	
	工業雷射	
	醫療雷射	
	光感測器	

表 1　光電產業定義分類一覽表(光電科技工業協進會 PIDA, 2000)

根據光電協進會(PIDA)統計，2014 年台灣光電產業總產值達新台幣 2.467 兆元，佔有全球光電產業 5,766 億美元總產值約 12%。其中產值排名前五大的產品分別為 TFT-LCD 面板、觸控面板、LCD 元件材料、太陽電池及精密光學元件與鏡頭；至於產值高成長率的前五大產品，分別為 LED 照明、太陽能矽晶材料、精密光學元件與鏡頭、LED 元件及光學治療。（光電科技工業協進會，2014）

近年來台灣的精密光學元件發展快速，由於智慧型手機興起，受惠於手機品牌紛紛推出新產品，帶動光電產業發展，台灣精密光學元件 2014 年產值達新台幣 977 億元。在產品技術趨勢上，800 萬素已成為智慧手機鏡頭的基本規格，千萬級像素更早已出現在高階機種，其他趨勢如薄型化、大光圈、防手震，更將成為精密光學元件未來要搶奪手機大廠訂單的關鍵技術。

第二節　　研究動機

過往研究指出，人們的情緒往往會影響股市的變動。投資者對於市場的信心、情緒有相當程度的相關性。因此，如何去快速分析大量的網路文本資料來推測其情緒成為重要的課題。

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

過去情感分析的相關研究使用傳統的機器學習方法，先透過非監督式方法進行文本主題標注與情緒傾向標注，而後再用監督式學習方法（如支援向量機SVM）建立分類模型以預測股價指數趨勢。而近年來硬體運算能力的上升，神經網路（Neural Network）與深度學習（Deep Learning）都有快速發展，在許多領域上都有傑出的表現，特別是在自然語言處理（Natural Language Processing，NLP）的領域中，包含了語音識別、文本摘要、文本分類、機器翻譯以及問答系統，都有著顯著的進步，因此本研究欲嘗試使用深度學習來進行文本情緒之分類，並驗證情感分析對於股價指數漲跌之預測準確度，是否有所幫助。

第三節　研究目的

本研究將蒐集各大財經新聞網的相關產業新聞及相關產業評論，進行以下步驟：

1. 先以非監督式的學習方法建立 LDA 主題模型，再針對財經文本內容將其標注不同的主題。
2. 將各主題財經文本作情緒指數計算與情感傾向的標注，標注其所屬正面、負面或中性情感傾向，並透過視覺化分析找出具有領先指標特性之主題。
3. 分別建立 SVM、LSTM 兩種分類模型，來對文本進行分類，並驗證兩者分類效果，檢驗情感分析對於股價指數漲跌之預測準確度的提升是否有所幫助。

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

第二章　　文獻探討

第一節　情感分析

情感分析又被稱為意見探勘(Opinion Mining)，其被定義為透過自動化技術找出作者對於特定主題或標的所表達的正負情感、意見與態度 (Pang and Lee, 2008; Liu, 2012; Feldman, 2013)。情感分析涉及自然語言處理、文字探勘、資料檢索以及機器學習等領域，透過應用情感分析於文本資料上，可快速分析出文本中所隱含的情緒傾向，以及針對某特定對象或主題所提出的正面或負面意見，無須再透過人工方式的去解析文本資料，無須耗時費力即可得知大眾對於特定事件或標的物的看法(張良杰, 2014; Liu, 2012; Mishne, 2006)，如李亭宜（2017）利用 Twitter 上消費者的言論進行文本情感分析，來得知消費者情緒和其品牌價值的關係。

許多情感分析之研究，是基於是先定義好的正面詞庫與負面詞庫來進行情感分數的計算（Taboada et al., 2011）。而中文的字詞情感除了使用已建好的詞庫外，也可以自行建置詞庫，首先會建立一個已知情感傾向的種子詞集(Seed Words)，再透過外部的語彙資料像是 WordNet、HowNet 等找出相關字詞，並對種子詞集進行擴充來完成自行建置之詞庫(林育龍, 2014)。

第二節　隱含狄利克雷分佈（LDA）

隱含狄利克雷分佈（Latent Dirichlet Allocation, LDA）是一種以統計為基礎、非監督式學習的主題分類模型，被廣泛應用於文本主題識別、文本分類以及文本相似度計算等領域。

LDA 主題模型透過 hyper-parameter α、β 來做機率分布控制，α 控制主題

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

於文章中的機率分布， β 控制字詞於主題中的機率分布， θ_i為主題於文件 i 的機率分布，φ_k 為字詞在主題 K 的機率分布， $Z_{i,j}$為文章 i 第 j 個字詞的主題， W 為文件中的字詞， K 為主題數量， N 為文件中字詞總數， M 為文件總數， 如圖所示。

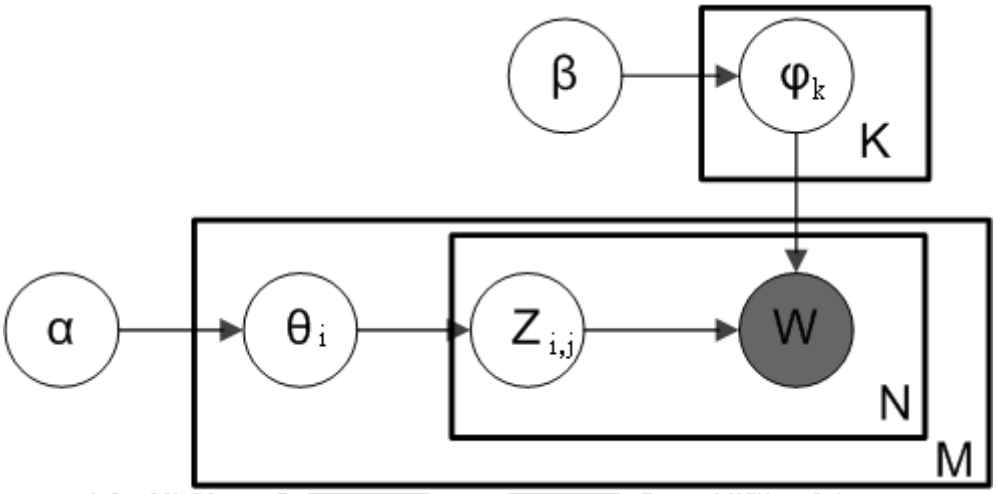

圖 1　LDA 主題模型結構（Blei, Ng, Jordan, 2003）

α、β 變數對於主題模型有很大的影響，α 值過大可能會造成文件上的主題分布效果不佳，難以判斷該文章屬於何種主題；β 過大可能會造成主題上的字詞分布不精確，進而影響主題判別的準確度，因此需不斷調整修正 α、β 變數來達到主題分類效果最佳化（洪崇洋，2012）。

機率模型的表現，常使用 Perplexity 來評估，當 Perplexity 值越小，模型表現越佳(Griffiths & Steyvers, 2004)。因此可經由不斷調整主題數量來產生不同 Perplexity 值，作為選擇主題數量的方式，當 Perplexity 收斂時，該主題數量即為最佳的主題數量。

過去許多研究指出 LDA 於未知主題文本資料分類都有良好的效果。洪崇洋（2012）利用 LDA 主題模型可精確地分類大量的電子書，找出其主題，進而作為一個輔助的資訊供讀者參考。著名社群網站 Twitter 本身並無提供文章分類的功能，而蕭昱維（2014）利用 LDA 找出社群網站 Twitter 中具有代表性的主題，作為提供使用者閱讀模式的需求。張日威（2014）利用 LDA 找出社群網

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

站噗浪（Plurk）中文章的相關主題，並透過情感分析對相關主題進行分類，讓使用者得以快速瞭解大眾對於相關主題的喜好程度。劉羿廷（2015）使用 LDA 來對財經文本進行標注，發現效果明顯優於 TFIDF-Kmeans 和 NPMI-Concor 等其他主題模型。

第三節　支援向量機（SVM）

支援向量機（Support Vector Machine, SVM）為 Vapnik 在 1995 年所提出的一種監督式學習的演算法。其概念為透過找出空間中的一個超平面（Hyperplane）使得空間中的點能夠被其區隔開來，如圖所示。

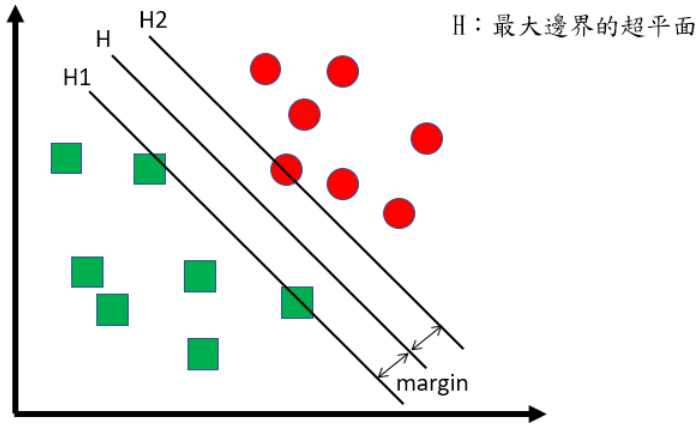

圖 2 支援向量機示意圖

能夠將資料點分開的超平面不只一個，而擁有最大邊界（Margin）的超平面稱為最佳分離超平面（Optimal Separating Hyperplane, OSH），SVM 的目標是要找出這個最佳分離超平面來達到最好的分類效果。

然而現實中並非所有的資料都能找得到其線性的超平面，針對這類問題 Bernhard E. Boser(1992) 發現如果將原始資料透過非線性的映射函數Φ轉換到一個更高維度的特徵空間（feature space）中，此時可將原本無法線性區分的資料，在高維度特徵空間中找到其線性超平面。如下圖，將原先無法線性區分

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

的二維資料，經由映射函數轉換到三維特徵空間中後，便可找到一超平面將資料區隔開來。

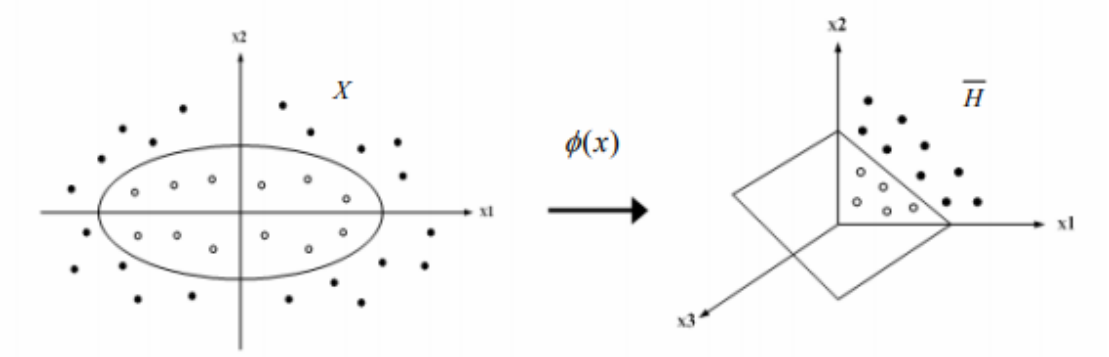

圖 3　將 2 維資料集經由映射函數轉移到 3 維特徵空間（雷祖強 et al., 2007）

第四節　Word2vec 模型

過往在處理文字資料時，常使用 one hot encoding 的方法來將文字數值化，即從 1 開始把辭典中每個詞都賦予一個編號，如「快樂」是 11 號，「高興」是 27 號，缺點是雖然我們知道兩個詞意思很相近，卻無法數值看出量者之間的關係。

詞向量（Word Vector, Word Embedding）是近年來在自然語言處理(Natural Language Processing, NLP)中被廣泛使用的一種將文字數值化的技術，好的詞向量具有「相似的詞其向量在空間中也很接近」的特性，因此彌補了上述的缺點。

Word2vec，為一群用來產生詞向量的相關模型，是 2013 年 Google 的 Tomas Mikolov 在 Google 帶領的研究團隊所提出的演算法。其訓練模型可分為 CBOW 和 Skip-grams 兩種。（Mikolov et al., 2013）

第一種是 Continuous Bag Of Words（CBOW），此方法會利用上下文的詞來當作神經網路的輸入，最後預測這個目標的詞是什麼，即「根據上下文的詞預測位置 t 的單詞」；第二種是 Skip-Gram 演算法，跟第一種演算法相反，它輸

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

入當前的詞來預測這一段的文章上下文的詞，即「根據位置 t 的單詞預測其上下文」。如下圖所示。

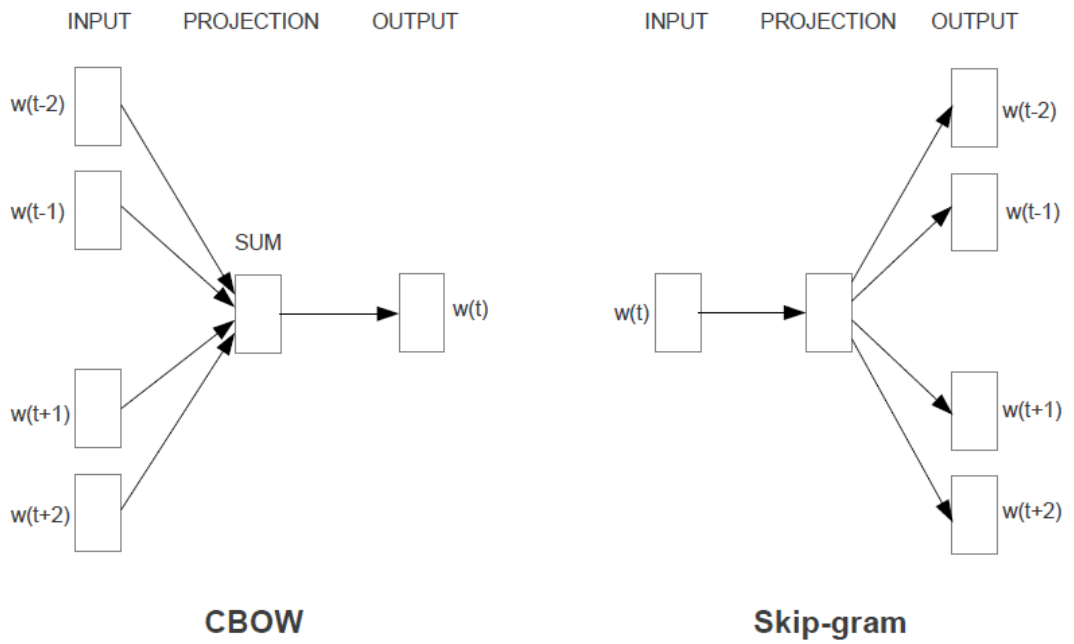

圖 4　CBOW 模型與 Skip-gram 模型（Mikolov et al., 2013）

CBOW 模型的訓練時間短，語法的資訊含量高；Skip-gram 的訓練時間長，但語意分析表現較好。（Mikolov et al., 2013）。故本研究使用 Skip-gram 來進行 Word2vec 的訓練。

第五節　深度學習（Deep Learning）

一、　神經網路

神經網路是一種對於生物神經元的簡單模擬，它從外界環境或其他神經元取得資訊，加以運算後，輸出其結果到外界環境或其他神經元。一般而言，神經網路由輸入層、隱藏層、輸出層所構成，其中輸入層、輸出層為一層，隱藏層數目則不一定，可視情況而定來做調整。其基本架構如圖所示。

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

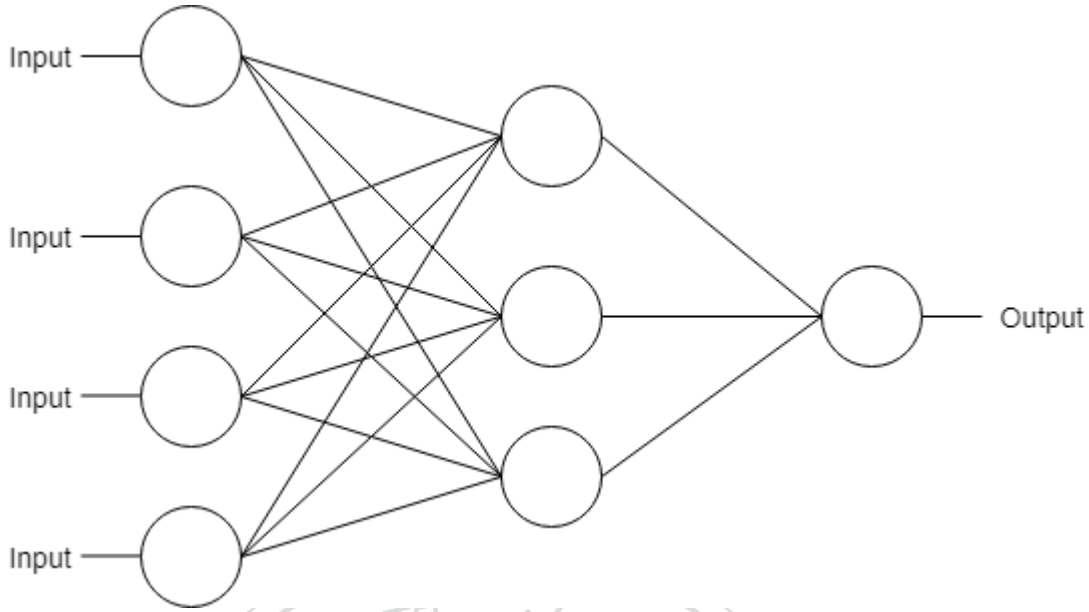

圖 5　神經網路架構示意圖

神經網路最早是由 Warren McCulloch 和 Walter Pitts 於 1943 年所提出的數學模型（W. McCulloch et al., 1943），但一開始受限於當時的計算能力不足，使得模型存在著非常大的侷限性。直到 1986 年 David Everett Rumelhart、Geoffrey Everest Hinton、Ronald J. Williams 提出了反向傳播（back propagation）的演算法後，大幅降低了訓練神經網路所需要的時間（D. Rumelhart et al., 1986），加上 1980 年代末電腦性能的快速發展，使得神經網路從此有了突飛猛進的進展，也就是現在所稱的深度學習（Deep Learning）。如今深度學習在許多機器學習的領域都有出色的表現，在影像識別、語音辨識、音訊處理、自然語言處理、機器人、生物資訊處理、搜尋引擎、醫學自動診斷、金融等各大領域都能看到其應用。

神經網路的訓練，主要是透過最小化損失函數（Loss function，或稱代價函數、成本函數 Cost function）來達成，損失函數為預測值和實際值的差異程度，其值越小則代表模型的表現越好。

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

二、 遞歸神經網絡（RNN）

遞歸神經網絡（Recurrent Neural Network, RNN）又稱遞迴神經網路、循環神經網路，是專門用於處理序列資料的深度學習模型，被廣泛應用於語音辨識、語言模型、機器翻譯等領域。

RNN 的架構如圖所示。RNN 與典型的神經網路最主要的差別就是 RNN 存在時間的概念，擁有「記憶」的能力。在每一個時刻，RNN 主題結構 A 會讀取 t 時刻的輸入X_t以及一上時刻的狀態作為其 Input，輸出一個值h_t並將結果傳到下一步。即每個時刻的結果，都是由此刻的輸入與上一時刻的結果所共同決定的，這樣的結構特徵讓它因此擅長解決時間序列相關的問題。

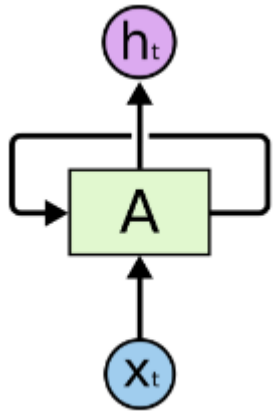

圖 6 RNN 架構示意圖

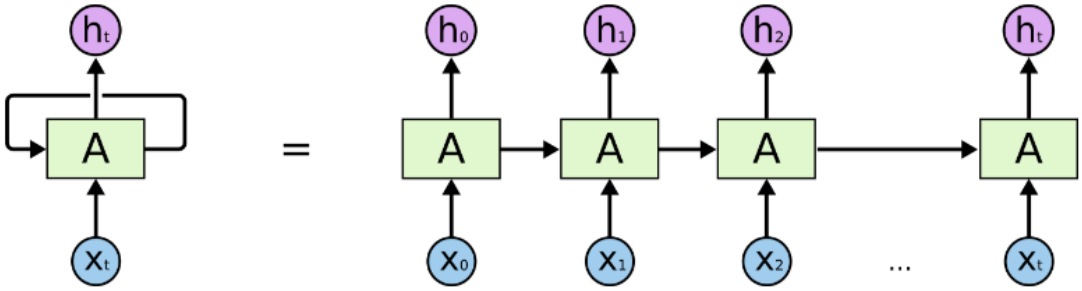

圖 7 RNN 按時間展開後的結構

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

三、 長短期記憶網路（LSTM）

RNN 是在有序的資料上進行學習，為了記住這些資料，RNN 會像人類一樣對之前發生的事件產生記憶，但 RNN 的結構使得它只能對距離較近的時刻有強烈的記憶，而年代久遠的事件則記不清楚，也就產生了這種長期依賴問題（Long-term dependencies）

長短期記憶網路（Long Short-Term Memory, LSTM）是基於 RNN 神經網路的一種改進，由 Sepp Hochreiter 和 Jurgen Schmidhuber 於 1997 年提出。（S. Hochreiter & J. Schmidhuber, 1997），主要是為了解決典型 RNN 網路可能會遇到的長期依賴問題，其架構如下圖。

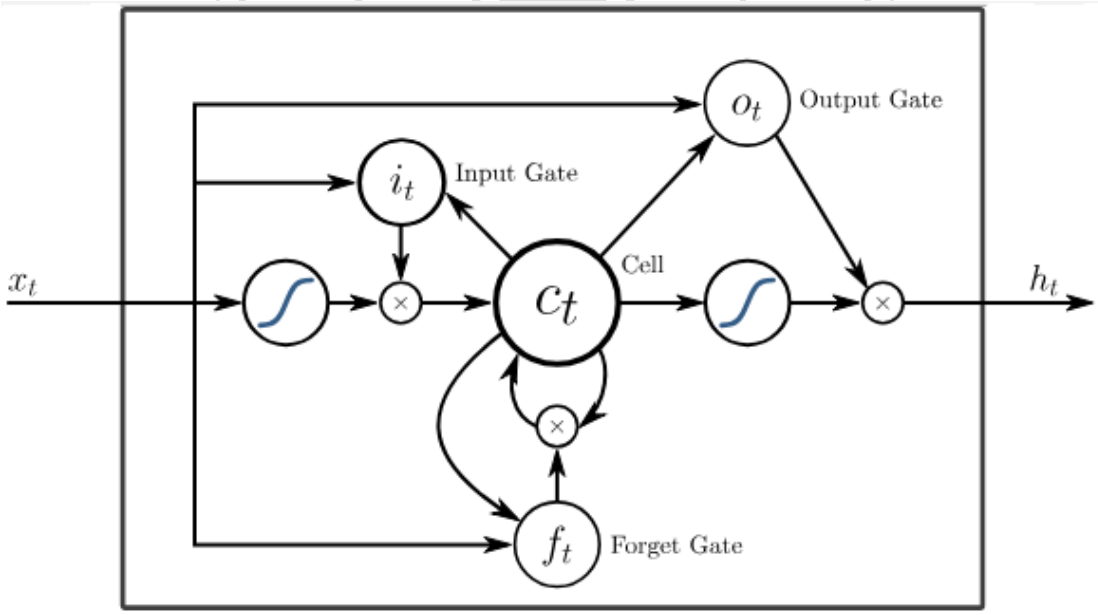

圖 8 LSTM 單元結構(wikipedia)

典型的 RNN，其每個時刻的狀態都由目前時刻的輸入和原有的記憶組合而成，然而記憶是有限的，早期的記憶會隨著時間而衰減。因此，LSTM 在原有的短期記憶單元上，增加了一個長期記憶和三個 gate 來進行控制：

1. Input Gate：控制新記憶寫入長期記憶的程度
2. Forget Gate：控制上一時刻記憶的遺忘程度

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

3. Output Gate：控制短期記憶如何受長期記憶影響

如果一個事件很重要，輸入門就針對其重要程度將短期記憶合併至長期記憶，或透過遺忘們忘記部分長期記憶後取代其成為新記憶。最後輸出門再基於長期記憶和短期記憶綜合判斷應該有什麼樣的輸出。

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

第三章　研究方法

本研究之流程圖，如下圖所示。

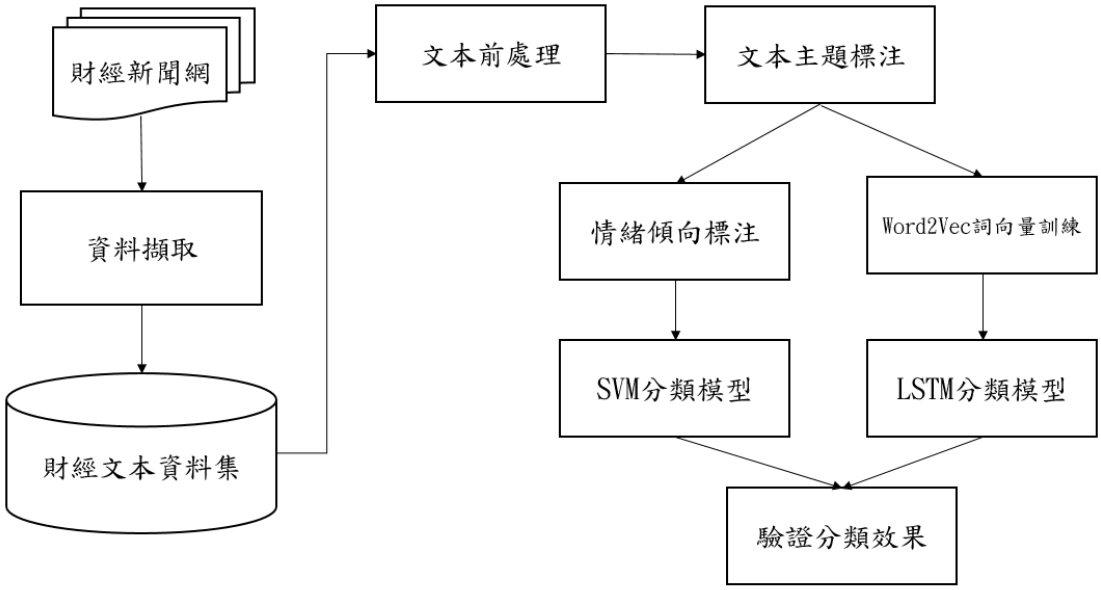

圖 9　研究流程圖

利用網路爬蟲於各大財經新聞網頁擷取股市文本資料，對其進行文字前處理後，使用非監督式學習建立主題模型，將文本依不同主題進行分群。之後利用情感分析的方式，以情感詞集計算出情緒指數並標注情緒傾向，將其作為 SVM 分類模型之輸入；使用 Word2vec 將文字做數值化後，將詞向量作為 LSTM 模型之輸入。最後驗證 SVM 模型與 LSTM 模型之分類效果。

第一節　資料蒐集

本研究使用 Python 撰寫之爬蟲程式，蒐集知名的財經網站如鉅亨網、經濟日報等 2017 年 07 月 01 日至 2018 年 04 月 30 日之財經新聞文本資料。由於光電類股所包含的公司眾多，本研究取市值最高的前 10 間公司作為新聞爬蟲的目標：

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

- 大立光(3008)
- 群創光電(3481)
- 友達(2409)
- GIS-KY(6456)
- 晶電(2448)
- TPK-KY(3673)
- 瑞儀光電(6176)
- 彩晶(6116)
- 玉晶光(3406)
- 亞洲光學(3019)

第二節　文本前處理

一、　中文斷詞

文章是由句子所組成，而句子是由字詞所組成。在英文文章中，可以很簡單地用空白來將字詞所分隔開來；但中文文章的字詞卻是相鄰的，因此本研究使用詞庫容易擴充的 Jieba 中文斷詞系統，來將取得之文本資料以字詞為單位斷開。

中文斷詞前
大立光第 1 季營運淡季效應顯著，表現不如法人預期
中文斷詞後
大立光　第一季　營運　淡季　效應　顯著　表現　不如　法人　預期

表 2　中文斷詞

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

二、 詞性標注

接著進行詞性標注，參照 Jieba 的字詞詞性表，對每個字詞進行中文詞性(Part-of-Speech)的標注。

三、 否定詞處理

否定詞的出現會導致語意相反，因此需要對否定詞進行處理。本研究參照李啟菁、王正豪(2010)所提出之採用區間 4 的範圍判斷的方式來調整，即搜尋述詞前 2 個位置與後 2 個位置是否有否定字存在，若存在則將該述詞作否定字處理。

四、 停用字過濾

有些詞雖然出現頻率很高，但並不會影響語意，對於文本意義也沒有貢獻，屬於無法提供重要資訊的字詞，可以將其去除，如：的、了。

五、 詞性過濾

為了增加後續分析之準確性，本研究將過濾其他多餘的詞性，僅保留具主題意義的體詞（名詞）和具有情緒的述詞（動詞），並移除其他多餘的詞性 (劉吉軒、吳建良，2007)。

第三節 文本主題標注

一、 建立 LDA 主題模型

LDA 主要是透過 hyper-parameter α和β來作機率分布控制，α控制主題

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

於文章中的機率分布，β控制字詞於主題中的機率分布，LDA 預設值為α=50/K、β=0.1，可依研究內容而自行調整至最佳的值，之後再藉由不斷調整主題數量來產生不同的 Perplexity 值，直到其收斂即可找到最佳的主題數量。LDA 會根據前一步驟的參數設定對熱門議題詞字進行運算，透過吉布斯採樣迭代計算出各個字詞於各主題中的機率分布，進而找出字詞對應的主題，建立 LDA 主題模型。

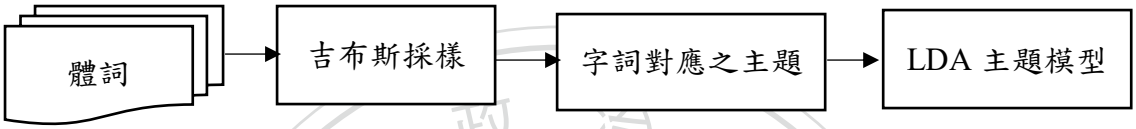

圖 10　LDA 主題模型流程圖

二、　判斷文本主題

根據文本出現各主題的議題詞數量，判斷文本屬於何種主題的機率最高，並將其做主題標注。如該文章中出現「人才」、「設備」、「管理」等等體詞，而這些議題詞都屬於「企業營運」主題，因此則判斷此文本主題為「企業營運」的機率較高。

第四節　情緒傾向標注

一、　建立情感詞集

本研究使用台灣大學所開發的中文情感詞庫 NTUSD 做為種子詞集，NTUSD 包含 2810 個正向字詞和 8276 個負向字詞，共 11086 個字詞。但 NTUSD 缺乏本研究所需財經方面相關之情感字詞，如「利空」、「利多」等，因此需針對財經領域相關字詞做擴充詞庫。經由擴充後，情感詞集內含 3589 個正

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

極字詞、8476 個負極字詞，共 12065 個字詞。

二、 情緒指數計算

透過情感詞集來計算各本文之情緒分數，計算方法為透過文本中述詞比對情感詞集來判斷，若該詞為正向詞，則情緒分數+1；若該詞為負向詞，則情緒分數-1。

計算完各文本的情緒分數後，將其加總得後利用 Z-score 正規化，即將其數值減去母體平均值μ再除以標準差σ，算出文本的情緒指數。

$$\text{Sentiment Index} = \frac{Score - \mu}{\sigma}$$

依據計算出之情緒指數，對各本本進行情緒傾向標注。若情緒指數＞0 則為正向情緒，若情緒指數＜0 則為負向情緒，若情緒指數＝0，則為中立情緒。

$$\text{Sentiment Index} = \begin{cases} > 0 \Rightarrow \text{正向情緒} \\ = 0 \Rightarrow \text{中性情緒} \\ < 0 \Rightarrow \text{負向情緒} \end{cases}$$

第五節 建立分類模型

本研究分別使用 SVM、神經網路兩種技術來建立分類模型，將文本分為上漲（正向）、下跌（負向）兩類，並驗證比較兩者分類效果。

一、 SVM 分類模型

過往研究指出，SVM 在進行文本的情緒分類時比起其他分類模型有更佳的分類效果（劉羿廷，2015），故選擇使用 SVM 來建立分類模型。將標注好情感類別與主題類別之文件，作為監督式學習之訓練資料進行訓練，並驗證其分類效果。

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

二、 LSTM 神經網路模型

首先使用 word2vec 模型來訓練詞向量，再將文本以詞向量的形式作為模型之輸入。

使用 RNN 的 LSTM 模型來建立分類模型，將標注好情感類別與主題類別之文件作為訓練資料進行訓練，每次以一個詞為單位作為輸入，如圖所示。

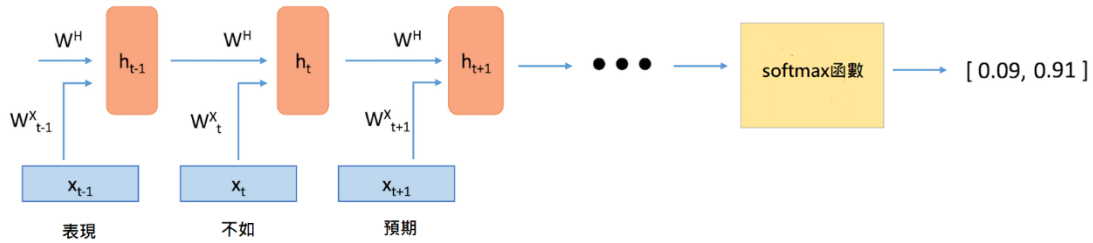

圖 11 LSTM 模型概念圖

在神經網路中如果不使用激勵函數，那麼在神經網路中皆是以上層輸入的線性組合作為這一層的輸出（也就是矩陣相乘），輸出和輸入依然脫離不了線性關係，做深度神經網路便失去意義。因此必須加入激勵函數（Activation Function）。激勵函數有許多種類，包括 sigmoid, tanh, ReLU 等，本研究使用分類問題中常使用之 softmax 函數作為激勵函數。其公式為：

$$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad for \quad j = 1, \ldots, K$$

經由 softmax 函數可以讓模型的輸出成為機率分布的形式，分別代表正向與負向之機率，並取機率大者最為分類之結果。

損失函數的選擇，本研究使用交叉熵（Cross Entropy）作為損失函數，交叉熵是一個源自於資訊理論的概念，原先是用來估算平均編碼長度。在機器學習的領域中，則常被用來預測兩個機率分布之間的差距，其公式如下：

$$\mathrm{H(p,q)} = -\sum_{x} p(x) \ log \ q(x)$$

P 為正確的機率分布，q 為預測的機率分布。透過交叉熵函數來定義模型所

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

預測的答案與真實答案之間的距離。

本研究使用了 LSTM 神經網路來進行文本情感分析，但單純使用文本情緒仍不足以能夠預測股價趨勢，因此為了優化模型，除使用 LSTM 外也加入了以技術指標所建立的前饋神經網絡(Feedforward Neural Network, FNN)，以 LSTM+FNN 相結合之神經網路模型作為本實驗之股價漲跌分類模型，如下圖所示。

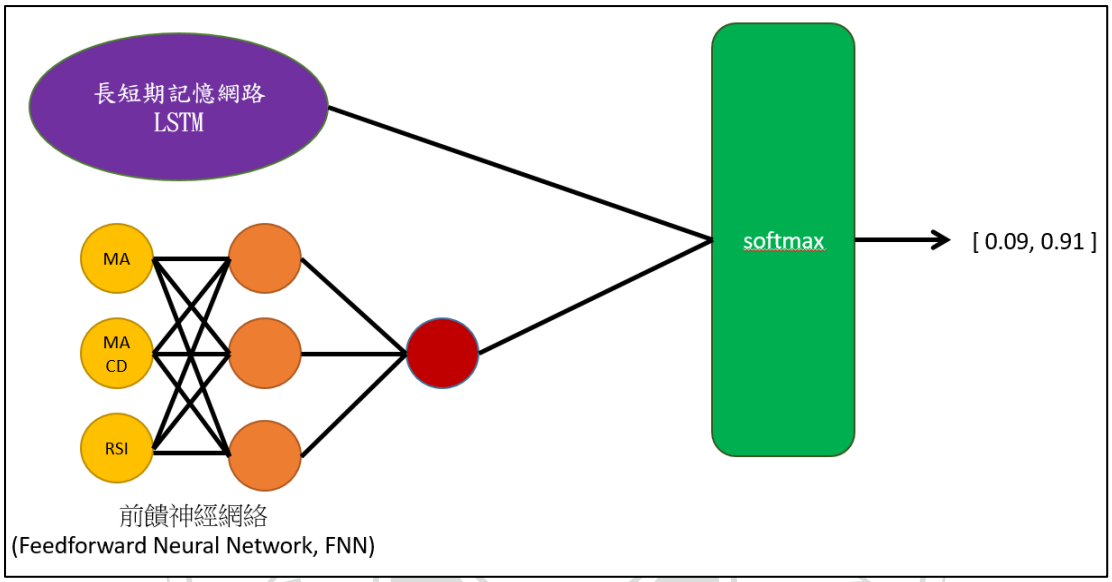

圖 12　LSTM+FNN 分類模型示意圖

三、　技術指標與間接情緒指標

為了提高分類模型預測準確度，除了考慮前面計算出的情緒指數外，亦加入一些技術指標及間接情緒指標作為輸入變數，建立分類模型，詳細如下：

(一) 移動平均（moving average，MA）

移動平均是技術分析中一種分析時間序列數據的工具，算法是將過去某段時間內的價格取平均值。利用兩條平均線中較短的一期作為訊號線，來做為股價漲跌的訊號，例如上升行情初期，短期移動平均線從下向上突破中長期移動平均線，預示股價將上漲，當短期移動平均線向下跌破中長期移動平均線，預示股價將下跌。

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

(二) 平滑異同移動平均線（Moving Average Convergence / Divergence, MACD）

MACD 指標是根據均線的構造原理，對股票價格的收盤價進行平滑處理，求出算術平均值以後再進行計算。MACD 指標是運用快速（短期）和慢速（長期）移動平均線及其聚合與分離的徵兆，加以雙重平滑運算。MACD 保留了移動平均線的效果，可以用來研判買賣股票的時機、預測股票價格漲跌的技術分析指標。

(三) 相對強弱指數（Relative Strength Index，RSI）

RSI 為一種可以判斷股市買氣與賣氣強弱的一種指標，藉由計算某一段時間上漲日的指數平均值與下跌日的指數平均值，推算股價上漲機率。RSI 大多在 30~70 之間波動，當 RSI 超過 80 時表示出現超買現象，股價可能會開始下跌；低於 20 時表示市場太過悲觀，之後可能止跌回升。

(四) 匯率

匯率可以反映我國經濟成長與國際貿易的狀況，匯率的變動會對股價的變動造成影響。本研究以美元匯率作為間接情緒指標。

(五) 原油價格

國際原油價格的波動會對全球經濟市場造成影響，間接影響各國股市，本研究將原油價格加入間接情緒指標。

(六) 台股加權指數(TAIEX)

台股加權指數是臺灣證券交易所所編製的股價指數，可用來呈現台灣的經濟走向，本研究將台股加權指數加入間接情緒指標

四、 分類效果驗證

在 Data Mining 領域中，常使用 Precision、Recall、F-measure 作為評估分類模型好壞之依據。Precision 為精確率，即預測上漲，實際結果也上漲的比率。Recall 為召回率，即實際上漲，且預測也為上漲的比率。F-measure 為整合

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

Precision 和 Recall 的結果。

		模型分類結果	
		上漲	下跌
實際結果	上漲	TP (True Positive)	FN (False Negative)
	下跌	FP (False Positive)	TN (True Negative)

表 3　混淆矩陣(Confusion Matrix)

TP (True Positive)：分類結果「上漲」，實際上也「上漲」

FP (False Positive)：分類結果「上漲」，實際上卻「下跌」

FN (False Negative)：分類結果「下跌」，實際上卻「上漲」

TN (True Negative)：分類結果「下跌」，實際上也「下跌」

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F} - \text{measure} = 2 \times \frac{Recall \times Precision}{Recall + Precision}$$

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

第四章　實驗結果與討論

第一節　財經文本資料蒐集結果

本研究使用 python 的 BeautifulSoup 與 Selenium 套件來撰寫爬蟲程式，蒐集國內知名網站如鉅亨網、ETtoday、經濟日報 2017 年 7 月 1 日至 2018 年 4 月 30 日之財經新聞文本，共 11797 篇，如下表所示。

時間	2017/07/01 – 2018/04/30		
資料來源	鉅亨網	ETtoday	經濟日報
文本數量	5254	418	6125

表 4　財經文本蒐集結果

第二節　文本主題標注結果

一、　詞向量訓練結果

透過 Word2Vec 模型的訓練後，將 84673 的不同字詞訓練成 100 維的向量。訓練好的 Word2Vec 模型，除了能夠將文字轉換成對應的詞向量之外，也能衡量兩個詞的相似程度，如下面列出「面板」這一詞所對應的詞向量，以及在所有文本中的字詞，與「面板」這個詞最相似的詞和兩者的相似度。

字詞	詞向量
面板	[0.28825834 -1.5086749 -0.90512884 3.9851687 0.51185954 -1.1335863 -4.2281904 -0.6821879 -1.0026735 4.081599 -2.0802324 0.8536029 0.94848543 -1.167045 0.74125916 -0.66381127 -0.07675572 -1.512117 -2.2375782 -2.014289

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

	-0.5969997 -4.490413 -1.9621731 -3.4338572 -2.4323812 -1.293377 1.7003638 2.149623 -1.7306939 1.1896832 0.12943096 -2.1850023 -2.647519 3.2156782 1.8841643 3.2527056 -5.1831737 -0.26691 -0.04503355 -0.6260971 0.6925822 0.6943185 1.004433 0.564042 -2.7253673 -1.862499 -5.201912 1.1214997 -1.1666199 -0.16591626 1.051902 1.00185 1.196004 -1.6545168 2.0815926 2.3229086 0.98307765 1.8902385 3.02768 -1.2888148 0.92112523 0.16941184 1.2159408 0.05491477 0.49376357 4.453815 4.192127 -1.3875039 4.285681 2.1615632 0.47571746 1.135167 -2.8749058 -0.78481287 -0.05159324 3.4657786 1.5661738 1.1935376 -0.92790353 -0.81229657 1.7844034 0.07245653 -4.760211 0.2989984 2.0340483 1.8770708 -0.8718976 4.2840734 2.6337733 -2.4908907 -2.6280317 -1.2620242 -0.3428923 -1.1583384 -0.73844504 -1.0813144 0.23927112 1.0207144 -0.28657818 0.8256913]

表 5 「面板」的詞向量

相似詞	相似度
面板廠	0.7195311188697815
電視	0.6510827541351318
顯示器	0.6222976446151733
偏光板	0.6117022037506104
韓廠	0.5812633633613586
背光	0.5739165544509888

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

尺寸	0.5617168545722961

表 6　和「面板」相似的詞

二、　LDA 主題模型

本研究使用 LDA 主題模型來找出各群體之議題詞，如下表所示。

群體 1 股市資訊				
股票	分析	成交量	整理	行情
跌幅	價差	賣權	買權	波段
群體 2 企業營運				
面板	群創	電視	合作	尺寸
董事	品牌	集團	人才	管理
設備	客戶	員工	旗下	成本
群體 3 公司與法人				
台積電	大立光	法人	族群	金融
權證	國泰	公司	布局	修正
群體 4 手機產業				
iphone	蘋果	手機	鏡頭	LED
玉晶光	光學	OLED	螢幕	三星

表 7　LDA 主題模型議題詞表

透過 LDA 主題模型將議題詞分為 4 個主題群，可發現群體 1 含有「行情」、「跌幅」、「波段」等字詞，是偏向股市資訊方面的主題；群體 2 含有「人才」、「管理」、「設備」等字詞，是偏向企業營運方面的主題；群體 3 含有「大立光」、「法人」等字詞，是偏向公司與法人的主題；群體 4 含有「蘋果」、「手機」、「鏡頭」，是偏向手機產業相關的主題。

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

三、 主題標注結果

建立 LDA 主題模型後，根據各主題群下之議題詞對文本進行主題標注，依據各主題群下的議題詞對文本進行標注，標注的結果如下表所示。

	群體 1	群體 2	群體 3	群體 4	總計
	股市資訊	企業營運	公司與法人	手機產業	
文章數量	3182	1723	4809	2083	11797
百分比	26.97%	14.61%	40.76%	17.66%	100%

表 8 文本主題標注結果

由表中可看出，四個主題類別中，主要以公司與法人相關的新聞文本占最多數，比例占所有被標注主題之文本中的 40.76%。其次為股市資訊方面的文本與手機產業相關的文本，比例分別占所有被標注主題之文本中的 26.97%與 17.66%。而企業營運方面的新聞文本數量最少，僅占所有被標注主題之文本中的 14.61%

第三節 情緒傾向標注結果

經由透過文本中述詞比對情感詞集來判斷，若該詞為正向詞，則情緒分數+1；若該詞為負向詞，則情緒分數-1。計算出情緒分數之後，進行 Z-score 得到文本的情緒指數，若情緒指數大於 0，則將文本標注為正向情緒；若情緒指數小於 0，則將文本標注為負向情緒；若情緒指數等於 0，則將文本標注為中性情緒。

	正向情緒傾向	負向情緒傾向	中性情緒傾向
篇數	8987	2209	601
百分比	76.18%	18.73%	5.09%

表 9 文本情緒傾向標注結果

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

根據表 4-6 可發現，除 5%的文本無法被情感詞庫標注情緒傾向外，大部分的文本皆有情緒傾向，證實本研究建立的情感詞庫確實有良好的情緒傾向判斷效果。

第四節　視覺化分析

利用 Python 的視覺化套件 matplotlib 將計算好之情緒指數與光電類股指數進行視覺化分析。觀察情緒指數與光電類股指數波峰與波谷趨勢轉折的時間點，是否存在領先或落後之關係，取 2018 年第一季為例。

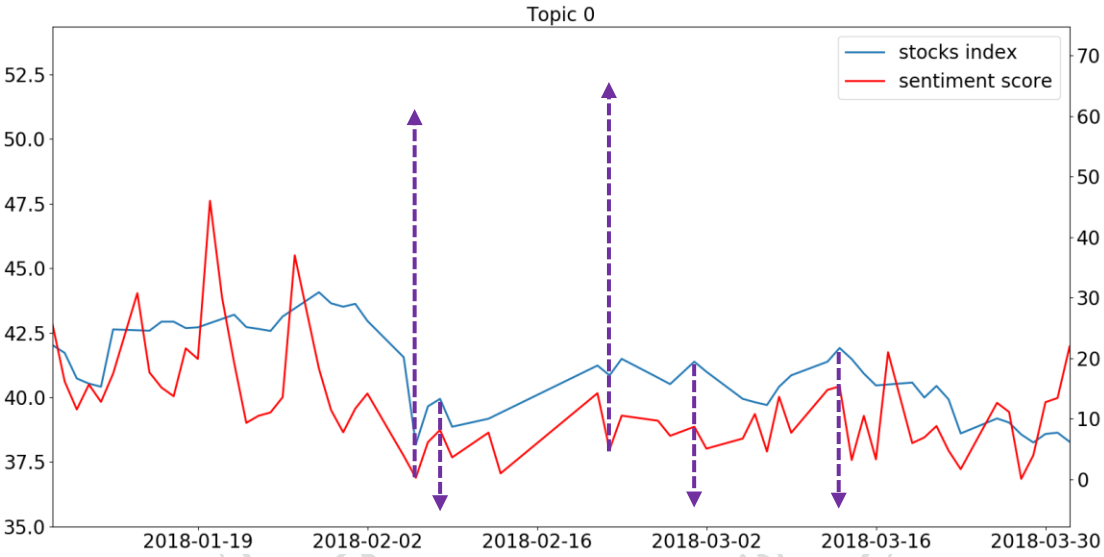

圖 13 股市資訊情緒指數與光電類股指數

上圖紅色線條為情緒指數、藍色線條為光電類股指數，觀察發現在 2 月 6 日、2 月 9 日、2 月 23 日、3 月 1 日、3 月 13 日時，兩折線位於波峰與波谷趨勢轉折的時間點幾乎一致，表示股市資訊主題文本之情緒指數轉折時間點大多與光電類股價指數一致，但並未有領先指標之特性。

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

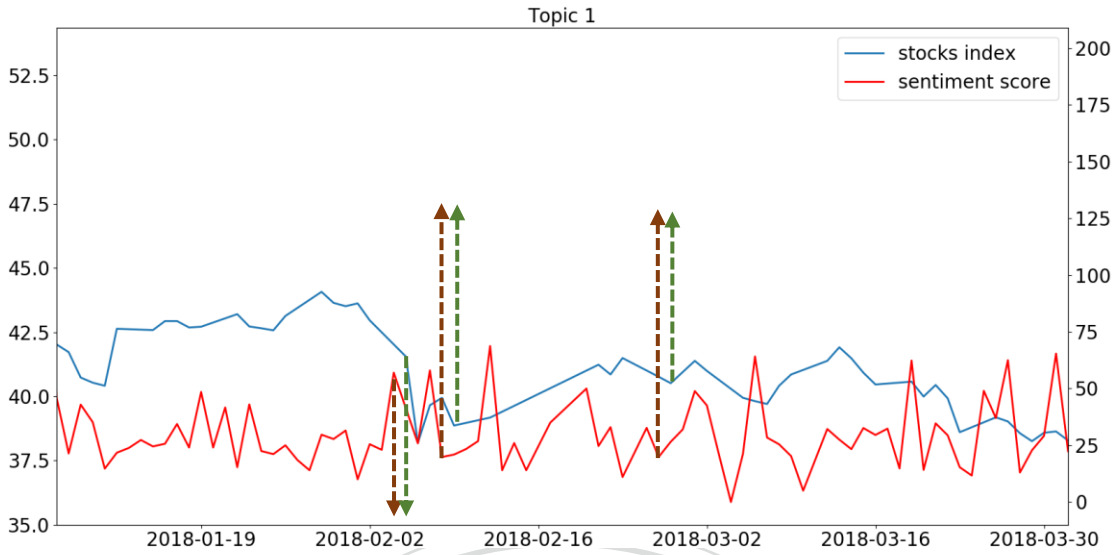

圖 14 企業營運情緒指數與光電類股指數

上圖紅色線條為情緒指數、藍色線條為光電類股指數，觀察發現在 2 月 5 日情緒指數出現轉折開始下降，2 月 6 日股價出現轉折開始下降；2 月 8 日情緒指數出現轉折開始上升，2 月 9 日股價出現轉折開始上升；2 月 26 日情緒指數出現轉折開始上升，2 月 27 日股價出現轉折開始上升。企業營運情緒指數之波峰與波谷趨勢轉折時間點約領先光電類股價指數 1 天，故此類主題之文本，有助於提升分類模型之準確度。

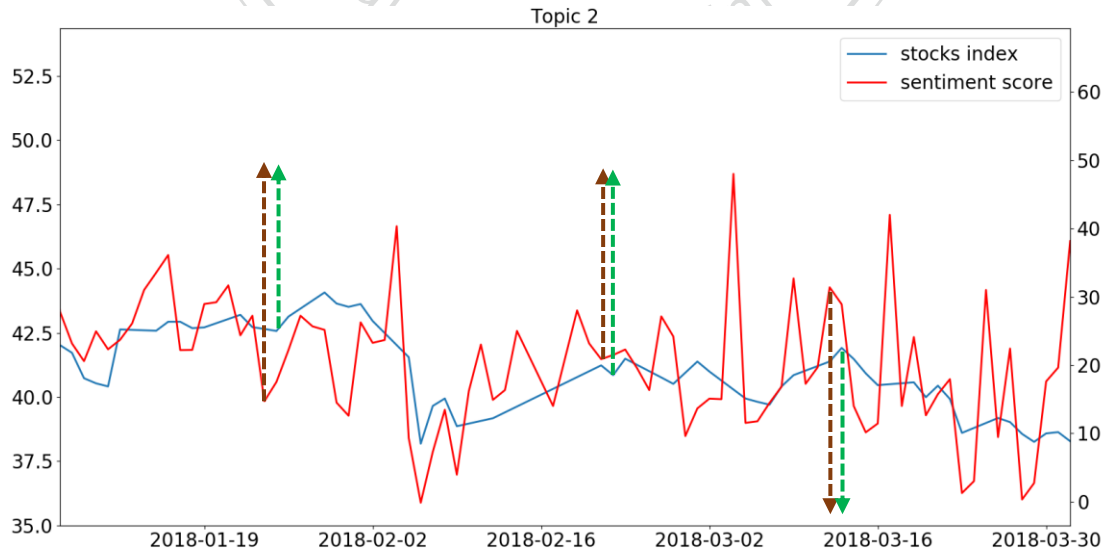

圖 15 公司與法人情緒指數與光電類股指數

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

上圖紅色線條為情緒指數、藍色線條為光電類股指數，觀察發現在 1 月 24 日情緒指數出現轉折開始上升，1 月 25 日股價出現轉折開始上升；2 月 21 日情緒指數出現轉折開始上升，2 月 22 日情緒出現轉折開始上升；3 月 12 日情緒指數出現轉折開始下降，3 月 13 日股價出現轉折開始下降。公司與法人情緒指數之波峰與波谷趨勢轉折時間點約領先光電類股價指數 1 天，故此類主題之文本，有助於提升分類模型之準確度。

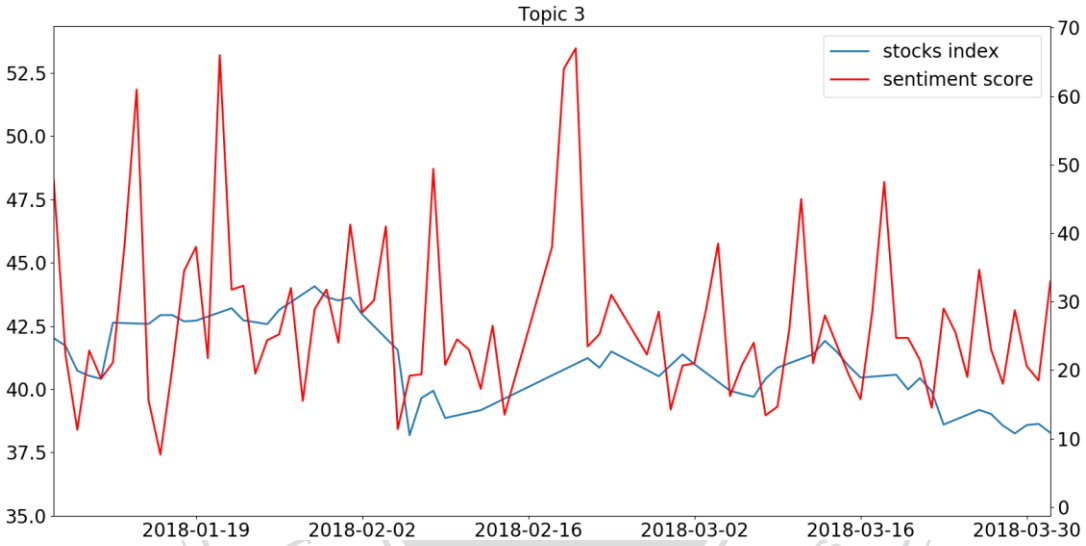

圖 16 手機產業情緒指數與光電類股指數

上圖紅色線條為情緒指數、藍色線條為光電類股指數。這類新聞文本的情緒指數看起來不具有領先指標的特性，因此對於分類模型的幫助有限，此主題文本之情緒指數較不適合作為分類模型之參數。

第五節　分類模型結果

本研究為驗證加入情緒作為參數是否對傳統單純以技術指標建立的分類模型之準確率的提升有所幫助，故分別建立單純技術指標之分類模型、技術指標結合情緒指標之分類模型，並將資料集的 70%作為訓練資料、30%作為測試資

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

料，最後透過計算 Precision、Recall、F1-score 來驗證模型的分類效果。

一、 SVM 分類模型結果

將先前所計算好之新聞文本情緒指數，搭配其他技術指標與間接情緒指標作為 SVM 分類模型的訓練資料，再分別計算其 Precision、Recall、F1-score，與純技術指標分類模型進行比較，其結果如下表。

		單純技術指標分類模型			結合情緒指數分類模型		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Topic 0 股市資訊	上漲	63%	42%	50%	67%	61%	64%
	下跌	58%	77%	66%	68%	73%	70%
	平均	60%	60%	58%	67%	67%	67%
Topic 1 企業營運	上漲	65%	53%	58%	70%	63%	66%
	下跌	65%	75%	70%	66%	74%	70%
	平均	65%	65%	64%	67%	68%	67%
Topic 2 公司與法人	上漲	65%	46%	54%	70%	63%	66%
	下跌	62%	78%	69%	71%	77%	74%

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

	平均	63%	63%	62%	**70%**	**70%**	**70%**
Topic 3 手機產業	上漲	65%	43%	52%	61%	55%	58%
	下跌	63%	82%	71%	64%	71%	67%
	平均	64%	64%	62%	63%	63%	63%

表 10　SVM 模型分類結果

由上表可知，單純使用技術指標的情況下，SVM 模型在各群的準確度為在 60~64%左右。再加入情緒指標後，優化後的 SVM 模型其準確率除 Topic3 手機產業文本外都有明顯提升，其中以 Topic2 公司與法人的文本的分類效果表現最好，準確率可達 70%。

二、　神經網路分類模型結果

將先前透過 Word2Vec 模型訓練好的新聞文本詞向量，作為 LSTM 之訓練資料，搭配其他技術指標作為前饋神經網絡(Feedforward Neural Network, FNN)之訓練資料。分別計算其 Precision、Recall、F1-score，與純技術指標的神經網路模型進行比較，其結果如下表。

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

		單純技術指標分類模型 (FNN)			結合情緒指數分類模型 (FNN+LSTM)		
		Precision	Recall	F1-score	Precision	Recall	F1-score
Topic 0 股市資訊	上漲	44%	34%	38%	53%	41%	47%
	下跌	52%	62%	56%	57%	69%	62%
	平均	48%	49%	48%	55%	56%	55%
Topic 1 企業營運	上漲	53%	40%	46%	55%	57%	56%
	下跌	51%	64%	57%	60%	59%	59%
	平均	52%	52%	51%	58%	58%	58%
Topic 2 公司與法人	上漲	54%	42%	47%	60%	47%	53%
	下跌	57%	69%	62%	59%	72%	65%
	平均	56%	56%	55%	**60%**	**60%**	**59%**
Topic 3 手機產業	上漲	46%	27%	34%	48%	49%	48%
	下跌	55%	74%	63%	57%	56%	56%

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

	平均	51%	52%	50%	53%	53%	53%

表 11　神經網路模型分類結果

由上表可知，單純使用技術指標的前饋神經網絡模型時，各群的準確度約在 48～55%左右。再加入情緒指標建立 LSTM 的神經網路模型後，各群的準確率皆有所提升，其中以 Topic2 公司與法人之文本的分類效果表現最好，準確率約為 60%。

三、　分類模型討論

本研究使用了 SVM 分類模型、神經網路分類模型兩種分類模型來進行股價漲跌之分類。這兩者在分類效果上，SVM 在準確率方面優於 LSTM 的神經網路模型。

相較於單純使用技術指標的分類模型，混和技術指標與情緒指數的分類模型在對股價的漲跌預測有更好的準確率。其中以 Topic2 公司與法人這類型的新聞文本對於分類模型之預測準確率提升最多，推測是光電類股仍舊是以大立光最具有影響力，故大立光的相關消息有其參考價值；又大立光身為股王，極高的股價非一般散戶所能購買，因此法人的投資消息也是很好的參考資訊。Topic3 手機產業這類型的新聞文本其分類模型的表現最差，推測是雖然光電類股與蘋果之間的關係密切，大立光、玉晶光等企業都屬蘋果 iPhone 鏡頭供應鏈的成員，但觀察其新聞文本，並未觀察出其具有領先指標之特性，故對於模型在預測方面的幫助有限。

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

第五章　研究結論與建議

本研究利用文字情感分析，針對 2017 年 07 月 01 日至 2018 年 04 月 30 日光電類股之財經新聞文本進行分析，並用其建立分類模型預測光電類股指數之漲跌，證實文本情緒傾向對於股價指數漲跌之預測準確度有所提升。歸納結論如下：

- 透過 LDA 主題模型，以非監督式的方法將文本進行主題分類。
- 透過比對情感詞集，計算出各文本之情感傾向。
- 利用 Word2Vec 模型，將文字型態的資料進行文字數值化，轉換成詞向量之形式。
- 以監督式的學習方法建立兩種不同的分類模型，SVM 分類模型與 LSTM+FNN 神經網路模型，來預測光電類股的漲跌趨勢。
- 就分類效果而言，SVM 模型優於 LSTM+FNN 神經網路模型。
- 各主題之文本，以公司與法人主題之文本分類效果最好，手機產業主題之文本效果最差
- 不論是 SVM 模型或是 LSTM+FNN 神經網路模型，相較於單純使用技術指標的分類模型，混和技術指標與情緒指數的分類模型在對股價的漲跌預測都有更好的準確率，最多可增加約 7%的準確率。

針對本研究不足之處，提出以下未來仍可改進與發展的方向：

- 在情緒指數的計算方面，本研究在情緒指數的計算時並未將副詞納入考量，而副詞亦能表達情緒以及情緒傾向的強度，不同的程度副詞對於情感表達的強烈程度亦有所不同，未來可考慮增加副詞的計算，使情緒指數更為精確。

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

- 神經網路模型方面，本研究在技術指標的部分是使用前饋神經網路 FNN 所建立。除了字詞的前後文是具有時間關係的資料外，技術指標本身亦是具有時間前後關係的資料類型，故未來可嘗試將技術指標的部分也使用 RNN 或 LSTM 神經網路來進行建立。

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

第六章　　參考文獻

[1] 王正豪, & 李啟菁. (2010). 中文部落格文章之意見分析. 國立台北科技大學資訊工程研究所碩士論文,

[2] 邱世芳(2008)。台灣地區光電產業之廠商衍生與空間擴散。成功大學都市計劃學系碩士論文

[3] 李啟菁. (2010). 中文部落格文章之意見分析. 臺北科技大學資訊工程系研究所學位論文, 1-44.

[4] 林育龍. (2014). 對使用者評論之情感分析研究-以 Google Play 市集為例

[5] 洪崇洋. (2012). LDA 和使用紀錄為基礎的線上電子書主題趨勢發掘方法. 國立中山大學資訊管理所碩士論文

[6] 張日威. (2014). 應用 LDA 進行 Plurk 主題分類及使用者情緒分析. 國立雲林科技大學資訊管理所碩士論文

[7] 雷祖強, 周天穎, 萬絢, 楊龍士, & 許晉嘉. (2007). 空間特徵分類器支援向量機之研究. Journal of Photogrammetry and Remote Sensing, 12(2), 145-163.

[8] 劉吉軒, & 吳建良. (2007). 以情緒為中心之情境資訊觀察與評估. Paper presented at the NCS 全國計算機會議.

[9] 劉羿廷. (2015). 運用財經文本情感分析於台灣電子類股價指數趨勢預測之研究

[10] 蕭昱維. (2014). 基於多階 LDA 技術尋找 Twitter 文章的隱含主題之研究. 樹德科技大學資訊工程系碩士班學位論文, 1-47.

[11] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[12] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In Proceedings of the fifth annual

DOI:10.6814/THE.NCCU.MIS.007.2018.A05

workshop on Computational learning theory (pp. 144-152). ACM.

[13] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. Proceedings of the National academy of Sciences, 101, 5228-5235.

[14] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[15] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis lectures on human language technologies, 5(1), 1-167.

[16] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[17] McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. The bulletin of mathematical biophysics, 5(4), 115-133.

[18] Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533.

DOI:10.6814/THE.NCCU.MIS.007.2018.A05