中華管理評論國際學報
Web Journal of Chinese Management Review

2007 年 2 月第十卷一期 • Vol. 10, No. 1, Feb 2007

運用文字探勘於日內股價漲跌趨勢預測之研究

鍾任明　李維平　吳澤民

http://cmr.ba.ouhk.edu.hk

運用文字探勘於日內股價漲跌趨勢預測之研究

鍾任明　李維平　吳澤民

摘要

如何選擇股票投資標的以取得投資利益，一直是股票投資者關心的議題，一般在選擇股票投資標的時，主要是採用基本面分析及技術面分析兩種分析方式。基本面分析主要考量上市公司的營運及財務狀況，藉以預測未來可能之盈虧，以作為選擇股票投資的依據;而技術面分析則著重過去歷史股價的變動，並從中找出股價趨勢間的特徵，藉以預測未來股價可能的漲跌趨勢，以作為股票買賣的依據。然而，不論基本面分析或技術面分析都忽略了消息面對短期股價的衝擊，故對於短期投資者而言，如何掌握消息面以便對股票買賣做出正確的決策，便成為重要的課題。

本研究針對短期股票投資議題，藉由整合歷史股價交易資料與中文財經新聞，建構台股個股日內股價漲跌之預測模型，對個股日內的股價受消息面的影響進行漲跌趨勢預測。經由實驗結果顯示，該模型利用中文新聞文件來預測台股個股日內漲跌趨勢可達到 81.48%的預測正確率，並經由模擬的股票市場交易可達到 5.33%的季平均報酬率，因此，我們認為本研究提出之方法，對於股票投資人在短期買賣的操作上具有參考的價值。

關鍵詞：文字探勘、日內股價預測、倒傳遞類神經網路

緒論

近年來，隨著經濟的蓬勃發展，各種金融商品不斷的推陳出新，包括股票、期貨、債券……等，這些金融商品的出現使得個人理財的管道也呈現多樣化的面貌，在這些已推出的金融商品中，由於上市公司的股票交易推出甚早並已廣為投資大眾所熟悉，故投資股票已成為目前重要的個人理財工具之一。

股票投資目的之一，為透過公開的市場交易機制，藉由股票的買進及賣出來賺取其間的價差而獲取利潤，雖然其操作方式包括現金買進、賣出，或是融資買進、賣出及融券買進賣出等信用交易模式，然而希望藉由股票交易來獲

鍾任明　中原大學資訊管理所碩士
李維平　中原大學資訊管理系助理教授
吳澤民　中華航空公司資訊工程師

利的目的卻是一致的，因此，如何選擇股票投資標的以取得投資利益，一直是股票投資者關心的議題。一般在選擇股票投資標的時，最常採用的分析方式主要有基本面分析及技術面分析兩種。基本面分析主要考量上市公司的營運及財務狀況，藉以預測未來可能之盈虧，以作為選擇股票投資的依據;而技術面分析則著重過去歷史股價的變動，並從中找出股價趨勢間的特徵，藉以預測未來股價可能的漲跌趨勢，以作為股票買賣的依據。然而，不論基本面分析或技術面分析都忽略了與股票上市公司相關新聞消息對短期股價的衝擊，故對於短期投資者而言，如何掌握消息面以便對股票買賣做出正確的決策，便成為重要的課題。

過去對於股票漲跌預測的相關研究，不論是使用統計或是人工智慧等方法，主要是以量化的資料作為分析來源；但投資者的決策過程，往往會被外在環境所影響，以股票交易而言，各種媒體所發佈的相關新聞皆有可能對股市造成某種程度的影響。而目前網際網路發達，越來越多的財經新聞以電子化型式呈現，且更新頻率也愈加快速，使得投資資訊愈加容易取得，然因資訊來源眾多，投資者通常沒有足夠的時間與能力來分析這些大量的資訊，並轉化為有效的投資決策，因此，如何有效率地分析這些資訊並轉化為投資決策，是目前亟需研究的議題。

目前有許多的財務服務公司提供許多財務訊息於網路中，例如：Wall Street Journal（www.wsj.com）、Financial Times（www.ft.com）提供每日相關的財經相關議題，Reuters（www.investools.com）、Dow Jones（www.asianupdate.com）則是提供即時的新聞資訊和股票、債券與期貨的量化資訊；而國內部分如 Yahoo!奇摩股市（tw.stock.yahoo.com）等都有提供類似的即時新聞資訊服務。儘管先前已有相當多的研究以非結構化資料（新聞文件、專家評論、網路論壇等）來預測個股當日的股價趨勢，然而這些研究主要為針對國外股市進行預測，而台灣與國外股市之結構並不盡相同，且在預測準確度上，仍有可提升的空間。因此，本研究希望探討運用於中文環境與台灣股市預測模型的可行性，並藉由模擬的股市交易結果，驗證本研究提出之預測模型應用於台灣股市短期投資的實用價值。

文獻探討

文字探勘

傳統的資料探勘主要是針對結構化的資料進行探勘，但以半結構化（semi-structured）或非結構（un-structured）格式儲存的文件資料，隱藏著許多重要的資訊，其重要性不容小覷，使得文字探勘技術成為近年重要的研究領域之一。文字探勘（Text Mining）是從半結構化或非結構化的文件當中，發掘出文件中隱含的、有意義且重要的資訊，透過分析文件、特徵擷取的過程，從中粹取出隱性資訊，進而處理儲存成為可被再用的知識。目前的應用相當廣泛，包括自動化分類技術（Automatic Classification）、網頁探勘（Web Mining）與文件分群（Document Clustering）等，這些找出文件中所包含樣式（pattern）的相關技術，都是文字探勘技術應用上相當普遍的例子。

現今新聞媒體充斥著許多以文字形式儲存的電子資料，以國內而言，經濟日報或是奇摩新聞等，不論是經濟面、政策面或是技術報告等都有大量的新聞發佈，然而這些資料雖然隱含相當有價值的資訊，卻無法透過一般的方法直接分析取得。Sullivan（2001）將文字探勘定義為一種編輯、組織及分析大量文件的過程，主要為提供特定使用者（如：決策者、分析師）特定的資訊（如：摘要、關鍵字），以及發現某些特徵及其關聯。近年來文字探勘的相關研究已有相當的發展，但因中文文件的句法結構與用詞皆不同於英文語系，也不像英文有明顯的分隔符號，故本研究整合中文斷詞處理與關鍵詞彙萃取策略，再配合文字探勘相關技術的運用，包括：關鍵字取樣、相關的文件自動分類方法等，結合個股股市分時資料來做台灣股市個股價格的趨勢預測。

新聞資料於股價預測之相關研究

Ahmad, Oliveira, Manomaisupat, Casey & Taskaya（2002）的研究指出，影響財務市場的資訊通常經由電子郵件、新聞、公司簡報與企業年度報告等形式發佈。研究中認為不論其資訊來源形式為何，新聞消息中所隱藏的資訊，對制定投資策略而言是相當重要的元素，如圖 1 所示。

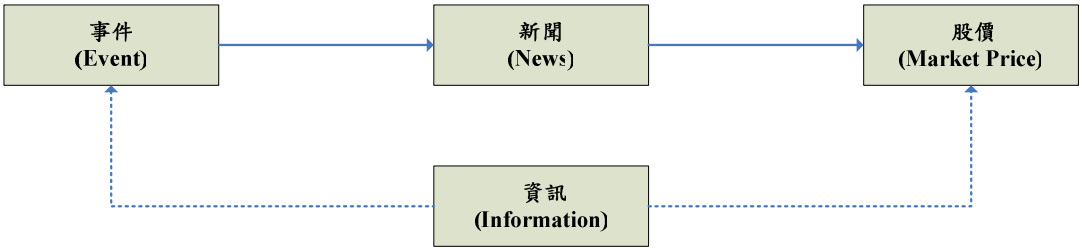

圖 1 資訊對事件、新聞、股價之互動關係圖

Wuthrich, Cho, Leung, Permunetilleke, Sankaran, Zhang & Lam（1998）選定五個主要股市指數作為趨勢預測之標的，有美國道瓊工業平均指數（Dow Jones Industrial Average）、香港恆生指數（Hang Seng Index）、日本日經指數（Nikkie 225 INDEX）、新加坡海峽時報指數（Singapore Strait Times Index）、倫敦金融時報指數（Financial Times 100 Index）等不同區域之國家股市；透過代理人（agent）自相關之專業財經網站，在股市開盤前蒐集大量且即時之財經新聞，以數種文字探勘技術 （k-Nearest Neighbor、neural network）作為分析，用以預測香港當日即將開盤的股價趨勢及可能的收盤價格（closing price）。預測結果為下列三種，股價為上漲趨勢（漲幅超過 0.5%）、下跌趨勢（跌幅超過 0.5%）或是持平（介於 0.5%～-0.5），研究結果證實平均準確率（average accuracy）比隨機投資策略的效果要好（46% > 33%）。其架構的特色在於分析文件前，先經由該領域的專家學者或投資分析師，訂出約 400 個與股價漲跌可能的關鍵字組合，作為後續訓練分類器時的詞庫，缺點在於需要事前人工建立篩選關鍵字。

在預測當日股價趨勢部份，Gidófalvi（2001）研究中利用 naïve Bayesian 文字分類器（text classifier）將新聞作漲跌的分類，研究中提出 *window of influence* 的概念，指出新聞中所包含的資訊在一定的時間間隔，會對股匯市造成相當程度的影響；對於新聞造成股票的波動部份，使用 ß 值（ß-value）加以評量與量化，該值能透過線性迴歸的方式加以計算而得，來預測中長期的股價趨勢。Mittermayer （2004）應用支援向量機（Support Vector Machine, SVM）來對新聞做分類，預測新聞在發布後 60 分鐘內的漲跌幅度與趨勢，SVM 為二元分類器，負責辨識好新聞（Good News）和壞新聞（Bad News），其餘的則被分類成不會對股價造成波動的新聞（No Movers），交易引擎在接受到系統的分類結果後，會產出對應的交易建議（買進或賣出），結果證明透

過該研究架構的每次交易的平均獲利(Average Profit)大於隨機投資(Random Trader)。

Lavrenko, Schmill, Lawrie, Ogilvie, Jensen & Allan(2000)提出 Language Model 的概念，來辨識對股價趨勢有影響的字詞，例如 loss（損失）、shortfall（虧損）和 banruptcy（破產）都與下跌趨勢有高度的相關性，反之像 merger（企業併購）、acquisition（收購）和 alliance（企業聯盟）等字詞則可能對股價有正面的訊息，透過訓練與建立 Language Model 來辨識這些與股價趨勢相關的詞彙，可協助股價趨勢之預測。

Fung, Yu & Lam (2002, 2003)則結合數種資料探勘與文字探勘技術來建立預測模型，提出 t 檢定為基礎(t-test based)的演算法來切割股價趨勢(漲與跌)，特色為使用兩個 SVM 分類器，一為專門辨識好消息（Good News）的新聞，另一個則用來辨識壞消息（Bad News）的新聞，其餘則為影響不大的新聞文件，並對以往的權重值計算公式加入 inter-cluster discrimination coefficient（CDC）和 intra-cluster similarity coefficient （CSC）做相似度的判斷，計算方式如公式（1）與公式（2）所示，且該預測模型不需要有特定區間的時間序列資料，交易策略為買進持有(Buy-and-Hold)的投資策略。

$$CDC = (\frac{n_{i,t}}{N_t})^2 \qquad (1)$$

$$CSC = \sqrt{\frac{n_{i,t}}{n_i}} \qquad (2)$$

由上述的文獻歸納可得新聞與股價波動的連動性，交易市場是一個有效率的訊息處理機制，可以立即將所發生的資訊吸收消化後反應至股價上，相較於國外相關文獻與學者的論點，傾向支持美國的股票市場具有弱式效率性，至於國內的台股市場，由於各學者取樣對象與時間差異，至今仍無法對台股是否具弱式效率性市場作一明確的結論，因此本研究假設在效率市場假說成立下，新聞會對應的反映在股價的漲跌上，投資者可以透過分析這些資訊，來獲取超額的報酬。

倒傳遞類神經網路

葉怡成(2000)指出，類神經網路是「一種基於腦與神經系統研究所啟發的資訊處理技術」，它可以利用一組輸入和輸出資料來建立系統模型，用來解決

分類、預測等問題。傳統上股價預測的相關研究，大多採用歷史股價資料訓練模型以預測未來之股價趨勢，鮮少有以文字資料為輸入之預測模型。

而倒傳遞類神經網路基本原理是利用目標輸出值（Desired Output）與實際輸出值（Actual Output）之間的差異，利用梯度遞減法（Gradient Descent Method）的觀念將誤差函數（Error Function）予以最小化，即每當輸入一個學習的例子時，網路即小幅調整權重值的大小。倒傳遞類神經網路的學習過程包含兩個階段，分別為順向傳遞（forward pass）與逆向傳遞（backward pass）；順向傳遞是從輸入層開始，一層一層向前傳遞並計算各層處理單元的輸入值，然後連接權重，並經由轉換函數得到該層各處理單元的輸出值，該輸出值將成為下層各處理單元的輸入值，以此類推，直至網路的最後一層。逆向傳遞則是由輸出層向後傳遞，這一階段在於計算誤差及更新連接的權重，其權重更新方法為將前一層的誤差值向後傳遞，並以此為依據修改連接權重，接著計算該層的誤差，再將其往後傳遞，如此逐層往後傳遞計算修改權重，而倒傳遞類神經網路架構如圖 2 所示。

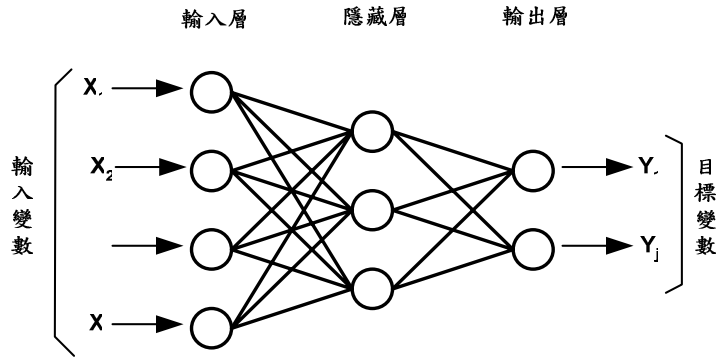

圖 2　倒傳遞網路架構

其中隱藏層中的神經元，亦稱為處理單元，為組成類神經網路的最基本單位。主要是接受網路前端的輸入值並彙總後輸出，該神經元主要包含集成函數（Summation function）與轉換函數（Transfer function），如圖 3 所示。

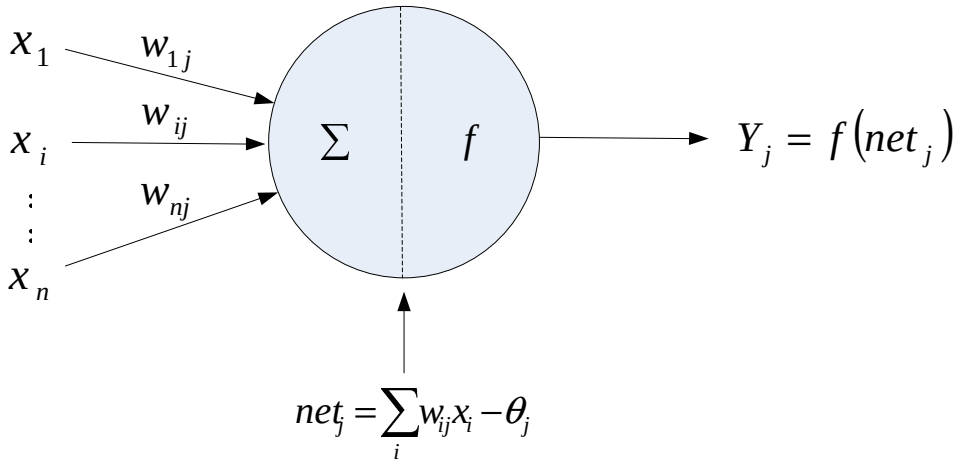

圖 3 神經元模型架構圖

本研究所採用之倒傳遞演算法的網路架構為多層前饋網路（Feed-forward backprop），如圖 3 所示，該神經元為一具有 ***n*** 個輸入的基本神經元，每一個輸入都用一適當的權重值 ***w*** 來加權，而加權後輸入與偏權值總和後形成轉換函數 ***f*** 的輸入，轉換函數則採用對數雙彎曲轉移曲線（Log-Sigmoid Transfer Function），計算方式如公式（3）所示，說明了輸入 ***n*** 可以是在正負無限間的任意值，透過計算後能將類神經的輸出 ***a*** 之範圍限制在 0 到 1 之間。

$$a = \frac{1}{1+e^{-n}} \tag{3}$$

倒傳遞網路演算法是一種廣義化的最小均方演算法（LMS），且這兩種演算法都是使用均方誤差（Mean Square Error, MSE）作為網路性能函數（Performance function），且倒傳遞網路為一監督式演算法，必須給予網路一連串對應的輸入值 ***p*** 與目標輸出值 ***t***，如公式（4）所示：

$$\{p_1,t_1\},\{p_2,t_2\},\ldots,\{p_Q,t_Q\} \tag{4}$$

其中，p_q 是網路的輸入值，而 t_q 則是與輸入相對應的目標輸出值。

當每一個輸入值 p_q 進入網路後，經過轉換函數的計算所得到的輸出值 a_q 便可以與目標輸出值 t_a 做比較，t_a 與 a_q 相減即為誤差，使得所有誤差平方總和的平均極小化，計算如公式（5）所示：

$$mse = \frac{1}{Q}\sum_{k=1}^{Q} e(k)^2 = \frac{1}{Q}\sum_{k=1}^{Q} (t(k) - a(k))^2 \quad (5)$$

研究設計

本研究整合自動化分類與中文斷詞技術，以不同的詞性組合策略，分析文字中所隱含的資訊，用來預測個股日內股價漲跌，模型以台股中的電子類股為對象，統計過上市上櫃的相關新聞資料與漲跌情形之後，選擇台積電作為預測模型的標的；整體之概念為利用個股分時資料與對應的新聞文件作為輸入資料，透過處理之後建立一股價趨勢預測器，預測未來新進新聞與股價漲跌趨勢之連動性如圖 4 所示，而整個系統概念可分為兩大階段，與其細部的子系統與相關的外部元件或資料庫，系統架構圖如圖 5 所示。

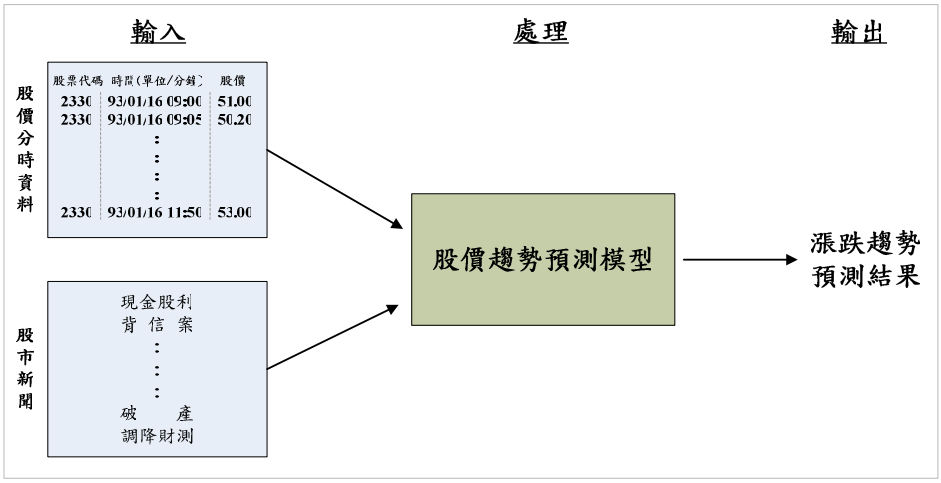

圖 4　研究概念圖

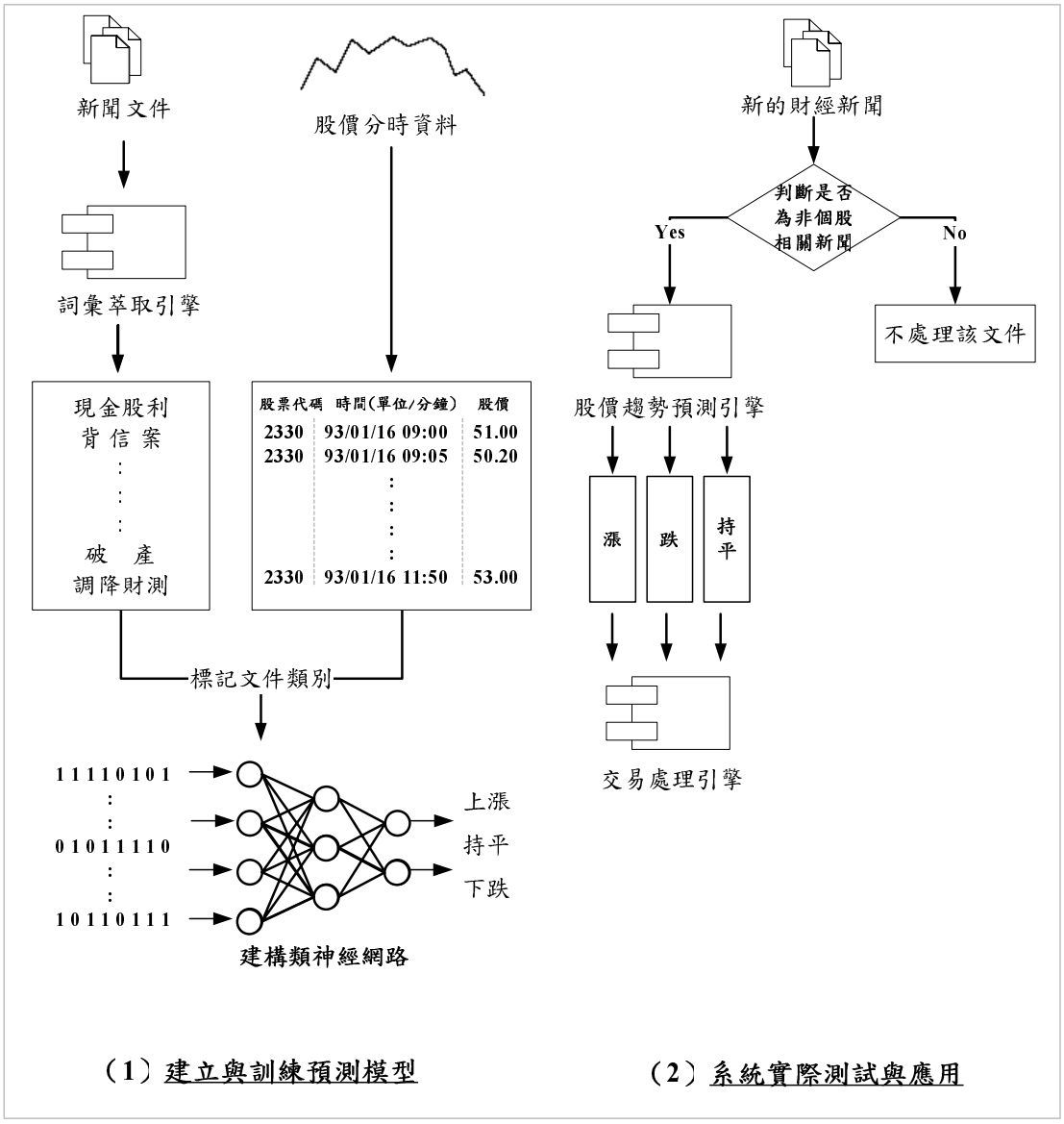

圖 5　系統設計架構圖

實驗資料收集

預測模型將整合中文新聞與股價分時資料建立個股股價漲跌趨勢預測模型，數值資料由台灣經濟新報文化事業（股）公司整理與收集，其包含了自 1999 年 5 月起每一分鐘指數的變化及委買委賣筆數與張數；中文新聞資料部分則是採用 Yahoo!奇摩股市新聞的個股相關新聞，實驗資料收集的時間範圍為 2004 年 10 月 1 日至 2005 年 3 月 31 日，訓練資料與測試資料的分佈如圖 6 所示；樣本數過少原因在於各家主要之電子媒體所發佈的新聞，時間資訊上只有發佈日期，但由於研究架構為為日內股價受新聞影響之波動情形，因此必須精確到以分鐘為單位，並於取樣上配合下列的限制：

- 新聞的發布時間限定於台股交易時間內（9：00AM～13：30PM）。
- 新聞中必須包含所選定實驗個股之股票代號（ticker symbol）或是同義的公司名稱（台積電、台灣積體電路製造股份有限公司或 TSMC 等）。
- 不能出現兩個或超過兩個的股票代號。

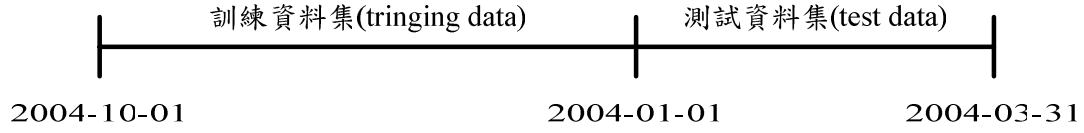

圖 6 實驗資料說明圖

詞彙萃取引擎

語彙萃取引擎部分，則是必須先將實驗文件輸入 CKIP 中文詞知識庫系統後，經過斷詞處理與詞性標記以解析出個別的字詞，而篩選出關鍵字詞可視為該篇文件之索引與重要的概念，並且配合股價資料來建立與訓練類神經網路；根據林厚誼、蔣岳霖與周世俊（2002）、許正欣（2004）的研究指出，由於中文斷詞後的候選詞太多，且單一的字詞較無意義，透過目前已整理出的幾種可能、常見的詞類序列之 Heuristic Rule 為基準，透過合併詞彙的方式降低候選詞的維度，並形成有意義的各類片語。下表 1 舉例說明詞性組合的範例。

表 **1** 詞性組合規則範例

組合規則	組合後詞性	組合範例
A + Na	Na	信託（A） + 股票（Na）
Na + Na	Na	現金（Na） + 股利（Na）
Nb + Na + Nc	Nb + Na	台積（Nb）+ 電（Na） + 公司（Nc）
Nc + Nc + Nc	Nc + Nc	中央研究院（Nc）+ 語言所（Nc）+ 語音實驗室（Nc）

建立「詞彙—文件矩陣」

本步驟先萃取出文件集內的辭彙及出現次數（Count）值。再藉由選擇詞彙權重或是詞彙次數值之大小，以決定每篇文件中將以哪些詞彙做為該文件的關鍵詞彙，最後將各文件以 Salton, Wong & Yang（1975）提出的向量空間法

來表示。在訓練與建立預測模型之前，需建立一「詞彙—文件矩陣」之向量空間，將文件特徵做為類神經網路的輸入資料。研究中將透過不同之門檻值來取出不同數量集合之關鍵詞彙，當某關鍵詞彙出現時，則標示為 1，反之則為 0，透過二元標記的方式表達該文件之特徵，如圖 7 所示。

	T1	**T2**	**…**	**Tn**
D1	T11	T21	…	Tnm
⋮	⋮	⋮	⋮	⋮
Dm	T1m	T2m	…	Tnm

圖 **7** 詞彙－文件矩陣

標記文件類別

倒傳遞類神經網路屬於監督式演算法，需要有明確的目標輸出變數，新聞文件雖然透過前置處理步驟轉換為結構化資料，仍須與歷史股價做配對，作為新聞漲跌標記的來源，以下透過範例說明預測模型的輸出入變數，並假設股價對新聞的反應時間為一小時，這時如果有一篇相關新聞在交易時間內發佈，將會比對所萃取出的關鍵字庫，用二元的方式表示成文件的特徵索引，如下所示：

$$D_1 = （101101 \cdots 011001）$$

接著系統在股價歷史資料庫中找尋滿足反應時間下，有撮和成功的交易紀錄，並配合漲跌條件的定義來判定漲跌，例如漲跌幅度大於 0.25%者，判定為漲，漲跌幅度在-0.25%～0.25%之間者為持平類別，漲跌幅度大於-0.25%者判定為跌，假設 D_1 的發佈時間為 2005/01/31 09:04:16，當時個股股價為 50.50，而一小時之後的最近撮合時間為 2005/01/31 10:04:32，當時個股股價為 51.50，股價兩相比較之下上漲了 1.98%，因此在所設定的漲跌幅度條件下，D_1 被表示成會對股價造成正面影響的新聞，輸出變數部分採用 3bytes 來表示上漲、持平與下跌情形，如下所示：

$$D_1 = （101101 \cdots 011001：100）$$

透過上述步驟可以將 D1 標記為對股價能造成上漲影響的類別。

建立倒傳遞類神經預測模型

研究模型中整合了文字與股價的量值資訊，而前文所述之詞彙—文件矩陣代表的新聞文件所內含的資訊，接著依據每篇新聞的發布時間，將對應的量值資訊作配對後，作為類神經網路的輸入資料（*Xi*）。而網路的輸出部分為三個類別（class），即為上漲、下跌與持平類別。上漲類別的定義為，在新聞發布後特定時間內，股價漲幅大於 0.25%以上者；反之，如果新聞發布後特定時間內，股價對應下跌幅度大於-0.25%以上者，則歸為下跌類別，若漲跌幅度居於-0.25%～0.25%者則分類為持平類別。

預測模型評估指標

一般而言，在評估分類模型的預測能力時，常以如表一的混亂矩陣(confusion matrix)來比較模型的個別與整體的正確率，藉以衡量預測的效果。並訂定投資獲利機率與投資損失機率來輔助評量模型的好壞，指標詳細定義如下。

- 投資獲利機率

  預測模型所造成的可能獲利為，將上漲類別正確的預測為上漲，以及將下跌類別正確的預測為下跌，上述兩類新聞數量總和與總測試新聞數量之比例，但持平類別不會造成獲利之情形，不列入計算。

- 投資損失機率

  預測模型所造成的可能損失為，將上漲類別錯誤的預測為下跌，以及將下跌類別正確的預測為上漲，上述兩類新聞數量總和與總測試新聞數量之比例，而持平類別的錯誤預測不造成投資損失，不列入計算。

模型假設的交易環境為，投資者手上沒有任何的持股，且當系統預測出買賣訊號時，都能立即的買賣手上的持股，交易方式為融資融券並包含真實之相關手續費用，透過系統來預測該股的漲跌趨勢，評估其獲利能力。

實驗評估與相關研究比較

實驗說明

本研究之目的在於如何應用文字探勘技術與自然語言處理架構，挖掘中文新聞中的潛在資訊以預測當日股價的漲跌趨勢。以下實驗中將調整可能影響預

測能力的參數－股價反應時間、關鍵資訊擷取門檻值類型與關鍵詞彙詞性組合類型，分別設計不同的實驗，提出較佳的參數組合模型，並應用於台灣之股市中做模擬交易，如圖 8 所示。

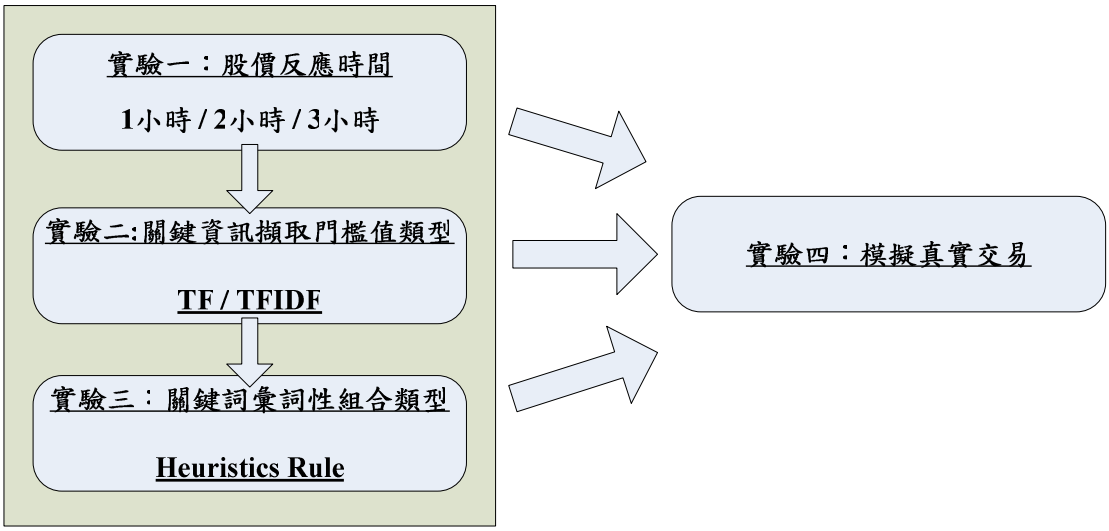

圖 8　實驗流程與相關參數說明圖

台股對於新聞的反應時間

本研究主要為探究財經新聞對於日內個股股價之影響，因此股價反應時間長度的差異，會造成交易時間內可用的文件數量集總數的差異，且反應時間設定越長，可能會造成此段時間內股價的波動反應數次，研究將在實驗中探討不同長度的反應時間對於預測正確率的影響。

關鍵資訊擷取門檻值類型

一般來說在關鍵資訊的擷取門檻值類型上，主要可區分為以關鍵詞彙出現次數作門檻值，或是以關鍵詞為權重值作為門檻值，以下依序介紹兩種門檻值。

(1)*TF*（關鍵詞彙出現次數）：以該詞彙出現次數為準，只擷取出現次數超過門檻值的詞彙當作關鍵資訊。

(2)*TFIDF*：由 Salton（1983）提出，以詞彙依出現次數換算成權重後為準，只擷取權重值高於門檻的詞彙。

由於出現次數多的詞彙不見得權重會比較高，反之亦然，實驗中將採用不同類型的門檻值設定，比較不同門檻值類型的分類效果。

關鍵詞彙詞性組合類型

由於研究是採用中文新聞文件，因此在斷詞過程之後，根據詞性規則做合併詞彙以萃取出可能的關鍵詞彙，此步驟是依照目前已整理出的幾種可能的、常見的詞類序列之 Heuristic Rule 為基準，將符合組合類型的候選詞加以合併，將比較不同組合規則對模型的影響。

實驗分析

實驗一：台灣股市對於新聞反應漲跌的時間

由於台灣股市結構與成熟度皆不同於歐美股票市場，過去有關新聞與日內股價漲跌的研究中，均設定其股價對新聞的反應時間為一小時，或是探討一小時內的漲跌幅度，但台灣股市未必適合此種反應時間的設定方式，故在實驗中，將檢驗不同的反應時間下，對預測個股股價漲跌的效果。以下依序介紹反應時間為一小時、兩小時與三小時預測結果。

股價反應時間設定為一小時

在此實驗組中，其整體平均預測正確率可達 76.00%，對於上漲類別的預測能力約六成左右，預測結果中分類錯誤到下跌類別的情形居多；而持平類別的預測正確率可達 83.33%，也出現分類至下跌類別的錯誤；而下跌類別的預測正確率為 71.43%，其中有少數的新聞被判定為上漲類別；由實驗結果可觀察到模型對於分辨上漲與下跌類能力較差，需要再加強這兩類別的訓練樣本比重。表 2 列出各種類別在這組實驗的預測正確率與整體平均預測正確率。

表 **2**　反應時間爲一小時之混亂矩陣

預測 / 實際	上漲	持平	下跌	總計	目標正確率	整體正確率
上漲	**4**	0	2	6	66.67%	76.00%
持平	0	**10**	2	12	83.33%	
下跌	1	1	**5**	7	71.43%	
總計	5	11	9	**25**		

股價反應時間設定為二小時

反應時間為二小時的實驗組與反應時間為一小時的實驗組相比，整體預測正確率有明顯的下降，預測上漲類別中，有將近 58.33%的文件被預測錯誤至持

平的類別；而持平類別有 62.50%的正確率；下跌類別是此時實驗組中最高的預測正確率類別，可達 66.67%；當反應時間拉長至二小時候，漲跌預測器只能較正確的預測下跌類別，其他類別的表現則較差，如表 3 所示。

表 **3**　反應時間爲二小時之混亂矩陣

預測 / 實際	上漲	持平	下跌	總計	目標正確率	整體正確率
上漲	**5**	7	0	12	41.67%	53.85%
持平	2	**5**	1	8	62.50%	
下跌	1	1	**4**	6	66.67%	
總計	8	13	5	**26**		

股價反應時間設定為三小時

反應時間三小時的個股漲跌預測實驗結果如表 4 所示，整體預測正確率相當的低，上漲類別多數被歸類到下跌類別，而持平類別也呈現混亂的預測結果，下跌類別雖有較高的預測正確率，但因測試文件數過少，無法驗證模型的有效性。

表 **4**　反應時間爲三小時之實驗結果

預測 / 實際	上漲	持平	下跌	總計	目標正確率	整體正確率
上漲	**0**	1	2	3	0.00%	42.86%
持平	1	**3**	3	7	42.86%	
下跌	1	0	**3**	4	75.00%	
總計	2	4	8	**14**		

從實驗一的結果可發現在不同的反應時間設定下，漲跌預測模型的效果有明顯的差異，在整體預測正確率方面，以一小時的股價反應時間為佳。探究其原因可能為採用不同的反應時間區間，其測試的文件集總數會隨著反應時間越久而減少，由於訓練文件集不夠多，導致在三小時的反應時間實驗組中，預測效果差強人意。各實驗組比較如圖 9 所示。

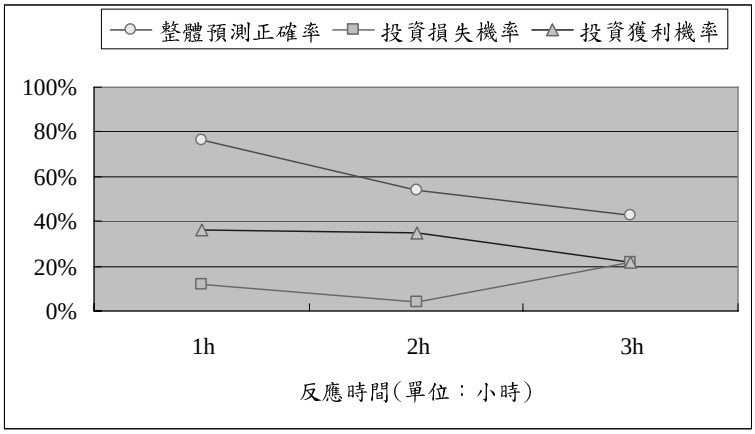

圖 **9**　不同反應時間下對預測正確率之影響

實驗二：關鍵資訊擷取門檻值類型

一般而言，用以表達文件的關鍵詞彙越多，越能描述文件的特徵，但越多的關鍵詞彙，卻也可能包含較多的雜訊，然而過少的關鍵詞彙，則又不足以表示該文件的特徵，無法明確的描述文件之意涵，因此關鍵詞彙的選擇是影響模型預測能力的重要因素。在關鍵詞彙的篩選技術方面，TF 表示某一詞彙出現的次數，當出現次數愈多則代表重要性愈高，愈能做為該文件的代表，如公式 4 中的 $\#\left(t_k, d_j\right)$。但若這個詞彙同時出現在許多篇文件中，則表示這個詞彙出現的範圍太廣泛而變得不具代表性。因此，在進行關鍵詞彙篩選時，可將反向文件頻率（Inverse Document Frequency, IDF）列入考量，亦即當一個詞彙在單一文件中出現很多次，且只出現在少數幾篇文件中，則這個詞彙將具有較高的權重。實驗將比較以 TF 與 TFIDF 為基準，在不同的關鍵資訊擷取門檻值類型下，對於預測效果的影響，其中 TFIDF 的計算公式(6)如下所示。

$$TFIDF\left(t_k, d_j\right) = \#\left(t_k, d_j\right) \times \log\left(\frac{Tr}{\#Tr\left(t_k\right)}\right) \tag{6}$$

以關鍵詞彙出現次數作為門檻值

由表 5 與圖 10 的實驗結果可知：以關鍵詞彙出現次數作為門檻值調整預測模型，確實會對其預測效果造成影響。當 TF>=1 時，由於關鍵詞彙總數過多，相對的包含較多的雜訊，造成該模型沒有預測上漲類別的能力；而 TF>=3 時，關鍵詞彙太少，無法充分表達該文件的特徵，雖其下跌預測能力達 100%，但可以發現其他類別的預測也偏向下跌類別，表示無法正確的判斷該

文件所隱含的漲跌資訊；TF>=2 的實驗中，雖然上漲類別的訓練文件相對而言較少，但該模型還是有 60.00%的預測正確率，且其他類別的預測正確率也相當的高，實驗中證實當關鍵詞彙出現門檻值設定為 2 時，該預測模型之整體預測正確率高達 82.86%。

表 **5**　不同 ***TF*** *設定*之混亂矩陣**(A:TF>=1, B:TF>=2,C: TF>=3)**

預測 / 實際	上漲			持平			下跌			總計	目標正確率%			整體正確率%		
	A	B	C	A	B	C	A	B	C		A	B	C	A	B	C
上漲	0	3	0	2	1	1	3	1	4	5	0	60	0	68.57	82.86	65.71
持平	2	0	0	16	17	13	2	3	7	20	80	85	65			
下跌	0	0	0	2	1	0	8	9	10	10	80	90	100			
總計	2	3	0	20	19	14	13	13	21	35						

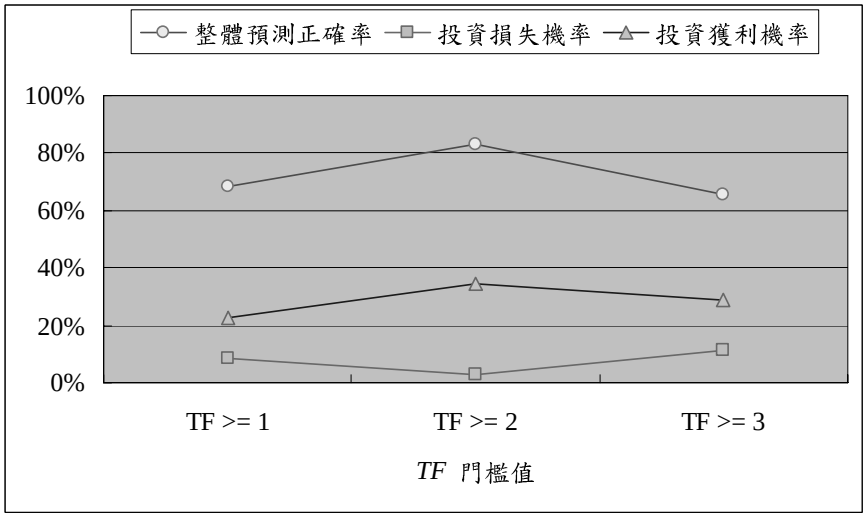

圖 **10**　不同 ***TF*** 門檻值對預測正確率之影響

以關鍵詞彙權重作為門檻值

由表 6 與圖 11 可發現，以關鍵詞彙權重做為門檻值的預測效果，整體而言沒有以關鍵詞彙出現次數作為門檻值的效果好，且預測正確率起伏相當大，對於上漲類別的預測能力最差，在三次的實驗中只有命中一篇，其餘多錯誤

分類到下跌類別，推測原因可能是訓練文件集中，上漲類別的文件較少所致，然而對於持平與下跌類別的預測效果卻沒有顯著的差異，由這次的實驗可得到與先前學者研究相異的結果，在國外的文獻中，實驗大多以關鍵詞彙權重來挑選關鍵詞彙，以建立文件索引來進行分類的動作，但本次實驗可以發現，在中文環境下以關鍵詞彙出現次數來挑選關鍵詞彙，會有較佳且平均的預測效果。

表 **6**　不同 ***TFIDF*** *設定之*混亂矩陣**(A:TFIDF>=1, B:TIDFF>=3,C:TFIDF>=4)**

預測 / 實際	上漲			持平			下跌			總計	目標正確率%			整體正確率%		
	A	B	C	A	B	C	A	B	C		A	B	C	A	B	C
上漲	0	1	0	2	3	0	3	1	5	5	0	20	0			
持平	1	1	0	15	14	20	4	5	0	20	75	70	100	65.71	68.57	85.71
下跌	0	0	0	2	1	0	8	9	10	10	80	90	100			
總計	1	2	0	19	18	20	15	15	15	35						

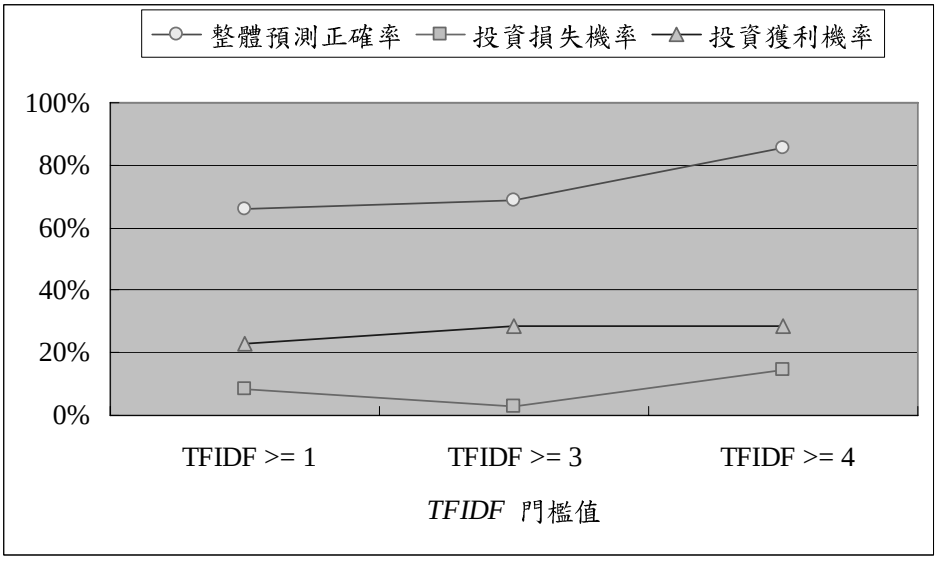

圖 **11**　不同 ***TFIDF*** 門檻值對預測正確率之影響

從上述的實驗結果中，可發現以詞彙權重當作門檻值時，雖整體預測正確率最高可達 85.71%，但細部觀察可以發現，模型對於上漲類別較沒有顯著的區

別能力，而以關鍵詞彙出現次數作為門檻值時，在門檻值為 2 時，有高達82.86%的預測正確率，且對於各類別都有顯著的預測效果。以研究模型的實用性與經濟性來歸納時，可以發現會導致投資人損失的投資決策為：當該新聞為下跌類別卻錯誤預測成上漲類別、當該新聞為下跌類別卻錯誤預測成持平類別、與該新聞為上漲類別卻預測為下跌類別時，但在這六組的實驗中，可以發現以關鍵詞彙出現次數的門檻值為 2 時，該模型可能造成的投資損失機率為 5.71%，相對於以關鍵詞彙權重門檻值大於等於 4 時的實驗結果，該模型可能造成的投資損失機率為 14.29%，因此可以明顯的發現以關鍵詞彙出現次數為門檻值來建立預測模型時，不但有較高的預測正確率，也能保有較低的投資錯誤機率，因此該模型應用於實際商業運作時，將更具實用性之價值。

實驗三：關鍵詞彙詞性組合類型

根據 Wu et al.（2002）、楊茂柱（2004）與許正欣（2004）的實驗指出，文章中的主要詞性組成多為名詞片語或是動詞片語，因此，此實驗組主要是透過先前研究所提出的合併規則，來找出較具語意之關鍵名詞，並透過實驗比較在不同詞性合併策略下，對預測效果之影響。透過不同的可能組合規則找出可能的關鍵詞彙，如表 7 所示。另外配合實驗二的實驗結果，我們設定關鍵詞彙的 TF 門檻值為 TF >= 2，以下分別就不同的組合類型探討模型的漲跌預測正確率。

表 7　關鍵詞彙組合規則

類型	組合規則	詞彙數	*TF* >= 2
KWa	非謂形容詞（A）+ 普通名詞（Na） 例：高階（A）+ 製程（Na）	75	48
KWb	普通名詞（Na）+ 普通名詞（Na） 例：晶圓（Na）+ 出貨量（Na）	773	244
KWc	普通名詞（Na）+ 動作及物動詞（VC） 例：技術（Na）+ 移轉（VC）	283	137
KWd	狀態及物動詞（VJ）+ 狀態不及物動詞（VH） 例：買超（VJ）+ 明顯（VH）	28	19
KWe	普通名詞（Na）+ 地方詞（Nc） 例：晶圓（Na）+ 代工廠（Nc）	73	51

KW_a 實驗組

由表 8 可以看出整體預測正確率並不高，只有 61.90%，而在上漲類別的預測中，可以發現本組合規則所萃取的關鍵詞彙並不能有效的辨識此類別的新聞，而下跌類別的正確率也不高，此實驗組所取出的複合名詞片語中所包含的「非謂形容詞（A）」，其意義主要是作名詞的修飾語，不具謂語作用，是純粹的形容詞，用來形容其後所接之普通名詞，例如：筆記型（A）電腦（Na）、共同（A）基金（Na）等，較無法表現出新聞中所隱含之漲跌資訊。另外就模型之實用性與經濟性而言，KW_a 實驗組之投資損失機率為 14.29%，而投資獲利機率亦為 14.29%，整體而言此實驗組的獲利或避險能力都表現不佳，容易造成投資上的損失。推測可能的原因為該類字詞較為中性，不能反映出漲跌情形，在模型中多被分類至持平的類別，使整體預測正確率差強人意。

表 **8** KW_a 實驗組之混亂矩陣表

預測 / 實際	上漲	持平	下跌	總計	目標正確率	整體正確率
上漲	0	4	1	5	0.00%	61.90%
持平	0	10	0	10	100.00%	
下跌	1	2	3	6	50.00%	
總計	1	16	4	21		

KW_b 實驗組

KW_b 實驗組所萃取出的關鍵詞彙為普通名詞與普通名詞的組合，以此規則所取出的關鍵詞彙數量在各實驗組中為最多者，可有較多的特徵作為新聞文件之描述，但相對的也會包含較多的雜訊。由表 9 可以發現 KW_b 實驗組在測試文件集的預測正確率為各實驗組之冠，整體預測正確率高達 80.95%，顯示以此種規則作為選詞的策略對於正確率有正面的影響與相關性。此類的關鍵詞彙包括現金（Na）股利（Na）、產能（Na）利用率（Na）、晶圓（Na）出貨量（Na）與記憶體（Na）晶片（Na）等，通常投資人對於該公司製程是否領先、或是本季的預期產能利用率與出貨量、或公司發放股利的政策、甚至職業災害等新聞，都會對其投資決策造成相當的影響，進而調整其投資比例，或決定買進與賣出持股。

KW_b 實驗組之投資損失機率為 9.52%，而投資獲利機率為 33.33%，顯示以此模型作為投資依據時，將有較佳的獲利機率與更低之投資風險。

表 **9** **KW_b** 實驗組混亂矩陣表

預測 / 實際	上漲	持平	下跌	總計	目標正確率	整體正確率
上漲	3	2	0	5	60.00%	80.95%
持平	0	10	0	10	100.00%	
下跌	1	1	4	6	66.67%	
總計	4	13	4	21		

KW_c 實驗組

從表 10 的實驗結果中可以觀察到，雖然透過普通名詞（Na）與動作及物動詞（VC）組合所取出的關鍵詞彙數量不少，但分析後發現該類詞彙多為殘缺不完整的詞組，或是單純敘述用的名詞片語，不能直接透過該字彙來看出本身所表示之意涵，例如：客戶（Na）量產（VC），新聞中投資者所關心的是廠商為客戶量產何種電子商品或是晶片等，但這類詞有些則為敘述之用，非新聞中所重要之關鍵詞彙；又如次長（Na）施（VC）、看法（Na）翻（VC）與月（Na）背（VC）等詞彙，被挑選出來的原因只是符合 KW_c 實驗組所設定的組合規則，但字彙本身並不能充分表達新聞之內容特徵。在 KW_c 實驗組的實驗結果中，上漲類別在該分類的預測正確率只達 40.00%，顯示該類關鍵詞彙對於上漲類別並沒有較強的正相關性，持平類別的預測效果仍是這三各類別中表現較佳，下跌類別在該類別的預測正確率為 33.33%，顯示模型不能正確的判別出可能造成股價負面下跌的新聞，整體而言，此實驗組的預測效果較差。而 KW_c 實驗組的投資損失機率高達 19.04%，遠高於 KW_a 與 KW_b 實驗組之預測模型，而投資獲利機率則為 19.04%，除顯示該模型的投資風險偏高，也顯示此類的詞性組合規則較不適用於本研究之個股樣本。

表 **10** KW_c 實驗組之混亂矩陣表

預測 / 實際	上漲	持平	下跌	總計	目標正確率	整體正確率
上漲	2	3	0	5	40.00%	57.14%
持平	2	8	0	10	80.00%	
下跌	0	4	2	6	33.33%	
總計	4	15	2	21		

KW_d 實驗組

此實驗組之關鍵詞彙萃取原則為狀態及物動詞(VJ)與狀態不及物動詞(VH)之詞性組合，該類詞彙多為一狀態或個股買賣趨勢，例如：買超（VJ）明顯（VH）、買超（VJ）不斷（VH）、展望（VJ）欠佳（VH）與呈現（VJ）走揚（VH）等，由上述的關鍵詞彙可以觀察到多由一些投資報告或財務報告中所萃取出，原先預期將有很高的預測正確率，但卻與實驗結果相異，由表 11 可以觀察到整體預測正確率達 52.38%，相當的不理想；而上漲類別的預測中，全部的測試新聞都被錯誤預測到持平類別中，而下跌類別的預測效果也是相當差，只有一篇新聞被正確辨識預測正確，其餘也是錯誤預測到持平類別中，雖然持平類別全部預測正確，但卻沒有獲利的空間，分析實驗結果後可歸納成下列兩點原因：

1. 關鍵詞彙數量過少

   由表 7 中可以發現，KW_d 實驗組所萃取的關鍵詞彙數量相對於其他實驗組為相對少數，且根據所設定的門檻值篩選之後，更是只剩下 19 個詞彙來做新聞文件之索引，過少的向量空間維度，不足以完整的描述新聞所隱含的資訊，也造成倒傳遞網路在訓練時無法收斂到所設定的目標（MES = 0.01），節點間的權重值較沒有明顯的差異化。

2. 股價已反應

   由於這些詞彙大多已是過時的資訊報導，投資人早已將所獲得的資訊反應至決策上，而新聞中也沒有包含相關的先進技術或是產能的相關字詞，所以無法正確的預測測試文件集中的未來個股股價之漲跌趨勢。

表 **11** KW_d 實驗組之混亂矩陣表

預測 / 實際	上漲	持平	下跌	總計	目標正確率	整體正確率
上漲	0	5	0	5	0.00%	52.38%
持平	0	10	0	10	100.00%	
下跌	0	5	1	6	16.67%	
總計	0	20	1	21		

KW_e 實驗組

此實驗組之詞性組合規則為普通名詞（Na）與地方詞（Nc）所形成的名詞組合，由表 12 觀察，47.62%之整體預測正確率為各實驗組之末，顯示以此規則作為關鍵詞彙的策略在本研究之特定產業個股樣本中，無法有效的建立預測模型；推論其可能原因如下：

1. 關鍵詞彙類型

由於根據此實驗組的詞性組合策略所挑選出的詞彙多為地方名詞，包含廠房、實驗室名稱、部會名稱或是廠商名稱等，例如：晶圓（Na）代工廠（Nc）、交通（Na）大學（Nc）、證券（Na）交易所（Nc）等，以這些字詞來對文件做索引時，無法表示該新聞所隱含之漲跌資訊。

2. 關鍵詞彙數量與語意

和 KW_d 實驗組的情形類似，以此策略所能符合的關鍵詞彙數量相對的少數，且地方名詞所包含的語意會依照新聞而有差異，也許是報導該廠房的公安事件，或是加碼投資該廠房的生產設備等，無法正確的辨識出其中所蘊含的資訊。

表 12　KW_e 實驗組之混亂矩陣表

預測 / 實際	上漲	持平	下跌	總計	目標正確率	整體正確率
上漲	2	3	0	5	40.00%	57.14%
持平	2	8	0	10	80.00%	
下跌	0	4	2	6	33.33%	
總計	4	15	2	21		

我們將各實驗組之投資損失機率、投資獲利機率與預測正確率比較列於圖 12，綜觀來說，採用 KW_b 類型的關鍵詞彙萃取策略的效果最好，該類詞彙所建立之股價漲跌預測模型有 80.95%的預測正確率，且投資損失機率為各實驗組中最好為 9.52%，另在投資獲利機率上更有高達 33.33%的表現，各項數據均優於其他的實驗組；而 KW_d 與 KW_e 實驗組則在各項評量數據上明顯較差，該兩組的預測正確率均為五成左右，且高達 23.81%的投資損失機率為各實驗組最高，而 4.76%的投資獲利機率也為各實驗組之末，研究所採用的樣本為電子業上市公司中的台灣積體電路公司，顯示該兩類的關鍵詞彙組合規則較不適用於在此樣本中，而 KW_a 與 KW_c 實驗組在各項數據的表現較為平均，但美中不足的則是這兩類型的預測正確率約只有六成左右，且投資損失機率相對來說還是偏高，整體而言，仍是以 KW_b 實驗組的表現最佳，因而在本研究中之關鍵詞彙組合類型較適合採用此類型的規則，來做關鍵詞彙的萃取。

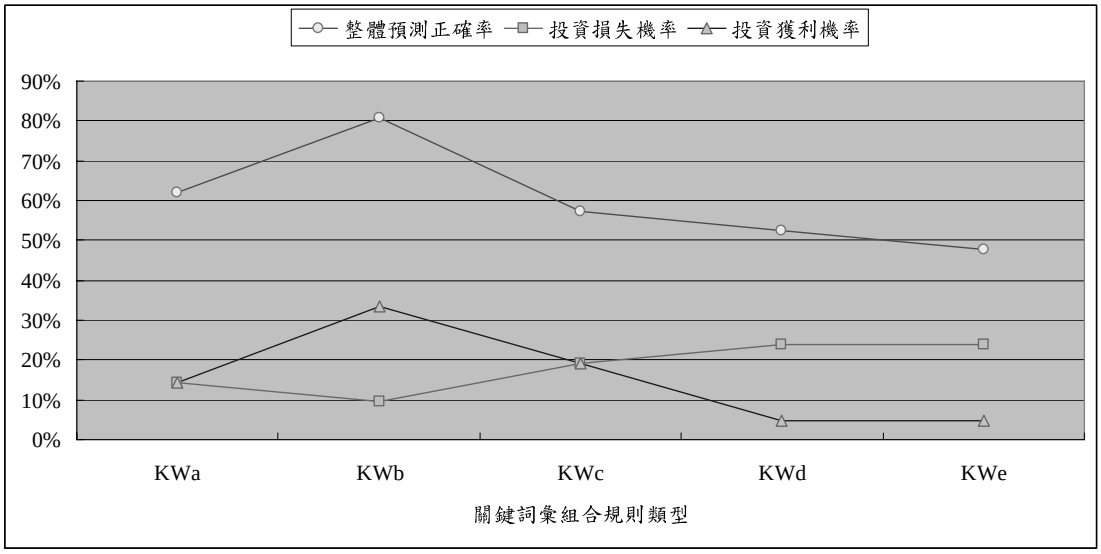

圖 12　關鍵詞彙組合規則類型於各項指標之實驗數據

實驗四：模擬真實交易

在此實驗組中，採用的參數設定為各實驗組中預測正確率為最佳的設定，因此本實驗組的股價漲跌反應時間設為一小時，在關鍵資訊擷取門檻值類型參數部分，採用關鍵詞彙的出現字數（*TF*）策略，並設定其門檻值為 *TF* >= 2，而在關鍵詞彙詞性組合類型參數設定上，採用的啟發式規則為普通名詞（Na）+ 普通名詞（Na）之詞性組合，訓練樣本為 2004/11/01～2004/12/31 日內的台積電相關新聞，測試資料區間為 2005/01/01～2005/03/31 之間的新聞，取樣限制如研究設計所述，由於實驗是採日內沖銷交易，必須採用信用交易之方式，而由於持平類別不能從中獲取利潤，因此模型只在預測股價趨勢為上漲或下跌時，會對應的發出交易訊號，在預測未來股價趨勢為上漲時，交易策略為先融資買進後融資賣出，而預測未來股價趨勢為下跌時，交易策略為先融券賣出後融券買進；每日只作一次進出場交易，並考慮真實交易環境下之所有利率與費用，透過多次交易後之總金額，計算由模型所得之報酬。

表 13 為此實驗組的每次交易時間、股價與資券相抵後之獲利一覽表，系統在新聞發佈後，會先對其內容作判定是否符合取樣的標準，例如交易時間、只能出現一個股代碼等限制，確認後即透過斷詞工具對文章作詞性標記的動作，透過先前所設定的關鍵詞彙詞性組合規則作萃取的動作，以空間向量索引的方式作為預測模型的輸入，透過實際的模擬當沖交易，來衡量本架構於實務面的貢獻與效能表現。綜合三個月內的 27 次模擬交易中，共有 5 次因為系統預測錯誤，造成套利不成反而蒙受龐大的成本損失，雖本研究的系統將交易手續費列入計算，但每次交易平均仍然可以獲得$987.62 的收益，顯示該模型確實在台股的電子股中有顯著的預測效果，證實本研究模型除架構之適用性之外，也能實際運用於真實之交易環境中。

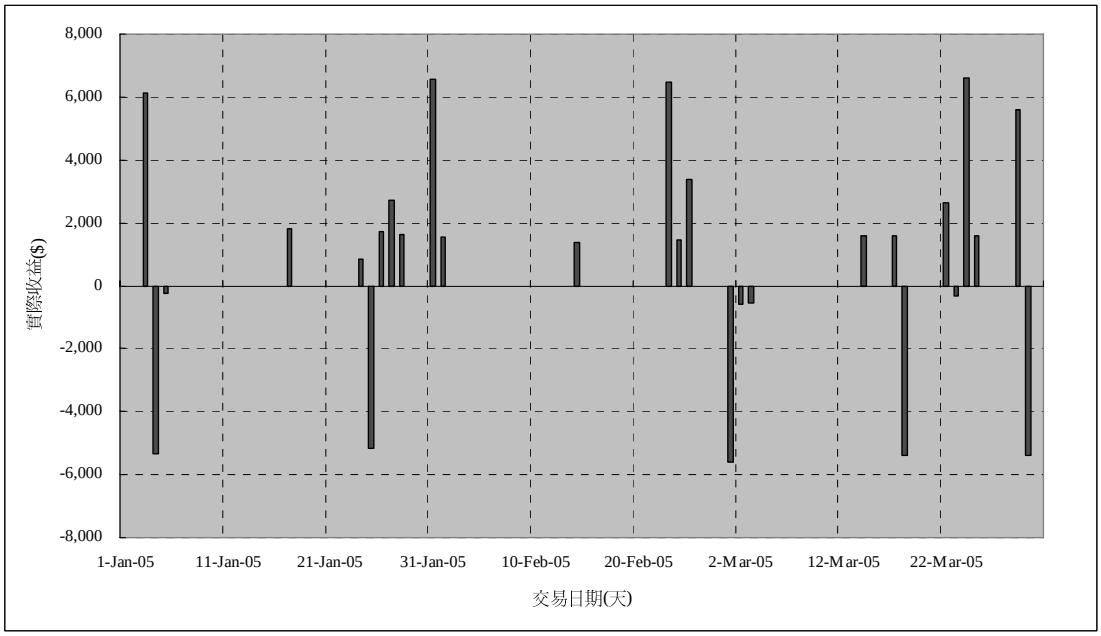

圖 13　預測模型於當日沖消機制下之效能表現

表 13　預測模型效能一覽表

獲利總額	$ 26,666
交易次數	27
預測正確率	81.48%
平均每次交易利潤	$ 987.62
單次交易最高利潤	$ 6,616
單次交易最高損失	$ -5,607
報酬率	5.33%（26,666/500,000）

與相關研究之比較

由文獻探討中可以發現先前研究中，並非都著重於日內股價漲跌預測之研究，不少研究多為大盤之中長期漲跌預測，標的多為歐美的股價指數，且財經新聞為英文新聞，不論是自然語言處理或是文字探勘過程的輔助工具都已發展多時，相對於中文新聞而言，萃取關鍵資訊的難度也不一，獲利評估方式也不盡相同。例如 Wuthrich et al.（1998）利用新聞來預測歐美主要指數的收盤價，但是卻沒有指出何時進出場做交易，較難評估其模型實用性，但仍有決策參考之價值，表 14 為與 Mittermayer（2004）的研究比較表，由於該研究也是運用文字探勘於美國股市的日內漲跌趨勢預測，因此作為與本研究之比較研究。

表 **14** 本研究與其他研究異同點比較表

	Mittermayer（2004）	本研究
分析方法	Support Vector Machine	Back-Propagation Network
新聞語系	英文	中文
關鍵詞取樣	word-based	phrase-based
關鍵資訊萃取方式	計算出各類別關鍵字的 TFIDF 值，取前 1000 個字作為代表	啟發式詞性組合規則與配合門檻值設定
關鍵資訊萃取難度	低	高
文件索引	二元方式表示的向量空間模型	二元方式表示的向量空間模型
平均交易獲利幅度	0.11%	0.20%
研究發現	不同的漲跌幅度設定，對平均獲利有顯著的影響	股價漲跌反應時間與詞性組合規則對正確率有顯著影響
研究貢獻	1. 建立一自動化的文件分類系統，並透過分類結果來發佈交易訊號。 2. 透過模型所產生的交易策略，比隨機交易的方式有更大的獲利。	1. 透過三項實驗評估各參數於台股環境的最佳設定。 2. 使用啟發式的詞性組合規則來萃取關鍵詞彙。 3. 提出一適用於中文新聞與台灣股市之預測模型。

結論與未來研究建議

投資股票為個人理財的重要工具之一，然而因股票價格隨市場機制波動卻也造成投資上的風險，因此如何選擇適當的股票買賣時機以確保獲利，一直是股票投資者所關心的議題。在傳統的分析技術上，雖然已有基本面分析及技術面分析等分析方式提供投資者在買賣股票上的建議，但是這些分析方式卻都忽略了消息面對短期股價的衝擊，使得短期投資者很難掌握股票的買賣時機。雖然現在已有許多的媒體提供相關新聞，但因消息來源太多，亦使得投資者沒有充分的時間來過濾相關資訊並迅速的取得投資決策，因此提供投資者一個簡單易用的工具來輔助投資決策，就成為近年來研究的重點。

先前的許多研究已針對新聞文件對股價影響的預測提供相當多的探討，而這些研究多以歐美股市作為探討的對象，然因台灣股市與歐美股市結構及語系並不相同，因此文件處理方式及預測模型架構亦不相同，針對中文語系的台灣股市受新聞影響的探討仍有其必要性。本研究提出一個台股日內股價漲跌的預測模型，藉由詞性標記方式將中文新聞文件斷詞後，透過可能的關鍵詞彙組合來萃取出每篇新聞之關鍵詞彙，再結合股價量化資料藉以建立預測模型。

藉由本研究模擬的市場交易機制，並以台股中的台積電股票為投資標的，實驗結果顯示，本研究提出之預測模型，在預測上漲或下跌之準確率上可達到81.48%的預測準確率，且季平均報酬率亦可達到 5.33%，若換算成年平均報酬率可達 21%，遠超過一般銀行的定存利率。對此，我們認為本研究提出之方法，對於股票投資人在短期買賣的操作上確實有其參考的價值。

本研究藉由模擬市場交易證實了所提架構確實能在股票市場的當沖機制中獲取利潤，然而本研究仍有研究上的限制可在後續的研究上繼續努力，如新聞樣本數量問題。

由於本研究是當日資券相抵交易，需要有較詳細的「時間」資訊作為和股價分時記錄的配對依據，然而目前所見的電子新聞媒體如中時電子報（http://news.chinatimes.com/）、聯合知識庫（http://udndata.com）等雖有龐大的新聞資料，但分析內容之後，發現都只有記載發佈日期，無法與每分鐘記錄一次的股價分時資料庫互相配對，取樣上也有諸多限制，雖然以此樣本

進行研究仍能獲得不錯的預測正確率，但測試樣本數量稍嫌不足。未來仍可持續搜尋相關網站以擴大新聞來源，並期能對預測的準確率有所提升。

參考文獻

林厚誼、蔣岳霖、周世俊(2002)〈The design and implementation of act e-service agent based on FAQ corpus〉，第七屆人工智慧與應用研討會（TAAI）。

許正欣（2004）《語意網上自動化建構本體論之研究》，輔仁大學。

楊茂柱（2004）《基於統計式語意相依關係之對話語句理解系統》，國立成功大學。

葉怡成（2000）《應用類神經網路（第二版）》，儒林圖書有限公司。

Ahmad, K., Oliveira, P. C. F. D., Manomaisupat, P., Casey, M. & Taskaya, T. (2002). Description of events: An analysis of keywords and indexical names. *Proceedings of the third international conference on language resources and evaluation, LREC 2002: Workshop on event modelling for multilingual document linking*, 29-35.

Fung, G. P. C., Yu, J. X. & Lam, W. (2002). News sensitive stock trend prediction. *Proceedings of the 6th Pacific-Asia conference on advances in knowledge discovery and data mining table of contents*, 481-493.

Fung, G. P. C., Yu, J. X. & Lam, W. (2003). Stock prediction: Integrating text mining approach using real-time news. *Proceedings of 2003 IEEE international conference on computational intelligence for financial engineering*, 395-402.

Gidófalvi, G. (2001). Using news articles to predict stock price movements. http://www.cs.ucsd.edu/users/gyozo/studies/cse254_AI/stock_price_prediction.pdf, 2004-06-15.

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D. & Allan, J. (2000). Mining of concurrent text and time series. *In: Proceedings of the 6th international conference on knowledge discovery and data mining*, 37-44.

Mittermayer, M. A. (2004). Forecasting intraday stock price trends with text mining techniques. *Proceedings of the 37th Hawaii international conference on system sciences*, 64-73.

Salton, G. & McGill, M. J. (1983). *Introduction to Modern Information Retrieval*, (2nd, ed.). NY: McGraw-Hill, Inc.

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

Sullivan, D. (2001). *Document warehousing and text mining*. Canada: Wiley Computer Publishing.

Wu, S. H., Day, M. Y., Tsai, T. H. & Hsu, W. L. (2002). FAQ-centered Organizational Memory, in Matta, N. and Dieng-Kuntz, R. (ed.), *Knowledge Management and Organizational Memories*, Massachusetts: Kluwer Academic Publishers.

Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K. & Zhang, J., et al. (1998). Daily stock market forecast from textual web data. *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*. CA: Los Alamitos, 2720-2725.