以文件分類技術預測股價趨勢

Predicting Trends of Stock Prices with Text Classification Techniques

陳俊達　Jiun-Da Chen
國立政治大學資訊科學系
Department of Computer Science
National Chengchi University
g9414@cs.nccu.edu.tw

王台平　Tai-Ping Wang
真理大學資訊管理學系
Department of Information Management
Aletheia University
tpwang@email.au.edu.tw

劉昭麟　Chao-Lin Liu
國立政治大學資訊科學系
Department of Computer Science
National Chengchi University
chaolin@nccu.edu.tw

摘要

股價的漲跌變化是由於證券市場中眾多不同投資人及其投資決策後所產生的結果。然而，影響股價變動的因素眾多且複雜，新聞也屬於其中一種，新聞事件不但是投資人用來得知該股票上市公司的相關營運資訊的主要媒介，同時也是影響投資人決定或變更其股票投資策略的主要因素之一。本研究提出以新聞文件做爲股價漲跌預測系統的基礎架構，透過文字探勘技術及分類技術來建置出能預測當日個股收盤股價漲跌趨勢之系統。

本研究共提出三種分類模型，分別是簡易貝氏模型、*k* 最近鄰居模型以及混合模型，並設計了三組實驗，分別是分類器效能的比較、新聞樣本資料深度的比較、以及新聞樣本資料廣度的比較來檢驗系統的預測效能。實驗結果顯示，本研究所提出的分類模型可以有效改善相關研究中整體正確率高但各個類別的預測效能卻差異甚大的情況。而對於影響投資人獲利與否的關鍵類別"漲"及類別"跌"的平均預測效能上，本研究所提出的這三種分類模型亦同時具有良好的成效，可以做爲投資人進行投資決策時的有效參考依據。

Abstract

Stocks' closing price levels can provide hints about investors' aggregate demands and aggregate supplies in the stock trading markets. If the level of a stock's closing price is higher than its previous closing price, it indicates that the aggregate demand is stronger than the aggregate supply in this trading day. Otherwise, the aggregate demand is weaker than the aggregate supply. It would be profitable if we can predict the individual stock's closing price level. For example, in case that one stock's current price is lower than its previous closing price. We can do the proper strategies(buy or sell) to gain profit if we can predict the stock's closing price level correctly in advance.

In this paper, we propose and evaluate three models for predicting individual stock's closing price in the Taiwan stock market. These models include a naïve Bayes model, a *k*-nearest neighbors model, and a hybrid model. Experimental results show the proposed methods perform better than the NewsCATS system for the "UP" and "DOWN" categories.

關鍵詞：股價預測，簡易貝氏模型，*k* 最近鄰居模型，混合模型。

Keywords: Stock Price Prediction, naïve Bayesian models, *k*NN models, hybrid models.

一、緒論

股價漲跌趨勢的預測是個令人感興趣的研究議題，然而影響股價變動的因素眾多且複雜。許多的相關研究使用技術分析或基本分析法[11]來做爲股價趨勢預測的特徵項目選取方式[6][8][12]。基本分析法著重於長期面的經濟因素變化，而對短期的證券市場的變動較不在乎。技術分析則著重於證券市場本身的變化，主要是透過圖表或技術指標的歷史資料及研究分析，從中找出規則並藉此來預測未來股價可能的趨勢變化；而不考慮其他可能也會對股價產生某種程度影響的外部因素，如經濟、政治、國際情勢…等其他各種方面的潛在影響。

對投資大眾而言，新聞是日常生活中非常容易接觸到的一種資訊來源；新聞是屬於會影響股價變化的非結構化資料，也是投資人可以用來得知該股票上市公司的相關資訊的主要媒介之一。新聞事件的本身也是影響投資人決定或變更其股票投資策略的其中一種考量因素；使其投資策略由賣方變爲買方，或由買方變爲賣方，導致交易市場中買賣雙方的力量發生變化，更進而影響股票價格的變動。因此，新聞文件除了是投資人在決定其投資策略的重要參考依據外，同時也隱藏著具有影響股價變化的可能性[13][14][17][18][21]。

對於新聞文件這種屬於非結構化的資料[15]，我們需要進行相關技術的處理才能將之轉化成半結構化或結構化資料，也才能從中進一步擷取出有用的資訊。然而，相關研究對於結合新聞與股價預測的研究議題上卻相對地著墨較少，且其在預測的成效上亦有其限制[14][17][21]。因此，本研究提出以結合文字探勘技術及分類技術來針對非結構化的新聞事件進行分析，建置出一個能預測個股當日收盤股價漲跌趨勢的系統，可同時改善相關研究中整體正確率高但各個類別的預測效能卻彼此差異甚大的情況提出改善之道，並可做爲輔助投資人進行投資決策時的有效參考依據。

本研究提出以預測個股股票當日收盤股價的漲跌趨勢變化來作爲本研究及系統建置的重心；透過整合股價資料與財經新聞事件，並結合文字探勘技術及分類技術來建置出預測台灣股市之個股當日收盤股價漲跌趨勢預測之系統模型，以提供投資人在股票交易時間內進行投資決策的參考依據。

我們提出三種不同分類模型來建立以新聞事件爲基礎的股價預測系統，分別是簡易貝氏模型、k 最近鄰居模型以及混合模型。在系統模型中，新聞事件的資料分析及處理是透過文字探勘技術來進行；而分類器則是用來整合個股股價資料與個股財經新聞事件，在一筆新的新聞事件發佈後，分類器可以自動將之分類並進行股價漲跌趨勢的預測。我們將新聞事件資料集分割成訓練資料及測試資料兩大類，利用訓練資料來訓練分類器，並用測試資料來驗證該分類器的成效好壞。透過整合股價資料與財經新聞事件，並結合文字探勘技術及分類技術來建置一個對證券市場中個別股票的當日收盤股價進行漲跌趨勢預測之系統模型。

本研究的實驗結果不但顯示新聞事件的本身是影響股價漲跌變化的主要因素之一，同時也顯示我們所提出的簡易貝氏模型、k 最近鄰居模型以及混合模型可以有效改善相關研究中整體正確率高，但各個類別的預測效能卻差異甚大的情況。而對於影響投資人獲利與否的關鍵類別"漲"及類別"跌"的平均預測效能上，本研究所提出的這三種分類模型亦同時具有良好的成效，可以做爲投資人進行投資決策時的有效參考依據。

在第二節中，我們將簡要地回顧相關文獻，說明股價預測的相關技術，及以新聞爲資料

來源的股價預測相關研究。第三節中則是說明我們的研究方法及系統架構。第四節中則是本研究的實驗及結論，我們將所收集的樣本資料集合進行相關實驗並分析實驗結果，最後提出本研究的結論。

二、文獻探討

本章節說明以新聞爲主要資料來源的股價漲跌趨勢預測相關研究文獻。

（一）、新聞對股價指數的預測

Wuthrich 等人[21]針對全球五個主要證券交易市場的股價指數(Dow Jones Industrial Average, Financial Times 100 Index, Nikkei225, Hang Seng Index, Straits Times Index)來進行當日股價指數漲跌趨勢的預測；透過收集並分析當日交易日開盤前的相關財經新聞網站所發佈的文章及新聞內容，並利用專家所建立的關鍵詞組資料庫及文字探勘技術，將該關鍵詞組利用權重方式設定其對股價的上漲或下跌的潛在影響力大小以進行證券市場股價指數的漲跌趨勢預測。該研究以正確率(Accuracy)做爲評估其系統效能的指標，實驗結果顯示 Wuthrich 等人的系統對全球五個主要證券交易市場股價指數的平均預測正確率爲 43.6%；實驗結果並顯示在假設市場中的交易成本爲零時，五個主要證券交易市場股價指數的平均報酬率 5.9%，而 Wuthrich 等人所建置的系統之平均報酬率達 20.8%。

（二）、新聞對個別股票股價的預測

Gidófalvi[14]則是探討新聞事件發佈後對相關股票即時股價變化的影響，其研究的基本假設是認爲在交易時間內所發佈的新聞事件會在其發佈後的某一段時間內對相關個股的股價具有影響力(window of influence)，而導致股價的變動，並提出新聞事件對股價變化的影響力時間間隔爲該新聞發佈的前後 20 分鐘之內。Gidófalvi 結合即時股價資料及即時新聞資料，並透過簡易貝氏文件分類器(naïve Bayesian text classifier)來對交易時間內所發佈的新的一筆新聞來進行分類，並預測該新聞可能對股價變化的影響。Gidófalvi 將新聞事件發佈後對股價的影響分爲三個類別: 「上漲(Up)」、「不變(Unchanged)」、及「下跌(Down)」，並透過這三個類別標籤來建立該筆新聞事件與股價變動程度之間的關係。

在 Mittermayer[17]研究中，其所提出的 NewsCATS(News Categorization and Trading System)是一個可以對新聞進行自動化的分析與分類的系統，該系統並可以主動提出投資策略的建議。實驗結果顯示 NewsCATS 投資策略建議具有比以隨機方式決定買賣投資策略更好的成效，隨機方式的每筆平均投資報酬率爲 0%，而 NewsCATS 的每筆平均投資報酬率則爲 0.11%。Mittermayer 認爲在 NewsCATS 中以新聞分類的方式可以提供比新聞本身更多的資訊來進行股價趨勢的預測。

（三）、中文文件前處理

中文斷詞方式主要可分成下列兩種方法[9]：詞庫比對法(Dictionary-Based Approach)以及統計分析法(Statistical Approach)。詞庫比對法是指透過事先建立的詞庫，對輸入文件中的詞彙進行比對，再擷取出文件中出現的詞彙，完成斷詞程序。統計分析法則是透過大量文件分析，經由分析結果取得統計參數後，擷取出統計參數滿足某些條件的詞，這些統計參數可以是詞彙發生的頻率，但此方法的缺點在於當關鍵詞出現的頻率極少時，可

能無法被擷取出來。本研究對中文斷詞的處理方法是選擇採用詞庫比對法；透過由中央研究院中文詞知識庫小組中文詞庫來進行新聞文件字句的中文斷詞處理程序。

三、研究方法

本章節說明我們的研究目標、步驟、系統架構，以及本研究所提出的簡易貝式模型、*k* 最近鄰居模型以及混合模型這三種分類器。

（一）、研究目標及系統架構

本研究的目標是結合文字探勘技術及分類技術來建立一個能預測股票當日收盤股價漲跌趨勢之系統模型，以提供投資人在股票交易時間內進行投資決策的參考依據。舉例來說，若系統預測當日該股票的收盤股價會是上漲時，則在當日交易時間內的股價波動若低於前一交易日的股價時，則可建議投資人買進，倘若當日該股票的收盤股價也確實是上漲時，則投資人將會因而獲利；反之亦然。

本研究所提出的系統模型架構如圖 1 所示。我們提出三種不同分類模型來建立以新聞事件爲基礎的股價預測系統。在系統模型中，新聞事件的資料分析及處理是透過文字探勘技術來進行；而分類器則是用來整合個股股價資料與個股財經新聞事件，在一筆新的新聞事件發佈後，分類器可以自動將之分類並進行股價漲跌趨勢的預測。我們將新聞事件資料集分割成訓練資料及測試資料兩大類，利用訓練資料來訓練分類器，並用測試資料來驗證該分類器的成效好壞。透過整合股價資料與財經新聞事件，並結合文字探勘技術及分類技術來建置一個對證券市場中個別股票的當日收盤股價進行漲跌趨勢預測之系統模型。

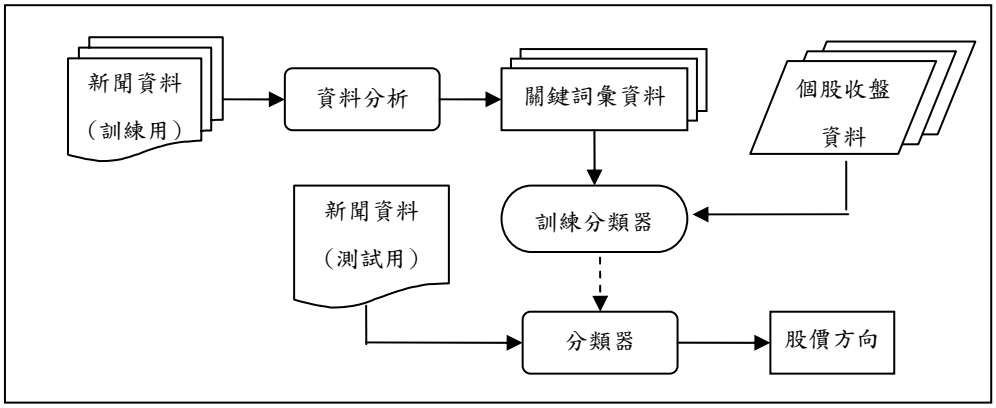

圖 1. 系統架構

（二）、分類器

分類(classification)[15][19]是指透過分類器將未知類別的資料依據其屬性值的不同來完成對該資料分派類別標籤的過程。分類器會先透過事先提供的訓練資料來學習分類規則，分類器訓練完畢後便可針對新的一筆未知類別的測試資料進行自動分類並建立該筆資料的類別標籤。在分類器的選擇上，在本研究中我們共設計三種不同的分類模型做爲分類器的核心，分別是簡易貝氏模型(naïve Bayes models)[15][20]、*k* 最近鄰居模型(*k*NN models)[15][20]以及混合模型(hybrid models)，以下分別敘述之。

1、簡易貝氏模型

簡易貝氏模型是以貝氏定理(Bayes' theorem)爲基礎，透過交換事前(prior probability)、事後機率(posterior probability)的方式來將未知類別的測試資料分派到類別機率最大的類別。簡易貝氏模型會先根據訓練資料的樣本來建立各類別機率表，對於之後所給予的測試資料則會依其的屬性值計算其歸屬於各個類別的機率值，並將具有最高機率值的類別作爲該測試資料的類別標籤。說明如下。

假設目前存在某一個特徵值x，且在樣本空間中可能出現的類別總共有k個$\{C_1, C_2, \ldots, C_k\}$，每個類別彼此間均互斥。P(C1)、P(C2)、…、P(Ck)分別爲其事前機率，則P(x)表示如下(公式 1)。

$$P(x) = \sum_{i=1}^{k} P(x \cap C_i) \cdots\cdots (公式 1)$$

條件機率(conditional probability)是指在已知出現某一特徵值x的條件下，某個類別C_i發生的機率記爲P(C_i|x)，其計算公式如下(公式 2)。

$$P(C_i \mid x) = \frac{P(C_i \cap x)}{P(x)} \cdots\cdots (公式 2)$$

事後機率是指當該特徵值x出現時，屬於類別C_i的機率，表示爲P(C_i|x)，其公式如下(公式 3)。

$$P(C_i \mid x) = \frac{P(C_i \cap x)}{P(x)} = \frac{P(C_i) \bullet P(x \mid C_i)}{P(x)} \cdots\cdots (公式 3)$$

假如目前是一組彼此互相條件獨立的特徵值$(x_1, x_2, \ldots, x_d)$時，則當給定某個類別C_i時，其條件機率可以表示如下(公式 4)。

$$P(x_1, x_2, \ldots, \mathrm{x}_d \mid \mathrm{C_i}) = \prod_{j=1}^{d} P(x_j \mid C_i) \cdots\cdots (公式 4)$$

依循上述的模式，我們可以得到 k 個類別中，包含 d 個特徵值的簡易貝式分類器模型，其公式如下(公式 5)。

$$P(C_i \mid x_1, x_2, \ldots, x_d) = \frac{P(C_i) \bullet \prod_{j=1}^{d} P(x_j \mid C_i)}{\sum_{i=1}^{k} \left(P(C_i) \bullet \prod_{j=1}^{d} P(x_j \mid C_i) \right)} \cdots\cdots (公式 5)$$

在特徵值選取的方法主要有資訊增益值(IG, Informatin Gain)、資訊增益比(Gain Ratio)、Gini-index、距離度量(Distance Measure)、J-measure、G 統計、$\chi 2$ 統計、最小描述長度(MLP)、正交法(Ortogonality Measure)、Relief…等，不同的度量方法有不同的分類效果，特別是對於高度分支的特徵值屬性(highly branching attributes)。在本實驗中我們嘗試以資訊增益值來做爲特徵值選取的方法，在未來我們認爲可以去嘗試不同的特徵值選取方法，以選取出更適當的特徵值來增進系統的分類效果。以下是信息增益值的簡要介紹。

資訊增益值主要是運用熵值(Entropy)的概念來做爲屬性選擇的評估依據，資訊增益值的計算方式是將未分類之前所獲得的資訊量減去分類後的資訊量，並以增益值的大小來做

爲特徵值選取的評估依據[19][20]，其計算公式如下(公式 6)。Ex 是原始樣本資料集合，H(Ex)是原始類別的熵值，H(Exla)則是考慮特徵值 a 後其不同屬性值下的熵值加總。

$$IG(Ex,a) = H(Ex) - H(Ex \mid a) \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(公式 6)}$$

因此，在簡易貝氏模型及 *k* 最近鄰居模型中的特徵值是以將新聞資料集合進行中文斷詞處理後所取得的關鍵詞，分別計算其資訊增益值後取出前 d 個關鍵詞來做爲其特徵值。而對於 d 值的設定上我們是透過對取樣的新聞資料集合進行初步實驗來決定，我們任意選取 3 個數量來比較其對系統效能的差異，分別是 25、50 及 100 個特徵值數量，並從中選出一個相對較好的來做爲 d 值的設定，以聯電新聞資料集之一爲例，實驗標的爲聯華電子股份有限公司，新聞資料來源爲台灣 Yahoo!奇摩股市新聞資料庫，資料取樣期間爲民國 95 年 2 月至民國 96 年 4 月，資料來源爲台灣 Yahoo!奇摩股市新聞資料庫，該新聞資料集經中文斷詞處理後共有 2890 個關鍵詞，透過上述方式的初步實驗結果顯示，以我們任意設定的這三個特徵值數量的整體效能而言，25 及 50 是差異不大，100 則相對較差一些，因此對於聯電新聞資料集之一的 d 值我們設定爲 50 個特徵值；對於本實驗其餘四組新聞取樣期間較短的新聞資料集合的 d 值設定則爲 25 個特徵值。

在本實驗中對於每一個特徵值的可能值只有兩種，出現(True)或未出現(False)，並假設這些特徵值彼此是條件獨立。對於某個特徵值在某一筆新聞中是否出現，我們可以透過檢查在該筆新聞文件中的該特徵值所代表的關鍵詞之詞頻是否大於 0，若該關鍵字至少出現一次則該特徵值的值爲出現，否則在該筆新聞文件中該特徵值的值則視爲未出現。在本研究中的類別數共有三個，分別是類別"漲"、類別"跌"及類別"持平"，透過訓練資料我們可以計算出每個類別出現的機率，表示爲P(C="漲")、P(C="跌")及P(C="持平")；並可以分別計算出在已知類別下第i個特徵值(x_i)出現及未出現的機率，表示爲P(x_i=出現|C="漲")、P(x_i=未出現|C="漲")、P(x_i=出現|C="跌")、P(x_i=未出現|C="跌")、P(x_i=出現|C="持平")、P(x_i =未出現|C="持平")。然而，在實際上可能會發生在已知類別下個某個特徵值都未出現而導致零機率的現象，進而在計算事後機率時(公式 5)會因爲該機率值爲零而導致無論其它機率值多大都還是會使機率相乘的結果爲零的絕對否定現象，因此我們採取將某個特徵值的所有可能值的出現次數都加上 $\frac{1}{\text{特徵值總數}}$ 的方式來避免零機率的現象[20]。

對於一筆未知類別但已知特徵值($x_1,x_2,\cdots,x_d$)的測試資料時，我們可以透過公式 5 來計算每個類別的事後機率，表示爲P(C="漲"| $x_1,x_2,\cdots,x_d$)、P(C="跌"| $x_1,x_2,\cdots,x_d$)及P(C="持平"| $x_1,x_2,\cdots,x_d$)，並將該筆未知類別的分類爲具有最大機率值的類別，而在實際計算上，由於分母都是相同的，因此我們可以僅計算分子並比較其大小即可。

2、*k* 最近鄰居模型

k 最近鄰居模型所根據的基礎是 *k* 最近鄰居分類法(*k*NN, *k*-Nearest Neighbor Algorithm)[20]。最近鄰居分類法是指相同一類的物件彼此應該會聚集在一起，即所謂的「物以類聚」。若以向量空間中的點來表示，則對於同一類別物件的這些點彼此間的距離應該會比較接近。所以對於一個未知類別的測試資料，我們只需要在訓練資料中找出和此筆資料最接近的點，就可以最近鄰居分類法來判定此筆未知類別的測試資料的類別應該和其最接近的點的類別是相同的。然而，在多數情況下若只有使用最近鄰居來決定類別可能並不恰當。因此，常見的做法是先求取最接近的 *k* 個資料點，再根據對應的 *k*

個類別資訊來進行投票，來決定最後的類別，這種方法稱爲 k 最近鄰居分類法，也就是以 k 個最靠近的鄰居來投票決定自己的類別，至於最好的 k 值，完全是取決於資料而定。

k 最近鄰居模型是根據 k 最近鄰居分類法來找出所有的訓練資料中和該筆測試資料距離最近的 k 個鄰居，並比較這 k 個鄰居的類別標籤何者類別爲最多數後，將以此類別做爲該筆測試資料的類別，即以多數決的方式將該筆測試資料歸類爲 k 個最近鄰居中所屬的類別中票數最高的類別。

本研究中對於距離的計算方式是採用歐幾里得距離(Euclidean distance)[20]，假設在 n 維的向量空間中有兩個點 $P=(p_1,p_2,...p_n)$ 、 $Q=(q_1,q_2,...,q_n)$ ，則歐幾里得距離的計算公式如下(公式 7)。

$$D_{Euclidean}=\sqrt{\sum_{i=1}^{n}(p_i-q_i)^2} \quad \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\text{(公式 7)}$$

在建置k最近鄰居模型過程中，我們將整個新聞資料集合進行中文斷詞及TFIDF處理後可以建立一個關鍵詞對應於新聞文件的矩陣，假設共有d個關鍵詞，則我們可將此矩陣視爲一個d維的向量空間，每一筆新聞文件代表此d維向量空間中的一個點。因此，對於對於一筆未知類別但已知特徵值($x_1,x_2,\cdots,x_d$)的測試資料時，我們會去計算和每一筆訓練資料的距離，透過公式 7 我們可以找出k個與該測試資料最近的鄰居，並以其中最多數的類別做爲該測試資料的類別。

3、混合模型

我們並同時提出一種結合簡易貝氏模型及 k 最近鄰居模型的混合模型，透過設定門檻值 ε 的方式使得混合模型可以判別並分派一筆新的測試資料到其所適合的分類模型中來進行分類。細節說明如下。

當一筆測試資料以混合模型來決定其類別時，我們會先分別計算該筆測試資料在簡易貝氏模型中類別機率最高及次高的機率值，分別以C_i及C_j代表之。接著，混合模型會去檢查該類別機率最高及次高的類別兩者機率值差距的比例大小Δp是否大於混合模型中所事先設定的門檻值 ε ，Δp的計算公式如下(公式 8)。

$$\Delta P=\frac{P(C_i)-P(C_j)}{P(C_i)}\text{，}C_i,C_j\in\{C_1,C_2,...,C_k\}\cdots\cdots\cdots\cdots\cdots\cdots\text{(公式 8)}$$

在Δp小於 ε 的情況下，代表在簡易貝氏模型中具有最高機率值的類別(C_i)和次高機率值的類別(C_j)對該筆測試資料而言是不相上下的，也就是說該測試資料的類別不是很明顯地應該被分類爲C_i，因爲該筆測試資料歸屬於C_j類別的機率也不小。因此，在Δp小於 ε 的情況下，混合模型會選擇以k最近鄰居的模式來將該筆測試資料進行分類；而在Δp大於 ε 的情況下，混合模型就會選擇以簡易貝氏的模式來將該筆測試資料進行分類，見公式 9。

$$\text{混合模型}=\begin{cases}\text{若 } \Delta P>\varepsilon\text{，則適用簡易貝氏模型}\\ \text{若 } \Delta P\le\varepsilon\text{，則適用}k\text{最近鄰居模型}\end{cases}\cdots\cdots\cdots\cdots\text{(公式 9)}$$

透過這種機制，我們可以避免在簡易貝氏模型中只能將新聞的類別標籤分派給具有最高機率的類別，即使是在最高與次高類別的機率值差距微乎其微時，此情況下所隱含的意義是該新聞並非可以被明顯歸類於具有最高機率的類別，而本研究所提出的混合模型就

可以在此種情況下提供另一個客觀的比較依據。

四、實驗及分析

本章節是介紹本研究的實驗資料來源、實驗設計、評估方法，以及實驗的結果與分析。

（一）、資料來源及實驗設計

本研究的實驗標的爲台灣證券市場的股票上市櫃公司，我們針對電子類股中的半導體類股及發光二極體類股中選出四家公司來進行本研究相關實驗，並收集與該公司相關的財經新聞料及該個股收盤股價資料，總共取得四組新聞資料集合(聯電新聞資料集、光寶新聞資料集、晶電新聞資料集、以及立碁新聞資料集)，總計 747 筆財經新聞資料。新聞資料來源爲台灣 Yahoo!奇摩股市新聞資料庫[1]，資料取樣期間爲民國 95 年 9 月至民國 96 年 4 月；個股收盤股價資料來源則爲台灣證券交易所的個股日收盤價資料庫[4][10]。

本研究的實驗標的則是針對相同的取樣期間但新聞樣本的股票取樣標的不同的資料樣本集，實驗目的則是用來檢驗並比較本研究所提出的簡易貝氏模型、*k* 最近鄰居模型以及混合模型是否具有提升系統整體預測效能的共通性，並比較不同樣本資料集合對於不同股票標的下的系統效能差異程度。在實驗中，我們採用三折式交叉驗證分析法(3-fold cross-validation)[20]來做實驗，也就是將所收集的資料樣本平均切成三等分，其中三分之二的資料樣本做爲訓練資料，三分之一的資料樣本作爲測試資料，並分別計算出該測試資料的精確率及召回率；之後將訓練、測試資料輪流對換，此步驟共執行 3 次，直到讓每一筆資料都當過一次測試資料，如此可得到整體偏差值較小且客觀的數據。

（二）、評估方法

本實驗的評估方法是採用精確率(Precision)、召回率(Recall)、以及 F-measure 來做爲評估系統成效的指標[19]。符號定義如下：

TP：文件實際爲該類別，而系統也正確地將文件分類爲該類別之個數
FP：文件實際非屬該類別，但系統將文件分類爲該類別之個數。
TN：文件實際非屬該類別，系統也正確地將文件分類成非該類別之個數。
FN：文件實際爲該類別，但系統將文件分類成非屬該類別之個數。

精確率是計算分類系統預測其爲某一類別時，且系統正確預測的百分比，其公式如下。

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{(公式 10)}$$

召回率是計算分類系統補捉到正確分類的百分比，其公式如下。

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad \text{(公式 11)}$$

F-measure 是依據精確率和召回率兩個指標加以綜合而成的評估指標，其計算公式見公式 12，其中 α 值是用來設定在 F-measure 中精確率、召回率重要程度高低的調整參數，在本研究中，我們將 F-measure 的 α 值設爲 1，也就是將精確率及召回率對於 F-measure 影響力的重要程度視爲是均等的。

$$\text{F - measure}_{\alpha} = \frac{(1+\alpha) \bullet \text{Precision} \bullet \text{Recall}}{\alpha \bullet \text{Precision} + \text{Recall}},\ \alpha \geq 0 \quad \text{.....(公式 12)}$$

此外，我們並以「所有類別平均」及「類別"漲"、"跌"平均」來做爲各屬性值離散化區間下的整體平均效能評估指標，以及「類別標準差」來做爲比較類別彼此間差異程度的評估指標[5]。

「所有類別平均」是指類別"漲"、類別"跌"、以及類別"持平"的這三個類別的整體平均精確率、整體平均召回率及整體平均 F-measure 值，此評估指標可以用來衡量系統整體的預測效能，指標值越高代表系統整體的平均預測效能越佳，對於投資人在進行投資決策時的參考價值也越高，其公式分別如公式 13、公式 14、公式 15 所示。

$$\text{所有類別平均}_{\text{精確率}} = \frac{\sum_{i=1}^{3} \text{Precision}_{\text{類別}i}}{3},\ \text{類別}i = \{\text{"漲","持平","跌"}\} \text{.....(公式 13)}$$

$$\text{所有類別平均}_{\text{召回率}} = \frac{\sum_{i=1}^{3} \text{Recall}_{\text{類別}i}}{3},\ \text{類別}i = \{\text{"漲","持平","跌"}\} \text{.....(公式 14)}$$

$$\text{所有類別平均}_{\text{F-measure}} = \frac{\sum_{i=1}^{3} \text{F - measure}_{\text{類別}i}}{3},\ \text{類別}i = \{\text{"漲","持平","跌"}\} \text{.....(公式 15)}$$

「類別"漲"、"跌"平均」則是指類別"漲"及類別"跌"這兩個類別的平均精確率、平均召回率及平均 F-measure 值，此評估指標可以用來衡量系統對於真正會影響投資獲利與否的類別"漲"及類別"跌"這兩個類別的平均預測效能，指標值越高代表系統對於類別"漲"及類別"跌"這兩個類別的平均預測效能也越佳，除了可提供投資人在進行買進或賣出時的投資決策輔助外，更是影響投資人獲利與否的關鍵指標，其公式分別如公式 16、公式 17、公式 18 所示。

$$\text{類別"漲"、"跌"平均}_{\text{精確率}} = \frac{\sum_{i=1}^{2} \text{Precision}_{\text{類別}i}}{2},\ \text{類別}i = \{\text{"漲","跌"}\} \text{.....(公式 16)}$$

$$\text{類別"漲"、"跌"平均}_{\text{召回率}} = \frac{\sum_{i=1}^{2} \text{Recall}_{\text{類別}i}}{2},\ \text{類別}i = \{\text{"漲","跌"}\} \text{.....(公式 17)}$$

$$\text{類別"漲"、"跌"平均}_{\text{F-measure}} = \frac{\sum_{i=1}^{2} \text{F - measure}_{\text{類別}i}}{2},\ \text{類別}i = \{\text{"漲","跌"}\} \text{...(公式 18)}$$

「類別標準差」是用來衡量各類別彼此間的差異程度大小的指標，類別標準差越小表示該系統對於各類別的預測效能越一致及穩定，越能提供投資人進行投資決策時的有效參考依據。反之，若類別標準差越大則代表該系統對於各類別的預測效能落差較大，較容易發生對於某一個類別的預測效能很高，但對另一個的預測效能卻可能非常低的情況發生，此情況會導致該系統的預測結果並不能提供投資人進行投資決策時的有效參考依據。對於分類器的整體預測效能而言，我們會希望其「類別標準差」越小越好，代表該

分類器對於各類別的預測效能較爲穩定及可靠。類別標準差的計算公式分別如公式 19、公式 20、公式 21 所示。

$$\text{類別標準差}_{\text{Precision}} = \sqrt{\frac{\sum_{i=1}^{3}(\text{Precision}_{\text{類別i}} - \overline{\text{Precision}_{\text{所有類別平均}}})}{3}}\text{，類別}i=\{\text{"漲","持平","跌"}\}$$

……(公式 19)

$$\text{類別標準差}_{\text{Recall}} = \sqrt{\frac{\sum_{i=1}^{3}(\text{Recall}_{\text{類別i}} - \overline{\text{Recall}_{\text{所有類別平均}}})}{3}}\text{，類別}i=\{\text{"漲","持平","跌"}\}$$

……(公式 20)

$$\text{類別標準差}_{\text{F-measure}} = \sqrt{\frac{\sum_{i=1}^{3}(\text{F-measure}_{\text{類別i}} - \overline{\text{F-measure}_{\text{所有類別平均}}})}{3}}\text{，類別}i=\{\text{"漲","持平","跌"}\}$$

……(公式 21)

（三）、模擬 NewsCATS 系統

本研究的重心是透過分結構化的新聞資訊來進行相關個股的當日收盤股價漲跌趨勢預測的研究。然而，在我們目前爲止所能找到的股價預測相關研究中，滿足同樣是透過分析新聞資訊並針對相關個股進行股價漲跌趨勢預測的研究限制者中，僅以 Mittermayer 的研究和本研究最爲接近。因此在本研究中的實驗比較標的爲 Mittermayer 所提出的 NewsCATS 系統。雖然 NewsCATS 的整體預測效果優異，且對於類別"持平"的平均預測精確率高達 98%，但對於類別"漲"、"跌"的平均預測精確率卻分別只有 5%及 6%，預測效果較類別"持平"差距非常顯著。這個現象顯示 NewsCATS 的系統限制是僅能提供投資人對於類別"持平"的預測來做爲其投資決策的參考依據，對於投資人更爲重視且影響其獲利與否的類別"漲"及類別"跌"這兩個類別上的預測效果，NewsCATS 系統卻不能提供投資人有效及可靠的預測參考依據。

基於 Mittermayer 的研究和本研究仍有許多不同之處，如：新聞資訊的語言不同、證券市場不同、新聞取樣期間不同…等諸多差異點。且受限於難以取得 Mittermayer 當時的樣本資料及其分類器的重建時的相關設定，因此，我們採取模擬 NewsCATS 的方式來做爲實驗的比較基礎。在之後進行相關的實驗中，本研究會以模擬的 NewCATS 系統來代替實際的 NewsCATS 系統，並和我們所提出的簡易貝氏模型、*k* 最近鄰居模型以及混合模型來進行彼此系統預測效能的比較。以下我們會以模擬的 NewsCATS 系統以代替實際的 NewsCATS 系統來做爲實驗比較的標的。

（四）、實驗結果及分析

本實驗的重心在於探討在相近的取樣期間下，對於不同新聞標的股票下的不同系統彼此間的整體效能差異程度爲何，並進一步探討本研究在不同的資料集合下是否仍具有提升系統整體預測效能的適用性。本實驗的樣本資料集共有四個，分別是聯電新聞資料集、光寶新聞資料集、晶電新聞資料集、以及立碁新聞資料集。

表 1 中分別爲聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集在建置簡易貝氏模型、*k* 最近鄰居模型與混合模型時的最適參數設定值。

表 1. 相近取樣期間不同樣本資料集的最適參數設定值

新聞樣本資料集合	最近鄰居數	門檻值設定	資料取樣期間	資料筆數
聯電新聞資料集	k= 1	ε = 20%	4 個月	197
光寶新聞資料集	k= 1	ε = 10%	5 個月	187
晶電新聞資料集	k= 1	ε = 90%	7 個月	291
立碁新聞資料集	k= 7	ε = 90%	5 個月	72

我們針對聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集這四組取樣期間相近且針對不同新聞標的股票的新聞樣本資料集合進行彼此間系統效能的比較與分析探討。在這四組新聞樣本資料集合中，我們以「所有類別平均」、「類別"漲"、"跌"平均」及「類別標準差」來評估在本研究所提出的簡易貝氏模型、k 最近鄰居模型、混合模型與模擬 NewCATS 系統的整體效能；「所有類別平均」可以顯示出系統整體的平均預測效能，「類別"漲"、"跌"平均」則可以顯示影響投資人獲利與否的類別"漲"及類別"跌"的平均預測效能，而「類別標準差」則可以顯示系統內的類別間彼此預測效能的差異程度。

在表 2、表 3 及表 4 中，分別顯示聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集在簡易貝氏模型下的系統預測精確率、召回率及 F-measure 值，其中的模擬 NewsCATS 系統欄位的數值是將每個新聞樣本資料集下在其系統下所模擬的 NewsCATS 系統分別加總平均，也就是在這四組新聞樣本資料集合下所模擬的平均 NewsCATS 系統效能。

實驗結果顯示，當以「所有類別平均」來做爲系統整體評估指標時，除了晶電電新聞資料集在精確率及 F-measure 較模擬 NewsCATS 系統分別低 2.66%及 0.61%外，其餘的新聞樣本資料集合的系統整體平均預測效能都是比模擬 NewsCATS 系統高。

而對於影響投資人獲利與否的類別"漲"及類別"跌"的評估指標「類別"漲"、"跌"平均」，聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集這四組新聞樣本資料集合的系統預測效能都是比模擬 NewsCATS 系統高；其中，精確率比模擬 NewsCATS 系統高 9.71%到 33.00%之間，召回率比模擬 NewsCATS 系統高 7.28%到 32.61%之間，F-measure 值比模擬 NewsCATS 系統高 17.45%到 27.36%之間。

當以「類別標準差」來評估在這四組新聞樣本資料集合下的簡易貝氏模型中的類別間彼此預測效能差異程度時，實驗數據顯示模擬 NewsCATS 的類別間的預測效能差異程度最大，顯示該系統的預測並不能有效投資人進行投資決策時的參考依據。

表 2. 相近取樣期間不同樣本資料集下之簡易貝氏模型精確率比較

精確率(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	49.24	73.81	53.03	24.07	80.41
類別"跌"	47.87	39.17	34.39	62.50	11.48
類別"漲"	45.70	44.62	24.44	42.93	27.95
所有類別平均	47.60	52.53	37.29	43.17	39.95

類別"漲"、"跌"平均	46.79	41.89	29.42	52.71	19.71
類別標準差	1.79	18.63	14.51	19.21	39.24

表 3. 相近取樣期間不同樣本資料集下之簡易貝氏模型召回率比較

召回率(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	49.54	18.67	33.90	77.78	75.72
類別"跌"	55.35	16.85	62.13	18.52	20.94
類別"漲"	36.51	91.72	25.94	39.39	22.43
所有類別平均	47.13	42.41	40.66	45.23	39.70
類別"漲"、"跌"平均	45.93	54.29	44.03	28.96	21.68
類別標準差	9.65	42.71	19.02	30.06	40.33

表 4. 相近取樣期間不同樣本資料集下之簡易貝氏模型 F-measure 值比較

F-measure(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	49.39	29.80	41.36	36.77	77.70
類別"跌"	51.34	23.56	44.27	28.57	14.61
類別"漲"	40.59	60.03	25.17	41.08	19.74
所有類別平均	47.37	46.93	38.90	44.17	39.51
類別"漲"、"跌"平均	46.35	47.29	35.27	37.38	19.93
類別標準差	5.73	19.51	10.29	6.36	38.71

在表 5、表 6 及表 7 中分別顯示聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集在 *k* 最近鄰居模型下的系統預測精確率、召回率及 F-measure 值。

實驗結果顯示，當以「所有類別平均」來做爲系統整體評估指標時，這四組新聞樣本資料集合的系統預測效能都是比模擬 NewsCATS 系統高；其中，精確率比模擬 NewsCATS 系統高 2.62%到 10.81%之間，召回率比模擬 NewsCATS 系統高 1.04%到 11.56%之間，F-measure 值則比模擬 NewsCATS 系統高 2.12%到 11.50%之間。而對於影響投資人獲利與否的類別"漲"及類別"跌"的評估指標「類別"漲"、"跌"平均」，除了立碁新聞資料集在召回率上較模擬 NewsCATS 系統低 2.24%外，其餘都是較模擬 NewsCATS 系統的預測效能高。當以「類別標準差」來評估在這四組新聞樣本資料集合下的 *k* 最近鄰居模型中的類別間彼此預測效能差異程度時，實驗數據也顯示模擬 NewsCATS 的類別間的預測效能差異程度最大，顯示該系統的預測並不能有效投資人進行投資決策時的參考依據。

表 5. 相近取樣期間不同樣本資料集下之 *k* 最近鄰居模型精確率比較

精確率(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	46.04	57.39	56.12	16.59	80.41

類別"跌"	46.52	33.54	44.52	61.11	11.48
類別"漲"	52.63	61.36	40.99	50.00	27.95
所有類別平均	48.40	50.76	47.21	42.57	39.95
類別"漲"、"跌"平均	49.57	47.45	42.76	55.56	19.71
類別標準差	3.67	15.05	7.92	23.17	39.24

表 6．相近取樣期間不同樣本資料集下之 *k* 最近鄰居模型召回率比較

召回率(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	48.00	57.33	43.03	83.33	75.72
類別"跌"	46.32	34.62	38.98	33.33	20.94
類別"漲"	49.21	61.83	57.97	5.55	22.43
所有類別平均	47.84	51.26	46.66	40.74	39.70
類別"漲"、"跌"平均	47.76	48.23	48.47	19.44	21.68
類別標準差	1.45	14.59	10.00	39.42	40.33

表 7．相近取樣期間不同樣本資料集下之 *k* 最近鄰居模型 F-measure 值比較

F-measure(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	47.00	57.36	48.71	27.67	77.70
類別"跌"	46.42	34.07	41.57	43.13	14.61
類別"漲"	50.86	61.59	48.03	10.00	19.74
所有類別平均	48.12	51.01	46.93	41.63	39.51
類別"跌"、"漲"平均	48.65	47.84	45.44	28.80	19.93
類別標準差	2.42	14.82	3.94	16.58	38.71

在表 8、表 9 及表 10 中分別顯示聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集在混合模型下的系統預測精確率、召回率及 F-measure 值。

表 8．相近取樣期間不同樣本資料集下之混合模型精確率比較

精確率(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	50.87	82.50	56.14	18.29	80.41
類別"跌"	54.45	43.89	44.46	56.67	11.48
類別"漲"	47.99	46.91	40.99	55.56	27.95
所有類別平均	51.10	57.77	47.20	43.50	39.95
類別"漲"、"跌"平均	51.22	45.40	42.73	56.11	19.71

類別標準差	3.23	21.47	7.94	21.84	39.24

表 9．相近取樣期間不同樣本資料集下之混合模型召回率比較

召回率(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	53.44	24.00	42.22	91.67	75.72
類別"跌"	50.26	19.41	40.17	25.92	20.94
類別"漲"	42.86	93.11	57.97	11.36	22.43
所有類別平均	48.85	45.51	46.78	42.98	39.70
類別"漲"、"跌"平均	46.56	56.26	49.07	18.64	21.68
類別標準差	5.43	41.29	9.74	42.78	40.33

表 10．相近取樣期間不同樣本資料集下之混合模型 F-measure 值比較

F-measure(%)	聯電新聞資料集	光寶新聞資料集	晶電新聞資料集	立碁新聞資料集	模擬 NewsCATS
類別"持平"	52.12	37.18	48.19	30.50	77.70
類別"跌"	52.27	26.92	42.21	35.57	14.61
類別"漲"	45.28	62.39	48.03	18.87	19.74
所有類別平均	49.95	50.91	46.99	43.24	39.51
類別"漲"、"跌"平均	48.78	50.25	45.68	27.99	19.93
類別標準差	3.99	18.25	3.41	8.56	38.71

在本實驗中我們針對聯電新聞資料集、光寶新聞資料集、晶電新聞資料集與立碁新聞資料集這四組取樣期間相近但新聞取樣標的爲不同股票進行彼此間整體效能差異程度的比較，並和模擬 NewsCATS 系統進行比較，藉以檢驗在實驗 A 中對於本研究所提出的簡易貝氏模型、*k* 最近鄰居模型及混合模型是否同樣適用於其它的樣本資料集合，且同樣能具有改善模擬 NewsCATS 系統的整體預測效能。本實驗顯示，不同的新聞的資料集合雖然彼此的預測效能並不會完全相同，不過透過最適參數的設定可以使其在本研究所提出的簡易貝氏模型、*k* 最近鄰居模型及混合模型中具有較 NewsCATS 系統爲佳的系統預測效能。

五、結論

股價的漲跌變化是由於證券市場中眾多不同投資人及其投資決策後所產生的結果。然而，影響股價變動的因素眾多且複雜，新聞也屬於其中一種，新聞事件不但是投資人用來得知該股票上市公司的相關營運資訊的主要媒介，同時也是影響投資人決定或變更其股票投資策略的主要因素之一。本研究提出以新聞文件做爲股價漲跌預測系統的基礎架構，透過文字探勘技術及分類技術來建置出能預測當日個股收盤股價漲跌趨勢之系統。本研究共提出了簡易貝氏模型、*k* 最近鄰居模型以及混合模型這三種分類模型，並透過實驗來檢驗系統的預測效能。

實驗結果顯示本研究所提出的簡易貝氏模型、*k* 最近鄰居模型以及混合模型這三種分類模型對於系統的整體平均預測效能及對於影響投資人獲利與否的類別"漲"及類別"跌"的平均預測效能都是比相關研究的系統預測效能爲佳，顯示本研究所提出的分類模型可以提供提資人穩定及可靠的預測品質。

參考文獻

[1] Yahoo!奇摩股市，http://tw.stock.yahoo.com/。

[2] 中文斷詞系統，http://ckipsvr.iis.sinica.edu.tw/。

[3] 中央研究院資訊科學所中文組實驗室中文詞知識庫小組，http://godel.iis.sinica.edu.tw/CKIP/index.htm。

[4] 中華民國證券櫃檯買賣中心，http://www.otc.org.tw/。

[5] 方世榮，*統計學導論*，華泰書局，頁 39-81、215-231，1993。

[6] 王春笙，*以技術指標預測台灣股市股價漲跌之實證研究－以類神經網路與複迴歸模式建構*，台灣大學資訊管理研究所碩士論文，1996。

[7] 王疏艷，基於決策樹方法的分類規則的挖掘，海鼎出版，2002，http://hd123.com/asprun/Message/MessageList.asp?gid=17658。

[8] 施正宏，*結合總體經濟指標及個股財報資料以預測個股漲跌－以台灣電子類股爲例*，中原大學資訊管理學系碩士論文，2004。

[9] 曾元顯，"關鍵詞自動擷取技術與相關詞回饋"，*中國圖書館學會會報 59 期*，頁 59-64，1997。

[10] 臺灣證券交易所，http://www.tse.com.tw/。

[11] 謝德宗，*投資學*，華泰書局，頁 235-253、324、403-418，1997。

[12] H. Braun and J. S. Chandler, "Predicting Stock Market Behavior through Rule Induction: An Application of the Learning-from-Example Approach," *Decision Sciences*, volume 18, number 3, pp. 415-429, 1987.

[13] G. P. C. Fung, J. X. Yu and W. Lam, "News Sensitive Stock Trend Prediction," *Proceedings of the Sixth Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 289-296, 2002.

[14] G. Gidófalvi, "Using News Articles to Predict Stock Price Movements," *Technical Report: CSE 254*, Department of Computer Science and Engineering, University of California, San Diego, La Jolla, CA, USA, 2001.

[15] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Second Edition, Morgan Kaufmann, pp. 614-626, 2006.

[16] H. P. Luhn, "The Automatic Creation of Literature Abstracts," *IBM of Research and Development*, pp. 159-165, 1958.

[17] M.-A. Mittermayer, "Forecasting Intraday Stock Price Trends with Text Mining Techniques," Proceedings of the Thirty-Seventh Annual Hawaii International Conference on System Sciences, Track 3, p. 30064b, 2004.

[18] R. P. Schumaker and H.-C. Chen, "Textual Analysis of Stock Market Prediction Using Financial News Articles," *Proceedings of the Twelfth Americas Conference on Information Systems*, Acapulco, Mexico, 2006.

[19] Wikipedia,http://www.wikipedia.org/.

[20] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, Morgan Kaufmann, pp. 76-80, pp. 88-96, pp. 149-151, pp. 296-304, 2000.

[21] B. Wüthrich, V. Cho, S. Leung, D. Permunetilleke, K. Sankaran, and J. Zhang, "Daily Stock Market Forecast from Textual Web Data," *Proceedings of the 1998 IEEE International Conference on Systems, Man, and Cybernetics*, pp. 2720-2725, 1998.