郝沛毅、歐仁彬、黃天受、林振穎、吳建生（2018），『透過新聞文章預測股價漲跌趨勢－結合情緒分析、主題模型與模糊支持向量機』，*中華民國資訊管理學報*，第二十五卷，第四期，頁 363-396。

透過新聞文章預測股價漲跌趨勢－結合情緒分析、主題模型與模糊支持向量機

郝沛毅*
國立高雄科技大學資訊管理系

歐仁彬
國立高雄科技大學資訊管理系

黃天受
國立高雄科技大學資訊管理系

林振穎
國立高雄科技大學資訊管理系

吳建生
國立高雄科技大學資訊管理系

摘要

能夠成功預測股票漲跌趨勢明顯地有許多好處，根據效率市場假設，公司股票的價值是由當前所有可用的信息給定。當分析師、投資者和機構交易者評估當前股價時，新聞在股價估值過程中發揮重要作用。事實上，金融新聞刊載有關於公司基本面的訊息，和影響市場參與者期望的質化訊息。在大數據時代，線上新聞文章的數量持續增長，在如此巨量的文字資料面前，越來越多的機構依靠現代計算機的高速處理能力來進行文字探勘與機器學習，以建構更準確的股價趨勢預測模型。使用文章中非結構化的數據，是最具挑戰性的研究方向，也將是本研究工作的重點，在本論文中，我們將從新聞文章中萃取出隱含的主題模型與情緒資訊，此外，我們將開發一個模糊支持向量機來融合線上新聞文章內含的豐富資訊，以預測股價的漲跌趨勢。我們認為模糊理論非常適用於本研究，因為文字本身就是模糊的（例如，高低、大小），而且在漲跌趨勢之間，存在一條曖昧的模

* 本文通訊作者。電子郵件信箱：haupy@cc.kuas.edu.tw
2017/07/09 投稿；2017/11/17 修訂；2018/06/28 接受

糊邊界（例如，漲 0.01%與漲 1%雖然都屬於上漲的類別，但是屬於的程度明顯不同）。本研究在食品類股的預測正確率最高為 87%，半導體類股的正確率最高為 71%，電腦周邊類股的預測正確率最高為 69%，相較於傳統支持向量機透過關鍵字來預測股價漲跌趨勢的正確率僅五成多（接近於隨機猜測），本研究所提出的方法明顯優於傳統的支持向量機預測模型。

關鍵詞：股價預測、情緒分析、潛在狄利克雷分配、文字探勘、模糊理論、支持向量機

Hao, P.Y., Ou, J.B., Huang, T.S., Lin,Z.Y. and Wu, J.S. (2018), 'Sentiment and topic analysis on financial news for stock movement prediction by using fuzzy support vector machine', *Journal of Information Management*, Vol. 25, No. 4, pp. 363-396.

Sentiment and Topic Analysis on Financial News for Stock Movement Prediction by Using Fuzzy Support Vector Machine

Pei-Yi Hao*
Department of Information Management, National Kaohsiung University of Science and Technology

Jen-Bing Ou
Department of Information Management, National Kaohsiung University of Science and Technology

Tien-Shou Huang
Department of Information Management, National Kaohsiung University of Science and Technology

Zhen-Ying Lin
Department of Information Management, National Kaohsiung University of Science and Technology

Jian-Sheng Wu
Department of Information Management, National Kaohsiung University of Science and Technology

Abstract

Purpose－In Big Data era, the amount of news articles has been increasing tremendously. In front of such a big volume of textual data, more and more institutions rely on the high processing power of modern computers for text mining and machine learning to make more accurate predictions of stock market. Discovering the fundamental data available in unstructured text is the most challenging research aspect

* Corresponding author. Email: haupy@cc.kuas.edu.tw
2017/07/09 received; 2017/11/17 revised; 2018/06/28 accepted

and therefore is the goal of this work.

Design/methodology/approach－In this study, we extracted the hidden topic model and emotional information from news articles. Besides, we developed a fuzzy support vector machine to merge the abundant information from the on-line news, which can be used to forecast the trend of stock prices. Fuzzy set theory is very useful for this study because the texts are fuzzy in itself (such as high/low and big/small), and there is an ambiguous boundary between rise and fall categories. For example, going up either 10% or 1% belongs to rise category, but is different in degree.

Findings－As for this study, the highest forecast accuracy rate was 87% for the food-related stocks, 71% for the semiconductors-related stocks, and 69% for the computer peripheral-related stocks. When compared with traditional support vector machine, which the forecast accuracy rates of stock price trends were just over 50% (nearly to random guess), the method proposed in this study is significantly better than the forecasting model of traditional support vector machine.

Research limitations/implications－This study focused only on accurately classifying the stock movement based on hidden topic and sentiment features. In our future work, we plan to investigate more complex semantic features.

Practical implications－Successful predictions of stock price movement tendency have obvious advantages. According to the Efficient Market Hypothesis, the price of a stock asset is given by all information available in the moment. Financial news carries information about the firm's fundamentals and qualitative information influencing expectations of market participants. This study employs sentiment and topic analysis on financial news to predict stock movement. This can help analysts, investors and institutional traders to effectively evaluate current stock prices.

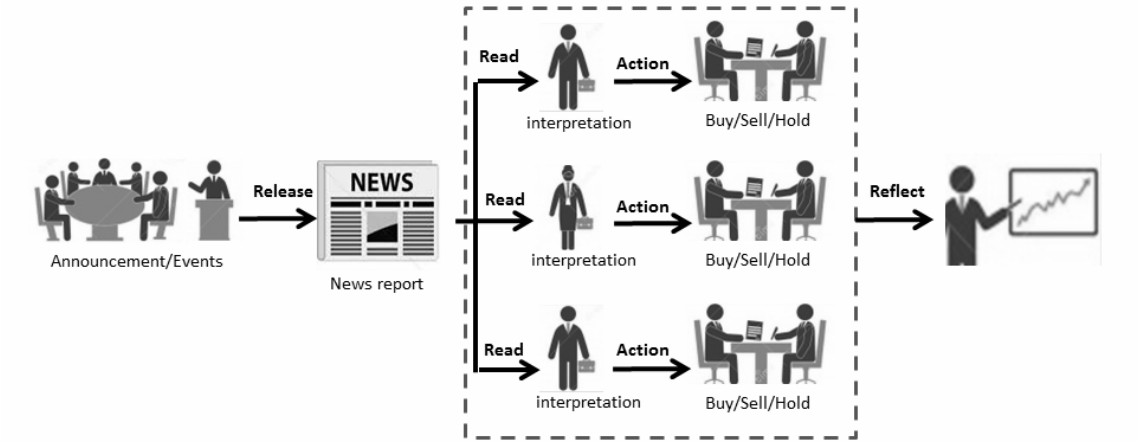
Originality/value－This study is, to the best of our knowledge, the first attempt to apply fuzzy support vector machine and hidden topic/semantic features for the prediction of stock movement in Taiwan.

Keywords: stock trend prediction, sentiment analysis, latent dirichlet allocation, text mining, fuzzy theory, support vector machine

壹、導論

股票市場是當今金融市場中最重要與活躍的一部分，能夠正確地預測股票漲跌的趨勢，明顯地能帶來許多好處。因此，透過資料探勘研究來建構自動化股票投資決策演算法，吸引許多學術界與投資人的關注。大多數預測股票價格未來趨勢的研究，主要依賴於股票市場的歷史數據，例如股票的價格和交易量（例如 Leigh et al. 2002; Lu et al.2009）。雖然，市場的歷史資訊的獲得與處理是相對簡單的，但實際上，它的缺點是忽略了關於真實世界中的重大事件（新聞），例如：併購和收購、獲利的變化、財報表現的升級或降級、管理階層的人事或制度變化等事件的發生，都已被證明對股票趨勢有影響（Ryan & Taffler 2004）。而近期關於公司的重大事件，通常是透過金融新聞提供商，例如路透社和彭博，以新聞文章的格式發布給投資大眾。這些資訊來源提供了真實世界事件的重要信息。市場上的投資者和投機客都希望透過分析市場信息來獲得更好的利潤。然而，在大數據時代，新聞文章的數量一直在增長。在這麼大量的新聞資料面前，越來越多的機構依靠現代計算機的高速處理能力進行分析，建立決策支持系統來提供未來趨勢的預測，並幫助投資者做出更明智的決策。

許多基於金融新聞文章建構的股價趨勢分析系統，已經被報導具有預測能力（Fung et al. 2002; Wu et al. 2009），如圖 1 所示，新聞事件影響股價變動的完整流程是：事件發生後透過財經新聞文章發布，新聞文章首先由投資者解釋並轉化為個人情緒與買入 / 賣出意願，投資者根據意願做出決定，最終，市場價格受到每個投資者的行動所影響，並且反映在最終的股票價格走勢。在本論文中，我們將提出一個基於文字探勘與機器學習的股價趨勢預測模型，本研究將導入新穎的模糊支持向量機作為核心的分類器，以判別股價的漲跌趨勢，我們認為模糊理論非常適合這方面的研究，首先，由於文字語言本身就是模糊的，例如高低、大小、快慢等，因此，Zadeh 教授於 1965 年提出了模糊理論，讓人類語言所表達的規則能夠被電腦如數值般計算（Zadeh 1965）。此外，在股票價格的漲與跌類別之間，存在一個模糊的邊界，公司股價漲 1%與漲 0.01%，雖然都是屬於漲的類別，但是屬於漲的類別的程度並不一樣。因此，在建構分類器的學習過程中，它們應該有不同的重要性與影響力。由於在金融市場中覆蓋著大量的雜訊資料，因此，融合模糊理論可以成功地避免雜訊所造成的不利影響，並且更準確地預測股價漲跌的趨勢。



(1)事件發生；(2)報導事件；(3)投資者閱讀新聞文章；(4)投資者根據自己的知識解釋信息並採取行動；和(5)各種行動被轉換成訂單並反映在股票價格變動中。

圖 1：新聞影響對市場價格產生影響的一般情況

貳、相關研究

一直以來，預測股票市場行為對研究人員有一定的吸引力。雖然已經進行了許多嘗試，但最困難的問題癥結點，是無法預測人類交易者的行為。更糟糕的是，投資者行為模式不斷變化，使得準確的預測變得相當困難。更進一步加深這種不確定性的來源，有幾種完全相反的股票市場研究哲學，例如效率市場假說 / 隨機遊走假說與基本 / 技術分析。最關鍵的問題是：股票價格是可以被預測嗎？以及如果可以的話，該如何預測？

一、效率市場假說與隨機遊走假說

許多研究都試圖解決股票市場是否可以預測的問題。股票市場是完全隨機與不可預測的想法，源於效率市場假說（efficient market hypothesis, EMH）與隨機遊走假說（random walk hypothesis），根據效率市場假說，當前的股票價格充分反映了所有可以獲得的信息（Fama 1991）。因此，價格變化只是由於新的信息或新聞。因為新聞本質上是隨機發生的，而且在現在是不可知的，股票價格應該遵循隨機遊走模式，下一個時間點股票價格的最佳賭注是當前的股票價格。因此，股票價格是不可預測的，預測精確度約 50%（Walczak 2001）。

在另一方面，各種研究指出，股市價格並未遵循隨機遊走模式，並且在某種程度上是可以被預測的。LeBaron 等學者曾經創造了一個模擬交易者的人工股票市場，其中參與的交易者的決策，是可以被操縱和剖析的（LeBaron et

al.1999）。當新的信息引入市場後，LeBaron 調整各個交易者其接收新信息並隨後採取行動之間的決策時間間隔。這項研究提出一個更重要的貢獻，他發現當新信息導入市場和市場自我糾正價格之間，存在一個停滯期。這種明顯的市場延遲行為，有助於抵消市場即時修正理論，並支持當引入新信息後，可以在短時間內預測市場的想法。

二、基本分析與技術分析

相信市場價格至少在某種程度上可以被預測的人，可被分成兩個陣營，那些相信市場變動歷史必定會重演的人，稱為技術分析師。他們深信在市場價格變動的圖表中，存在著某種模式，唯有經驗老道的眼睛才能發現其蛛絲馬跡。基於這種信念，技術分析法（technical analysis）是使用市場過去的歷史資料，以「價」與「量」為基礎，透過圖表型態及計量化的技術指標進行分析。圖表型態分析是透過市場供需交易行為構成的圖表型態，推估未來價格變動之趨勢，例如、道氏理論（Dow Theory）、艾略特波浪論（Wave Principle）、K 線分析及股價排列型態分析等。而技術指標分析是利用價格及成交量為基礎，經特定公式計算出各種指標，運用指標數據預測未來股價變動之趨勢，如移動平均線（moving average; MA）、濾嘴法則（filter rules）、相對強弱指標（relative strength index; RSI）、隨機指標（Stochastic Index）、心理值（PSY）、乖離率（BIAS）、指數平滑異同移動平均線（Moving Average Convergence and Divergence; MACD）、威廉指標（WMS%R）與能量潮（On Balance Volume; OBV）等。雖然，技術分析在許多市場經紀人和參與者中，是最為廣泛被使用的（Bollen et al. 2011; Vu et al. 2012; Si et al. 2013），但是從科學的角度來看，技術分析本身似乎並不吸引人，因為除了說明模式存在之外（如果真的有的話），大多數的技術分析師並不會去解釋模式存在的背後原因。在最近的一項研究中，對於技術分析的有效性和局限性再次進行了測試，並且證明在許多情況下，技術分析指標並不具有很強的預測能力（Yu et al. 2013）。然而，運用各種類型的機器學習算法，對於複雜的財務問題建立預測模型的努力，仍持續進行著，例如運用類神經網路（Sermpinis et al. 2012）、模糊邏輯（Bahrepour et al. 2011）、支持向量回歸機（Premanode & Toumazou 2013）、以及規則式遺傳網路規劃（Mabu et al. 2013）。

另一個被稱為基本分析（fundamental analysis）的學派似乎更有前途，在基本分析中，分析人員查審閱來自不同來源的基本數據，並根據這些數據對市場價格進行估計。我們可以提出至少 5 個基本數據的主要來源：(1)公司的財務數據，如資產負債表中的數據或外匯市場中貨幣的財務數據；(2)有關市場的財務數據，如其指數；(3)關於政府活動和銀行的財務數據；(4)政治情況；(5)自然或非自然

災害等地理和氣象情況。技術與基本分析的差異是其輸入的數據，技術分析使用市場的股價歷史數據，基本分析則是利用關於國家、社會、公司等各種信息或新聞。過去的大部分研究，都是針對技術分析方法進行的，主要是由於市場的歷史數據是可輕易獲得的。然而，基本分析是更具挑戰性，尤其當作為輸入使用的基本數據是文章中非結構化的文字資訊時，而這也是本研究的重點，本研究將使用在財經新聞網站中可獲得的基本數據，來預測股價的漲跌趨勢。

三、從財經新聞預測股價漲跌趨勢

金融分析師在股票市場的工作，是根據對於股票價值的上升或下降的期望，來建議他的客戶買入或賣出股票。這種預期是融合各種的信息來源，並根據這些信息定義股票的價格。當分析師、投資者和機構交易者評估當前股價時，新聞在股價估值過程中，發揮重要作用。事實上，金融新聞刊載有關於公司基本面的訊息，和影響市場參與者期望的質化訊息。如果金融新聞所傳達的新穎信息，將導致對公司現金流或投資者貼現率的預期發生調整，則它會影響股票回報（Nizer & Nievola 2012; Tetlock 2011）。隨著訊息乘載介質的不斷進步，以及網際網路蓬勃發展的推波助瀾下，可用訊息量在過去幾十年中急劇增加。在如此大數據的年代，投資者越來越難以遵循並考慮所有可用的訊息，因此對重要訊息的自動分類變得更加關鍵。然而，對金融新聞文章的自動分類的研究還處於起步階段。對於金融文章的挖掘探勘，現有研究文獻通常依賴非常簡單的文章表示方式，例如詞袋模型（bag-of-words）。甚者，建構文章向量表示式的單詞列表，是基於詞典而創建的，或者是從文章語料庫中實際出現的字詞檢索而得（Schumaker & Chen 2009; Hagenau et al. 2013）。儘管有許多嘗試（Li 2010; Nassirtoussi et al. 2014），在公司財經新聞發布後，股票價格漲跌趨勢的預測準確性很少超過 58%，準確度僅略高於隨機猜測概率（50%），留下了許多實質性改進的空間。考慮到現今金融文章挖掘中使用的方法皆非常簡單，並未使用最先進的方法，本研究預期在兩個方面有改進的潛力：首先，我們需要探索更複雜和高階的表達特徵（例如，主題與情緒），使其可以能夠捕獲新聞文章的基礎語義。第二，這些特徵應當與強健的特徵選擇過程結合，以選擇那些能夠最佳地區分具有影響股價漲或跌效應的新聞消息的特徵。

四、情緒分析與股價漲跌趨勢預測

情緒分析（sentiment analysis），也稱為意見挖掘（opinion mining）、評論挖掘、評價提取或態度分析，它是進行檢測、提取和分類關於不同主題的意見、情緒和態度的任務（Ravi & Ravi 2015）。情緒分析可以幫助實現各種目標，例如觀

察關於政治運動（Li & Li 2013）、產品評論和餐廳評論（Liu & Zhang 2012）、客戶滿意度測量（Kang & Park 2014）、電影銷售預測（Rui et al. 2013）等等的公眾情緒。由於電子商務的蓬勃發展，挖掘使用者對商品的評價和評論中隱含的情緒變得非常關鍵，這也是表達和分析意見的重要來源。最早將財經新聞中的情緒極性與股價趨勢聯繫起來的研究，是 Tetlock 等學者收集了華爾街日報（WSJ）和道瓊斯新聞服務（DJNS）中的新聞文章，他使用哈佛 IV-4 心理社會詞典，將所有詞語分類為積極或消極。他們發現，金融新聞中使用的負面詞彙較多，通常可預測公司的收益是較低（Tetlock et al. 2008）。換句話說，股票市場的商業新聞文章內隱含的情緒，可能會影響投資者的決定，因此商業新聞文章內的情緒是影響股價的另一個重要因素（Yu et al. 2013）。例如，積極的消息可能鼓勵投資者購買股票，從而帶動股價上漲，而負面消息則會產生相反的效果。然而，面對龐大數量的新聞文章，使得投資者難以從每日新聞來源中，找到這樣有用的信息。因此，自動化區分新聞文章的正面和負面情緒，已經成為股票趨勢預測中頗具前途的技術（Schumaker et al. 2012; Li et al. 2014）。

不過，除了文章的情緒之外，新聞文章對於股票價格的影響，可能與新聞所隱含的主題而有所不同，即使它們傳達類似的情緒。例如，常規的每季新產品發表公告，與公司的重大獲利發布，即便這兩個事件通常都帶有積極的語調與正面的情緒，但是這兩個消息可能對股價產生不同的影響，因此，本研究將融合新聞文章的主題與情緒，以獲得更能區分股價漲跌的混合特徵。

五、支持向量機

支持向量機（support vector machine; SVM）是一種以 Vapnik 的統計學習理論為基礎的分類技術（Vapnik 1995），並且具有極優良的推理能力（generalization ability）。SVM 不像傳統的機器學習（machine learning）技術以最小化經驗風險（empirical risk）為目標－即使得訓練資料的分類誤差最小化，SVM 以最小化結構風險（structural risk）為目標－即使得未知的資料（測試資料）的分類誤差在一個機率上界以下。根據統計學習理論，支持向量機建構出一個具有最大邊界（margin）的最佳超平面來分割兩個類別的資料，Vapnik 亦證明了最大化邊界等同於最小化推理誤差的上界（Vapnik 1995）。本研究將導入新穎的模糊支持向量機，作為核心的分類器，並用它來判別股價的漲跌趨勢。我們預期模糊理論能從兩個方向改善股價趨勢預測的能力，首先，文字語言本身就是模糊的，例如高低、大小、快慢等，同時還有同義字與一字多義的問題，因此，需要模糊理論能夠處理曖昧不精確的現實世界的特性的能力。此外，在股價漲跌趨勢預測的分類問題中，公司股票的漲與跌類別之間存在一個模糊的邊界，公司股價漲 1%與漲

0.01%，雖然都是屬於漲的類別，但是屬於漲的類別的程度並不一樣，也就是說，在建構分類器的學習過程中，它們應該有不同的重要性與影響力。由於在金融市場中覆蓋著大量的雜訊資料，因此融合模糊理論可以成功地避免雜訊所造成的不利影響。

參、研究方法

效率市場假說（efficient-market hypothesis; EMH）指出，股票的價值是由當前時刻所有可用的信息所給定。然而，單一金融分析師不可能完全知道目前所有最新發布的新聞消息，因此，應用文字探勘技術自動化處理大量與即時的新聞文章，可以幫助分析師和投資者掌握投資的先機，並且帶來可觀的獲利。在本論文中，我們將應用文字探勘與機器學習技術，自動化分析在新聞網站上巨量的線上文章資料，我們將從新聞文章中萃取高階的主題與情緒資訊，融合這些資訊後，將導入新穎的模糊支持向量分類機，來預測股票價格未來變動的趨勢，詳細流程請參見圖 2。

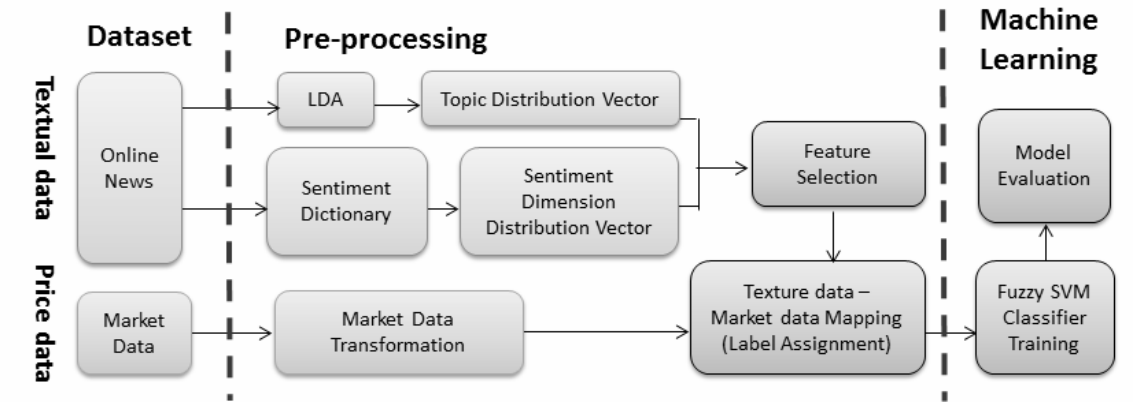


圖 2：系統流程圖

一、資料蒐集

本研究試圖透過文字探勘與機器學習的技術來預測台灣公司股票價格的變動趨勢。要建構自動化股價漲跌趨勢預測模型的分類器，必需預先準備標記好類別標籤的訓練資料集。然而，新聞網站所張貼的文章，本身並不帶有可用於分類目標的類別標籤。因此，為了判別新聞文章對股票價格的影響，我們將結合文字資料與股市的價格資料（如圖 2 所示），並考慮新聞發佈的時間點，以新聞公佈特定公司的消息之後，該公司股價的開盤與收盤價格之間的差距，代表該新聞對於

股價的影響。此外，根據 Li 等學者研究所指出，當股票市值大時，日交易量和參考新聞文章數量之間的相關性變得更大，這進一步意味著財經新聞積極報導股票交易活躍的公司的新聞，而交易量小的公司則新聞報導數量較小（Li et al. 2014）。因此，我們刪除低交易量的股票，並且針對不同的公司族群: 包含食品類股（味全、統一、南僑等公司）、半導體類股（台積電、聯電、聯發科等公司）與電腦周邊類股（宏碁、華碩、英業達等公司）進行研究，以探討文字探勘技術在不同類型公司股票之間的預測能力。

二、特徵挑選

當分析師、投資者和機構交易者評估當前股價時，新聞在股價估值過程中，發揮重要作用。金融新聞刊載有關於公司基本面的訊息，和影響市場參與者期望的質化訊息。然而，對於財經新聞文章的挖掘探勘，過去研究文獻通常依賴非常簡單的文章表示方式，例如，使用詞袋模型（bag-of-words）的文章向量表示式。因此，本研究將探索更複雜和高階的表達特徵（例如，主題與情緒維度），使其可以能夠捕獲文章消息的基礎語義。本研究利用情緒分析以及潛在狄利克雷分配（latent dirichlet allocation; LDA）來獲得文章中隱含的高階特徵，我們首先會對新聞文章執行情緒分析，透過 C-LIWC 中文詞典來獲得新聞文章所攜帶的情緒詞的類別（如正面情緒詞、負面情緒詞、情感歷程詞、焦慮詞、生氣詞、悲傷詞及認知歷程詞等），此外，文章除了作者表達出的情感外，文章的主題內容更是傳達出豐富的資訊，本研究利用 LDA 獲得能夠代表文章內容的主題模型，再使用粒子群演算法過濾預測效果不佳的主題，並將新聞的主題模型結合情緒分析結果來預測股票的漲跌趨勢。

（一）新聞文章的情緒維度強度分布特徵

新聞文章對股票市場影響的其中一個潛在因素，是新聞文章所傳達的情緒或所使用的語氣。過去的研究已經證實，積極的消息可能鼓勵投資者購買股票，從而刺激股價上漲，而負面消息則會產生相反的效果。本研究將使用情緒分析，透過情緒維度來對新聞文章建構向量模型，以代替傳統透過關鍵字頻率建構的文章向量。文章中的每個字詞（特別是具有情感極性彩色的）將被分解數個情緒特徵，並且表示為情緒特徵組成的向量。本研究採用中文 C-LIWC 詞典[1]，該詞典是黃金蘭與林以正等學者基於 C-LIWC 詞典所建立的中文詞典（黃金蘭等 2012），C-LIWC 詞典刪除中文不適用的類別，並且加入中文特有的類別（如數量單位詞、語助詞），其包含 30 個語言類別及 42 個心理類別，表 1 摘錄 C-LIWC 中部

1 http://cliwc.weebly.com/

分的語言類別與範例。

表 1：C-LIWC 詞典中的語言類別與範例

類別	縮寫	數目	範例	類別	縮寫	數目	範例
正向情緒詞	posemo	564	信心、滿足、祝福	負向情緒詞	negemo	925	擔憂、猜疑、報復
情感歷程詞	affect	1563	氣憤、感恩、失望	認知歷程詞	cogmech	1248	理解、選擇、質疑
焦慮詞	anx	129	不安、掙扎、緊繃	生氣詞	anger	284	可惡、抱怨、破壞
悲傷詞	sad	165	心痛、沮喪、無力	否定詞	negative	262	不要、未必、沒有
概數詞	quant	103	一些、所有、眾多	休閒詞	leisure	335	唱歌、輕鬆、假期
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
洞察詞	insight	327	了解、恍然大悟、體會	停頓贅詞	nonfl	7	呃、然後、那

過去情緒分析研究大多使用二元情緒來分析文章，然而實際上，文章不只含有正向與負向這兩種情緒，而是由多種情緒混雜而成，為了獲得更佳豐富的情緒資訊，本研究利用 C-LIWC 詞典中多種不同的情緒類型，包含正向情緒詞、負向情緒詞、情感歷程詞、焦慮詞、生氣詞、認知歷程詞及悲傷詞等七種情緒詞。我們認為基於情緒維度表達的文章向量，比傳統詞袋模型有以下的優點：

- 減少維度：使用關鍵字作為特徵時，通常需要數萬個關鍵字特徵，與之相比，情感維度表示式將大幅地將尺寸減小，例如，本研究使用 C-LIWC 詞典的正向情緒詞、負向情緒詞、情感歷程詞、焦慮詞、生氣詞、認知歷程詞及悲傷詞等七種情緒詞，共有七個情緒維度；
- 情感解釋：通常很難解釋在詞袋空間中生成的文章向量模型。相比之下，不同維度上的情緒分數，有時可以讓人們對文章有更直觀的感受。

我們舉一個簡單的例子，說明如何將新聞文章轉換成情緒維度強度分布的特徵向量，考慮表 2 顯示的財經新聞文章。該篇新聞文章出現的關鍵字中，有 9 個被 C-LIWC 詞典收錄為正向情緒詞（posemo），2 個為負向情緒詞（negemo），12 個情感歷程詞（affect），1 個悲傷詞（sad），35 個認知歷程詞（cogmech），0 個焦慮詞（anx）與 2 個生氣詞（anger），故此範例文章的情緒維度分布向量為（9, 2, 12, 1, 35, 0, 2）。注意，僅使用情緒特徵維度，可能無法有效的預測股票價格的

變動趨勢。例如，常規的每季新產品發表公告，與公司的重大獲利消息發布，這兩個事件通常都帶有相似的積極語調與正面情緒，但是，明顯地這兩個消息對股價會產生不同的影響，因此，本研究將整合文章隱含的主題模型與情緒維度特徵，並融合這些資訊代入我們的股價趨勢預測模型。

表 2：新聞文章的情緒詞分布

美國掀起穿戴攝影風潮，GoPro、安霸創歷史收盤新高							
情緒詞類別	posemo	negemo	affect	sad	cogmech	anx	anger
計數	9	2	12	1	35	0	2

MoneyDJ 新聞　2014-09-09　記者　賴宏昌　報導

marketwatch.com 8 日報導，FBN 證券分析師 Shebly Seyrafi 初評多功能攝影機製造商 GoPro Inc.（NASDAQ 股票代號：GPRO）、給予「表現優於指數」評等，目標價設定在 70 美元。Seyrafi 指出，GoPro 去年就有 380 萬台的銷售成績、顯示這家公司未來還會有好幾年的強勁成長榮景。Seyrafi 同時也看好 GoPro 進軍美國以外市場以及數位內容管理領域。

CNNMoney 報導，Seyrafi 指出，去年全球數位相機、數位攝錄影機銷售量超過 9 千萬台，顯示 GoPro 未來還有很大的成長空間。barrons.com 部落格報導，JP 摩根證券分析師 Paul Coster 上週將 GoPro 評等自「加碼」降至「中立」。彭博社 6 日報導，安全監控攝影機系統單晶片開發商安霸（Ambarella, Inc.；股票代號：AMBA.us）2015 會計年度第 2 季（2014 年 5-7 月）本業每股盈餘自 2-4 月的 0.25 美元升至 0.37 美元、優於市場預期的 0.28 美元。Ambarella 預估本季（8-10 月）營收將達 6,000 萬-6,400 萬美元，優於市場預期的 5,170 萬美元。

Ascendiant Capital Markets LLC 分析師 David Williams 指出，Ambarella 目前約有 20-25%營收是來自 GoPro，未來可望受惠於美國密蘇里州佛格森黑人少年意外遭警察槍殺事件所引發的安全監控攝影需求。GoPro 執行長 Nicholas Woodman 8 月 1 日在接受電話訪問時透露，像群光（2385）這樣的代工廠商就可以滿足 GoPro 的需求。

GoPro 8 日大漲 8.12%、收 63.52 美元，創 6 月 26 日初次公開發行（IPO）以來收盤新高；較 IPO 價格（24.00 美元）高出 165%、8 日收盤市值達 80 億美元。鴻海在 6 月底代子公司 Foxteq Holdings Inc.公告，6 月 26 日以每股 24 美元的價格處分 GoPro 737,079 股、處分利益為 5,642,542.43 美元；累積持有數量為 10,972,248 股、持股比例 8.88%。

Ambarella 8 日上漲 5.50%、收 38.93 美元，創歷史收盤新高；今年迄今漲幅達 14.87%。

（二）新聞文章的主題機率分布特徵

過去透過新聞文章預測股價趨勢的研究文獻，通常使用非常簡單的文章表示

方式，例如詞袋模型（bag-of-words）。它們透過 TF-IDF 或是 information gain 來挑選具有鑑別能力的關鍵字，並且使用這些關鍵字來建立文章向量。然而，僅單純使用個別的關鍵字，可能會忽略文章中更高階層的主題資訊。舉例來說，僅使用關鍵字「上揚」，可能沒有辦法鑑別股價的漲跌趨勢，因為有可能是成本上揚或是獲利上揚，所以底層的關鍵字「上揚」，在關鍵字篩選階段就可能被剔除。但是，關鍵字「成本」與「上揚」結合在一起，對於股價的漲跌就有鑑別能力了，因為他揭露了成本上揚將導致獲利縮減的主題。所謂的主題（topic），就是指能夠闡述高階語義概念的關鍵字集合，例如，{成本,上揚}或{獲利,上揚}。

本研究將使用潛在狄利克雷分配（latent dirichlet allocation; LDA）主題生成模型（Blei et al. 2003）來產生主題特徵，這一種嘗試發現文章集合中抽象主題的方法。其背後的直覺是，影響股價變化的新聞文章可能表現出特定的抽象主題，而這些主題可以被捕獲，並且應用於新聞文章的股價漲跌趨勢分類。LDA 之基礎概念為：每一個文件是由數個主題所構成的機率分佈，而每一個主題又是由數個關鍵字所構成的機率分佈。圖 3 顯示 LDA 的生成模型，其中，θ_i 為文件主題分佈，$i \in \{1,...,M\}$，ϕ_k 為主題 k 的單字分佈，α 為每一篇文件的主題的事前狄力克雷參數，β 為每一篇主題的單字分佈的事前狄力克雷參數，z_{ij} 為文件 i 中第 j 個單字的主題，w_{ij} 則是指特定的單字。α 和 β 分別為 θ 和 ϕ 的狄力克雷參數，M 為文章總篇數及 N 為文件的單字總數。在 M 篇文件中，從該文件所對應的 θ 抽出一個主題 Z，再從主題 Z 中所對應的 ϕ 抽出一個單字 W，重複上述過程 N 次，就產生主題 k 及主題 k 所包含之單字。

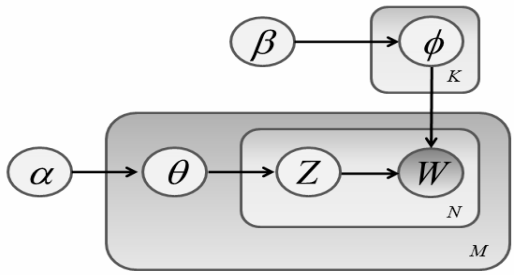


圖 3：LDA 生成模型

本研究將以 JAVA 版本的 LDA 程式（即 JGibbLDA[2]）來獲得隱藏的主題模式，由於不同類型的公司，其新聞文章中所出現的主題也不盡相同，因此本研究將分別探討食品類股、半導體類股與電腦周邊類股的新聞文章，並從中擷取文章內出現的主題，表 3 顯示針對食品類股的中文財經新聞執行 LDA，並且設定主題

2 http://jgibblda.sourceforge.net/

數目為 30 時，所得到的結果，它列出主題 Topic 1-Topic 30 中對應的關鍵字列表，以及這些關鍵字對應該主題的機率值（我們僅列出機率最大的前 30 個關鍵字），由表 3 中各個主題所對應的關鍵字可知，主題 Topic 1 應與食品銷售特賣活動有關，主題 Topic 2 應與公司股東會改選有關，主題 Topic 3 應與食品黑心油事件相關，主題 Topic 30 應與公司經營權變動有關。

表 3：LDA 找出的潛在主題

Topic 1		Topic 2		Topic 3		…	Topic 30	
關鍵字	機率值	關鍵字	機率值	關鍵字	機率值	…	關鍵字	機率值
成長	0.03357132	新任	0.02114803	消費者	0.0358528		台灣	0.0311147
業績	0.01838067	政府	0.01417615	採購	0.0234362		集團	0.0311147
商品	0.01686161	董事	0.01185219	食品	0.020332		員工	0.0125695
營收	0.01534254	交易	0.00952823	事件	0.0172279		可能	0.0125695
年增	0.01078535	現任	0.00952823	黑心油	0.0156759		投資	0.0125695
持續	0.01078535	食品	0.00720427	通報	0.0141238		抵制	0.0105089
銷售	0.00926629	今天	0.00720427	主管	0.0125717		經營權	0.0084483
商機	0.00926629	知道	0.00720427	生產線	0.0110197		行動	0.0084483
超商	0.00926629	指派	0.00720427	價格	0.0110197		股東	0.0084483
帶動	0.00926629	專長	0.00720427	檢驗	0.0094676		計畫	0.0084483
春節	0.00926629	漲停	0.00488031	封存	0.0094676		併購	0.0063878
活動	0.00774722	進行	0.00488031	機關	0.0094676		營運	0.0063878
增加	0.00774722	造成	0.00488031	油品	0.0079155		維持	0.0063878
優惠	0.00774722	官股	0.00488031	製成	0.0079155		不變	0.0063878
合併	0.00774722	內線	0.00488031	回收	0.0079155		通知	0.0063878
表示	0.00622816	消基會	0.0048803	認證	0.0063634		服務	0.0063878
去年	0.00622816	異常	0.00488031	供應	0.0063634	…	國內	0.0063878
顯示	0.00622816	出任	0.00488031	原料	0.0063634		態度	0.0043272
食品	0.00470909	介入	0.00488031	疑慮	0.0063634		媒體	0.0043272
同步	0.00470909	代表人	0.0048803	服務	0.0063634		事宜	0.0043272
使用	0.00470909	董監	0.00488031	合作	0.0063634		抵制	0.0043272
合作	0.00470909	治理	0.00255635	退貨	0.0048114		席捲	0.0043272
掌握	0.00470909	立法院	0.0025563	維護	0.0048114		詢問	0.0043272
轉投資	0.00470909	進一步	0.00255635	權益	0.0048114		災難	0.0043272
熱潮	0.0047090	改選	0.00255635	香豬油	0.0048114		淡化	0.0043272
高峰	0.00470909	完全	0.00255635	聲明	0.0048114		發出	0.0043272
鮮食	0.00470909	引爆	0.00255635	ISO	0.0048114		證交所	0.0043272
暖食	0.00470909	立委	0.00255635	供應商	0.0048114		市場	0.0043272
類別	0.00470909	財委會	0.00255635	不良	0.0048114		成本	0.0043272
年菜	0.00470909	輪番	0.00255635	SGS	0.0048114		決定	0.0043272
檔期	0.00470909	背書	0.00255635	客服	0.0032593		出包	0.0022666
連鎖	0.00319003	門神	0.00255635	專線	0.0032593		怒火	0.0022666

表 3 可以幫助我們計算文章屬於某特定主題的機率值，以表 4 為例子，該篇新聞文章是由多個主題混合而成，而對應不同主題的關鍵字，我們使用不同的顏色標記出來。使用表 3，我們可以知道該篇新聞文章屬於各個主題的機率分布，如表 4 所顯示。由表 4 的文章內容可知，它與黑心油事件以及經營權改組是較為相關的，所以該文章屬於主題 Topic 3 與 Topic 30 的機率較高，而屬於 Topic 1 與 Topic 2 的機率則是較低的，本研究將以新聞文章中包含各主題出現的機率值，建構以主題為基礎的文章向量。然而，LDA 是採取非監督式的策略從大型語料庫中生成隱藏的主題，為了篩選能真正鑑別股價漲跌趨勢的主題，本研究將搭配粒子群演算法來進行主題特徵篩選。

表 4：新聞文章的主題分布

魏應充曾表示味全董事長接任者不姓魏。

頂新集團黑心油風暴延燒，民眾及縣市政府、學校群抵制，味全（1201-TW）形象重創、股價大跌，全國教師工會總聯合會 呼籲頂新魏家退出 台灣食品業，據指出，魏家已在研究退出 味全經營權的可能，轉讓所有持股引進新的經營 團隊，針對這項消息，味全今（14）日指出，並未接到任何相關的訊息。頂新集團 1998 年回台入主味全，取得經營權，根據公開觀測站資料，截至今年 9 月，頂新魏家持有味全股權約 40.4%，法人持股部分，匯豐 高林基金以 46576 張，持股 9.2%最多，另外，富邦人壽持股 2.5%，德意志銀行持股 2.38%、三商人壽 持股約 1.7%。

頂新集團以製油起家，在中國大陸發跡後鮭魚返鄉，取得味全經營權之外，還跨足國內地產、電信等領域，但是近年來 3 次油品出包，頂新全都有份，引來民眾怒火發動抵制，甚至傳出銀行開始緊縮銀根，重新考慮融資案的聲音，處在風暴核心的味全股價再度遭重擊，今日持續跌停開出，統計近 1 年來，市值縮水高達 145 億元，減幅 51%。

黑心油風暴重創頂新集團形象，在民眾的抵制下，包括台北 101 明年的改選，購買中嘉案，及三重開發案都可能受到影響，另外，剛起步的台灣之星 4G 業務也遭到波及，味全高達 200 多項產品遭到抵制，除了消費者拒買之外，各縣市政府也紛紛決定將味全及頂新產品逐出校園。

在這起事件中，讓味全遭遇相當大的困境，60 年的招牌及 6000 名員工都可能受到影響，市場傳出，為了挽救味全，保護員工生計，魏家兄弟已在評估出脫持股，引進新經營團隊的可能。日前魏應充出面道歉時也曾表示，未來味全董事長不會姓魏，似乎已透露出些許端倪。

主題編號	Topic 1	Topic 2	Topic 3	....	Topic 30
機率值	0.041167	0.050895	0.157225	....	0.229137

（三）粒子群演算法進行特徵挑選

粒子群演算法（particle swarm optimization; PSO）是由 Kennedy 和 Eberhart 在 1995 年提出（Kennedy & Eberhart 1995），它是以鳥群覓食的社會行為之概念，所建構出的最佳化演算法。如同進化演算法（evolutionary algorithms）一般，PSO 透過對一個族群（稱為 Swarm，群體）中的個體（稱為 Particle，粒子）進行疊代運算，以找出 D 維空間中的最佳解。PSO 假設鳥群在一個區域中任意飛行尋找食物，鳥會由身邊搜尋食物，並向群體傳遞自身區域的食物，鳥群會透過共享群體內的知識，得知哪些區域的食物數量最多，並且傾向於往食物多的區域搜索，透過不斷的飛行搜索，鳥群就會得到最佳的食物地點。在 PSO 中，每個粒子（Particle）代表一個候選位置（即，問題解），粒子是 D 維空間中的一個點，其狀態可以根據其位置與速度來表徵，粒子 i 在第 t 次疊代時的 D 維位置可以表示為 $x_i^t = \{x_{i1}^t, x_{i2}^t, ..., x_{iD}^t\}$，同樣地，速度（即距離的變化）也是個 D 維向量，粒子 i 在第 t 次疊代時的速度可以表示為 $v_i^t = \{v_{i1}^t, v_{i2}^t, ..., v_{iD}^t\}$，每個粒子皆會透過適應函數知道自己的適應值（ftness value），且每個粒子皆會記憶過往自身最佳的適應值為區域最佳解（pbest），而群體（swarm）會記憶群體中最佳的適應值為全域最佳解（gbest）。每個粒子會根據兩個因素來改變其搜尋的方向，即自身過去最好的經驗（pbest）和所有其他成員的最佳經驗（gbest），Shi 與 Eberhart（1998）稱 pbest 為認知因素（cognition part），而 gbest 為社會因素（social part）。令 $p_i^t = \{p_{i1}^t, p_{i2}^t, ..., p_{iD}^t\}$ 表示粒子 i 至疊代 t 時最好的解（pbest），而 $p_g^t = \{p_{g1}^t, p_{g2}^t, ..., p_{gD}^t\}$ 表示至疊代 t 時群體最好的解（gbest）。每次疊代時，粒子會利用公式(1)來更新速度，並由更新速度來獲得新的移動位置，如公式(2)所示

$$v_{id}^{t+1} = w \cdot v_{id}^t + c_1 \cdot rand() \cdot (p_{id}^t - x_{id}^t) + c_2 \cdot rand() \cdot (p_{gd}^t - x_{id}^t), d = 1, 2, ..., D \quad (1)$$

$$x_{id}^{t+1} = x_{id}^t + v_{id}^{t+1}, d = 1, 2, ..., D \quad (2)$$

其中 c_1 為認知因素的學習因子，c_2 為社會因素的學習因子，$rand()$為[0,1]區間內的隨機值，此隨機值使群體在搜尋最佳覓食地點時富有多樣性，w 為慣性權重，它是由 Shi 與 Eberhart 學者於 1998 年提出（Shi & Eberhart 1998），透過此慣性權重變數，使粒子更富有探索能力，透過上述幾個變數能夠使得粒子的移動更趨向不同的搜尋方向，例如設置較高的 c_1 及 c_2 分別能夠更加趨近區域最佳解及全域最佳解，較高的 w 值會使粒子群演算法傾向全域搜索，較低的 w 值使其更傾向局部搜索，粒子群演算法的流程圖如圖 4 所示。

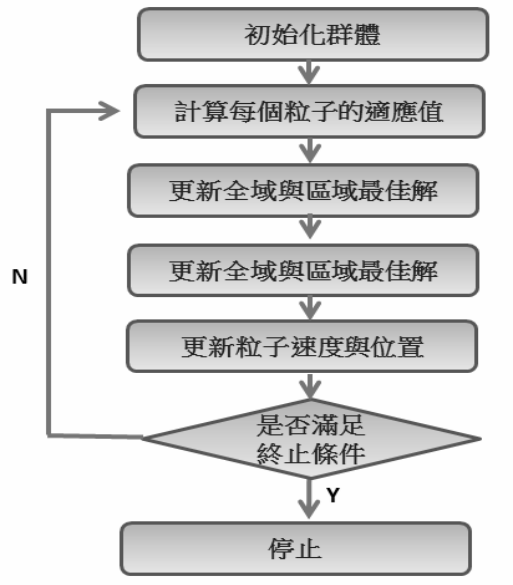


圖 4：粒子群演算法流程圖

由於本研究透過 LDA 獲取大量的主題數量，但是並非每一個特徵都有足夠的辨識能力，而且過多的特徵會使分類器過度學習（overfitting）。因此，本研究藉由特徵選取（feature selection）將不足夠辨識的特徵刪除，以提高預測的準確率。本研究將利用二元粒子群演算法（binary particle swarm optimization; BPSO）進行特徵選取，並挑選出擁有最佳辨識能力的特徵組合，使股票趨勢預測達到最佳化。二元粒子群演算法是 Kennedy 和 Eberhart 於 1997 年提出（Kennedy & Eberhart 1997），為了完成特徵選取的任務，如果目標是從 n 個特徵（主題）中挑選最佳的特徵子集合，則我們必須使用 n 個決策變數，x_{id}, d=1,…,n，其中，x_{id} 的值為在{0,1}中的二元變數，如果 x_{id}=0，代表對應的特徵 d（主題 d）沒有被選取，反之，如果 x_{id}=1，代表對應的特徵 d（主題 d）有被選取。所以，一個粒子在狀態空間中的移動被限制在每個維度上只能是 0 或 1，而每個 v_{id} 代表 x_{id} 取值為 1 的機率，換句話說，如果 v_{id}=0.2，則 x_{id} 有 20%的機率取值為 1，80%的機率其值為 0，如果之前該位元的最佳解為 0（即 p_{id}=0），則 $p_{id} - x_{id}$ 的值可以為 0 或-1，並且用來對下一步的 v_{id} 機率變化進行加權。總結來說，粒子的速度更新公式維持不變（如公式(1)），只是現在限制 x_{id}、p_{id} 與 p_{gd} 的值為在{0,1}中的整數，而 v_{id} 表示機率，所以其值限制在[0,1]的區間中，這項要求可以透過 Sigmoid 函數完成，亦即是說，v_{id} 會透過 Sigmoid 函數轉換為介於 0 至 1 之間的機率值，Sigmoid 函數如公式(3)：

$$S(v_{id}^{t+1}) = \frac{1}{1+e^{-v_{id}^{t+1}}} \tag{3}$$

新的粒子位置則是會根據速度機率值與 *rand*()函數來判斷，位置更新公式如下：

$$if\,(rand() < S(v_{id}^{t+1}))\ then\quad x_{id}^{t+1} = 1;\ \ else\quad x_{id}^{t+1} = 0; \tag{4}$$

三、使用以模糊超平面為基礎的模糊支持向量機作為核心分類器

本研究將提出使用最大模糊邊界支持向量機作為我們的核心分類器，支持向量機由於具有優異的分類預測能力，因此經常被應用在文字探勘的任務，而我們所提出的最大模糊邊界支持向量機，不但使用統計學習理論中的最大邊界來降低推理誤差的機率上界，同時使用最佳模糊超平面來分割正負類別的資料，這將同時融合統計學習理論的優異推理能力與模糊理論適合處理雜訊資料的優點（Hao 2016）。我們認為模糊理論是很適合作為股價趨勢變動預測的應用，因為股價的漲與跌之間，就是存在一條模糊的分割線，股價漲 3%與漲 0.03%，雖然都是屬於漲的類別，但是屬於漲的類別的程度值並不一樣。假設給定一組訓練資料$\{\mathbf{x}_i,\ y_i,\ \mu_i\}$, i=1,…,N, $y_i \in \{-1,1\}$, $\mu_i \in (0,1]$，其中 $\mathbf{x}_i \in R^n$ 代表第 i 筆資料向量，y_i 為其對應的類別標籤。每一筆資料向量皆賦予一個模糊歸屬程度（fuzzy membership），以 μ_i 表示，它代表資料向量 $\mathbf{x}_i$ 屬於該對應類別的信心強度（Lin & Wang 2002），本研究將找出一個模糊超平面 $\langle \mathbf{W} \cdot \mathbf{x} \rangle + \mathbf{B} = \Theta$ ，它以最大的邊界（margin）來分隔正負類別，其中 Θ 代表「模糊零」，它是一個三角形模糊數（triangular fuzzy number），其中心值為 0，寬度為 O_w，而模糊超平面中要被估計的參數，例如權重向量 **W** 中的元素與偏移量 **B**，亦皆設定為三角形模糊數字。模糊分割超平面中的模糊權重向量定義為 **W**=(**w**, **c**)，其中 $\mathbf{w} = [w_1,...,w_n]^t$ 與 $\mathbf{c} = [c_1,...,c_n]^t$ ，表示近似於 **w**，模糊度為 **c**。模糊偏移量定義為 **B**=(b, d)，代表近似於 b，模糊度為 d。模糊超平面是由底下的歸屬函數所描述：

$$\mu_Y(y) = \begin{cases} 1 - \dfrac{\left| y - \left(\langle \mathbf{w} \cdot \mathbf{x} \rangle + b \right) \right|}{\langle \mathbf{c} \cdot |\mathbf{x}| \rangle + d} & \mathbf{x} \neq 0 \\ 1 & \mathbf{x} = 0, y = 0 \\ 0 & \mathbf{x} = 0, y \neq 0 \end{cases} \tag{5}$$

其中 $\mu_Y(y)=0$ 當 $\langle \mathbf{c}\cdot|\mathbf{x}|\rangle+d \le \left|y-\left(\langle \mathbf{w}\cdot\mathbf{x}\rangle+b\right)\right|$，表示近似於 $\langle \mathbf{w}\cdot\mathbf{x}\rangle+b$，模糊度為 $\langle \mathbf{c}\cdot|\mathbf{x}|\rangle+d$。對於兩個三角形模糊數字 $A=(m_A,c_A)$ 與 $B=(m_B,c_B)$，其中 m 為中心，c 為寬度，則 A 大於 B，標記為 $A \underset{f}{\ge} B$，若且唯若

$$m_A+c_A \ge m_B+c_B \text{ 與 } m_A-c_A \ge m_B-c_B \tag{6}$$

這組訓練資料被稱為「模糊線性可分割（fuzzy linear separable）」，若且唯若底下這個不等式成立

$$y_i\left(\langle \mathbf{W}\cdot\mathbf{x}_i\rangle+\mathbf{B}\right) \underset{f}{\ge} \mathrm{I}_F \tag{7}$$

其中 $\mathbf{I}_F$ 代表「模糊壹」，它也是一個三角形模糊數，以 1 為中心，I_w 為寬度，根據公式(5)-(6)，要找出一個具有最大邊界的模糊超平面來最佳地模糊分割正負兩個類別，等同於求解下列二次最佳化問題（quadratic programming problem; QPP）：

$$\underset{\mathbf{w},\mathbf{c},b,d,\xi_{1i},\xi_{2i}}{\text{minimize}} \quad J=\frac{1}{2}\|\mathbf{w}\|^2+C\left(v\left(\frac{1}{2}\|\mathbf{c}\|^2+d\right)+\frac{1}{N}\sum_{i=1}^{N}\mu_i(\xi_{1i}+\xi_{2i})\right) \tag{8}$$

$$\text{subject to } y_i\left(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle+b\right)+\left(\langle \mathbf{c}\cdot|\mathbf{x}_i|\rangle+d\right) \ge 1+\mathrm{I}_w-\xi_{1i}$$

$$y_i\left(\langle \mathbf{w}\cdot\mathbf{x}_i\rangle+b\right)-\left(\langle \mathbf{c}\cdot|\mathbf{x}_i|\rangle+d\right) \ge 1-\mathrm{I}_w-\xi_{2i} \text{ and } d\ge 0,\ \xi_{1i},\xi_{2i}\ge 0 \text{ for } i=1,\ldots,N.$$

其中 $\|\mathbf{w}\|^2$ 表示模型的複雜度，最小化 $\|\mathbf{w}\|^2$ 保留統計學習（statistical learning）的基本精神–要獲得較佳的推理能力（generalization ability），則必須同時降低分類模型的複雜度與訓練誤差，同時，最小化 $\|\mathbf{w}\|^2$ 亦等同於最大化分割正負樣本的邊界（marign），而 $\frac{1}{2}\|\mathbf{c}\|^2+d$ 則表示模型的模糊程度，預測模型越模糊，則預測結果越不精確，而參數 v 是二者之間的調控參數。差額變數 $\{\xi_i\}_{i=1,..,N}$ 測量限制條件(7)被違反的程度，而參數 C 則是使用者給定的懲罰參數，C 值越大越不允許限制條件被違反。模糊歸屬程度 μ_i 表示訓練樣本點 $\mathbf{x}_i$ 屬於所對應類別的信心程度，在本研究中，股價漲（或跌）的幅度越大，它屬於漲（或跌）類別的信心程度越大。根據拉格朗日（Lagrangian）理論，我們得到底下的對偶問題（dual problem）：

$$\underset{\alpha_{1i},\alpha_{2i}}{\text{maximize}} \quad \frac{-1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} y_i y_j(\alpha_{1i}+\alpha_{2i})(\alpha_{1j}+\alpha_{2j})\langle \mathbf{x}_i \cdot \mathbf{x}_j \rangle$$
$$-\frac{1}{2Cv}\sum_{i=1}^{N}\sum_{j=1}^{N}(\alpha_{1i}-\alpha_{2i})(\alpha_{1j}-\alpha_{2j})\langle |\mathbf{x}_i|\cdot|\mathbf{x}_j| \rangle + \sum_{i=1}^{N}(\alpha_{1i}-\alpha_{2i})\mathrm{I}_w + \sum_{i=1}^{N}(\alpha_{1i}+\alpha_{2i}) \tag{9}$$
$$\text{subject to} \quad \sum_{i=1}^{N} y_i(\alpha_{1i}+\alpha_{2i})=0,\ \sum_{i=1}^{N}(\alpha_{1i}-\alpha_{2i}) \le Cv,\quad \alpha_{1i},\alpha_{2i} \in \left[0,\frac{C\mu_i}{N}\right]\ \ i=1,...,N$$

求解出上式後，我們得到拉格朗日乘數（Lagrange multipliers）α_i，權重向量 **w** 與 **c** 是 $\mathbf{x}_i$ 與 $|\mathbf{x}_i|$ 的線性組合：

$$\mathbf{w}=\sum_{i=1}^{N} y_i(\alpha_{1i}+\alpha_{2i})\mathbf{x}_i \quad \text{與} \quad \mathbf{c}=\frac{1}{Cv}\sum_{i=1}^{N}(\alpha_{1i}-\alpha_{2i})|\mathbf{x}_i|. \tag{10}$$

根據 Karush-Kuhn-Tucker（KKT）最佳化條件，偏移量參數 b 與 d 的計算公式為：

$$b=\frac{-1}{y_i+y_j}\left(y_i\langle \mathbf{w}\cdot\mathbf{x}_i\rangle + y_j\langle \mathbf{w}\cdot\mathbf{x}_j\rangle + \langle \mathbf{c}\cdot|\mathbf{x}_i|\rangle - \langle \mathbf{c}\cdot|\mathbf{x}_j|\rangle - 2\right) \tag{11}$$

$$d=\frac{-1}{2}\left(y_i\langle \mathbf{w}\cdot\mathbf{x}_i\rangle - y_j\langle \mathbf{w}\cdot\mathbf{x}_j\rangle + \langle \mathbf{c}\cdot|\mathbf{x}_i|\rangle + \langle \mathbf{c}\cdot|\mathbf{x}_j|\rangle - 2I_w\right) \tag{12}$$

當 i, j 滿足 $\alpha_{1i}\in\left(0,\frac{C\mu_i}{N}\right)$, $\alpha_{2j}\in\left(0,\frac{C\mu_j}{N}\right)$ 與 $y_i\cdot y_j=1$ 時。求解出模糊超平面的參數（$\mathbf{w}, \mathbf{c}, b, d$）後，模糊超平面可以使用底下的歸屬函數來定義

$$\mu_{Y_i^*}(y)=1-\frac{\left|y-\left(\sum_{k=1}^{N} y_k(\alpha_{1k}+\alpha_{2k})\langle \mathbf{x}_i\cdot\mathbf{x}_k\rangle + b\right)\right|}{\left(\frac{1}{Cv}\sum_{k=1}^{N}(\alpha_{1k}-\alpha_{2k})\langle |\mathbf{x}_i|\cdot|\mathbf{x}_k|\rangle\right)+d}. \tag{13}$$

求解出模糊超平面 $Y=\langle \mathbf{W}\cdot\mathbf{x}\rangle+\mathbf{B}$ 後，對任何輸入 $\mathbf{x}_i$，$Y_i=\langle \mathbf{W}\cdot\mathbf{x}_i\rangle+\mathbf{B}$ 是一個對稱三角形模糊數，其中心為 $\langle \mathbf{w}\cdot\mathbf{x}\rangle+b$ 且寬度為 $\langle \mathbf{c}\cdot|\mathbf{x}|\rangle+d$。而模糊零 Θ 是一個對稱三角形模糊數，其中心為 0 且寬度為 O_w。對一個新進測試樣本點 **x**，我們必須評估他是在模糊超平面哪一邊，亦即，我們必須定義二個三角形模糊數字比較大小的方式，對於任意二個對稱三角形模糊數字 $A=(m_A,c_A)$ 與 $B=(m_B,c_B)$，模糊數 A

大於 B 的模糊程度（亦即 A 位在 B 右邊的模糊程度），是由下列模糊歸屬函數定義

$$R_{\geq B}(A)=R(A,B)=\begin{cases}1 & \text{if } \alpha>0 \text{ and } \beta>0\\ 0 & \text{if } \alpha<0 \text{ and } \beta<0\\ 0.5\left(1+\dfrac{\alpha+\beta}{\max(|\alpha|,|\beta|)}\right) & \text{o.w.}\end{cases}, \quad (14)$$

其中 $\alpha=(m_A+c_A)-(m_B+c_B)$ 與 $\beta=(m_A-c_A)-(m_B-c_B)$。注意，當 $m_A=m_B$ 時，$R_{\geq B}(A)=0.5$，當 $m_A<m_B$ 時，$R_{\geq B}(A)<0.5$，反之，當 $m_A>m_B$ 時，$R_{\geq B}(A)>0.5$。因此，本研究中所提出的模糊支持向量機，其模糊決策函數為

$$f(\mathbf{x})=R_{\geq\Theta}\left(\langle\mathbf{W}\cdot\mathbf{x}\rangle+\mathbf{B}\right)=R\left(\langle\mathbf{W}\cdot\mathbf{x}\rangle+\mathbf{B},\Theta\right) \quad (15)$$

此決策函數傳回樣本點 **x** 屬於正類別的程度，其值在 0 與 1 之間，其值越大，代表樣本點 **x** 屬於正類別的程度值越大。如今，我們使用一個模糊的邊界來分割正類別與負類別，這樣更能解決現實世界中資料不精確的問題。基於模糊支持向量機優異的性質，本研究將應用他到股價趨勢預測的分類問題上。要將此模糊超平面延伸到非線性分割的方式十分簡單，使用核心函數（kernel function）式學習的概念，我們僅需將資料向量透過一個非線性轉換 Φ 映射至高維度的特徵空間，我們在高維度特徵空間求得的模糊線性超平面，在原始空間則為模糊非線性超平面，因為支持向量機演算法的特點，所有對於資料的計算都是以內積（inner product）的方式計算，因此，我們僅需要定義核心函數 $k(\mathbf{x},\mathbf{y})=\Phi(\mathbf{x})\cdot\Phi(\mathbf{y})$，而不需定義 Φ 的函數形式。將公式(9)與(13)中的 $\langle\mathbf{x}_i\cdot\mathbf{x}_j\rangle$ 與 $\langle|\mathbf{x}_i|\cdot|\mathbf{x}_j|\rangle$ 分別用 $k(\mathbf{x}_i,\mathbf{x}_j)$ 與 $k(|\mathbf{x}_i|,|\mathbf{x}_j|)$ 取代，我們即可得到非線性的模糊超平面。

肆、實驗結果

本研究蒐集鉅亨網[3]上的股票新聞，其中包含食品類公司、半導體類公司以及電腦周邊類公司的新聞文章，資料蒐集時間為西元 2014 年 9 月至 2015 年 2 月，各類股新聞文章數量的資料統計摘要顯示於表 5。

3 http://www.cnyes.com

表 5：各類股新聞資料的統計摘要

產業別	個股公司名稱 & 股票代號（新聞篇數）	新聞總篇數
食品類股	統一 1216（19）、南僑 1702（7）、味全 1201（45）	71
半導體類股	日月光 2311（12）、台積電 2330（58）、光罩 2338（2）、宇瞻 8271（3）、旺宏 2337（5）、矽品 2325（12）、南茂 8150（8）、盛群 6202（9）、創見 2451（4）、華亞科 3474（22）、華泰 2329（1）、聯發科 2454（42）、聯電 2303（11）	189
電腦周邊類股	F-鎧勝 5264（21）、仁寶 2324（17）、光寶科 2301（14）、宏碁 2353（45）、佳世達 2352（5）、和碩 4938（25）、研華 2395（4）、英業達 2356（6）、神達 3706（8）、華碩 2357（31）、群光 2385（16）、精英 2331（2）、廣達 2382（4）、緯創 3231（3）	201

本研究首先利用 jieba 斷詞系統[4]將新聞文章斷詞，以供接續的文字探勘技術使用。本研究利用情緒分析及 LDA 來獲得文章隱含內容，我們首先會對新聞文章執行情緒分析，透過 C-LIWC 中文詞典來獲得文章所攜帶的情緒詞類別（如正面情緒詞、負面情緒詞、情感歷程詞、焦慮詞、生氣詞、悲傷詞及認知歷程詞等），接著，本研究使用 JGibbLDA 獲取新聞文章的主題向量，本研究設定主題的數目為 30 個，並且使用粒子群演算法進行特徵挑選，以找出最能影響股價漲跌的隱含主題。所有的資料都進行正規化處理，並將資料值的範圍局限於-1 至 1 中，最終，我們將使用傳統的支持向量機與模糊支持向量機來預測在新聞文章發布後，該篇新聞所涉及的公司，在未來一日與三日的股票價格的漲跌趨勢（相對於新聞發布的時間點的股價），以探討股價預測的反應時間為長期較優異，或是短期較為適當。未來一日與三日股價漲跌趨勢的定義，與新聞發布的時間點有關，本研究是根據新聞是否在交易時段內發布，來定義未來一日與三日股價漲跌趨勢，詳細定義方式如表 6 所示。

表 6：未來一日與三日股價漲跌趨勢的定義

新聞發布時間	未來一日股價漲跌趨勢	未來三日股價漲跌趨勢
在交易時段外	下一個交易時段的開盤價與收盤價之高低比較	下一個交易時段的開盤價與其後第二個交易時段的收盤價之高低比較

[4] https://github.com/fxsjy/jieba

在交易時段內	新聞發布時間點的股價與當前交易時段的收盤價之高低比較	新聞發布時間點的股價與其後第二個交易時段的收盤價之高低比較

註：交易時段為周一至五的 09:00-13:30

舉例來說，如果新聞是在周一早上六點發布（在交易時段外），則未來一日股價漲跌趨勢定義為周一上午 09:00 至下午 13:30 的開盤價與收盤價之高低比較，未來三日股價漲跌趨勢則定義為周一上午 09:00 的開盤價與周三下午 13:30 的收盤價之高低比較；如果新聞是在周二上午十一點發布（在交易時段內），則未來一日股價漲跌趨勢定義為周二上午 11:00 的股價至下午 13:30 的收盤價之高低比較，未來三日股價漲跌趨勢則定義為周二上午 11:00 的股價至週四下午 13:30 的收盤價之高低比較。表 7 顯示本研究所蒐集的各個產業類股在未來一日與三日漲跌幅度的基本統計數據，包含上漲比率、平均數、標準差、第 25 個百分位數（Q1）、第 75 個百分位數（Q3）、最大值與最小值。

表 7：本研究所蒐集的各個產業類股在未來一日與三日漲跌幅度的基本統計數據

產業別	反應時間	上漲比率	平均數	標準差	Q1	Q3	最大值	最小值
食品類股	一日	42.22%	-0.009529	0.039371	-0.03218	0.010661	0.069915	-0.06974
	三日	42.22%	-0.010119	0.057031	-0.02982	0.017682	0.098485	-0.18899
半導體類股	一日	41.79%	-0.002783	0.022457	-0.01493	0.007813	0.068796	-0.06967
	三日	48.68%	0.000404	0.038938	-0.02033	0.017986	0.110594	-0.11526
電腦周邊類股	一日	39.15%	-0.001094	0.024464	-0.01225	0.009662	0.069252	-0.07437
	三日	47.62%	-0.001722	0.034584	-0.01923	0.015038	0.119929	-0.08413

我們使用 RBF（radial basis function）函數 $k(\mathbf{x},\mathbf{y}) = \exp(-q\|\mathbf{x}-\mathbf{y}\|^2)$ 作為我們的核心函數，支持向量機的預測正確率與模型參數（懲罰參數 C 與核心函數參數 q）是有關的，因此，使用相同的模型參數來比較傳統支持向量機與本研究所提出的模糊支持向量機是不公平的，所以本研究將使用方格式搜尋（grid-based searching）與十折交叉驗證評估（ten-folds cross validation），分別對傳統支持向量機與模糊支持向量機挑選最佳的模型參數，懲罰參數 C 的選取範圍為 C=1, 10, 10^2, …, 10^8，核心函數參數 q 的選取範圍為 q=2^3, 2^2, 2, 1, 2^{-1}, …, 2^{-10}，而模糊支持向量機中的參數 v, I_w 與 O_w 選取範圍為 0.1, 0.2,…,0.7，此外，為了簡化參數選取過程，我們設定 I_w=O_w。基於新聞事件的發生是完全隨機的，而且不受時間影響的假設，Nizer 與 Nievola（2012）和 Hagenau 等（2013）等學者的研究，皆是將資料隨機抽取（不論時間順序）其中一部分當作訓練資料，剩下的部分當作測試

資料。而 Groth 與 Muntermann（2011）、Ranco 等（2015）、Weng 等（2017）等學者的研究，則是採用 N 折交叉驗證來評估預測的正確率。為了讓本研究能夠更公平地與上述研究做比較，因此本研究採取十折交叉驗證來評估預測正確率。

股價趨勢預測是一個明顯的模糊分類問題，股價漲+5%的資料樣本，它屬於上漲類別的歸屬程度，比漲 0.05%的資料樣本還要更高，而漲 0.0001%的資料樣本則應位於上漲與下跌類別之間的模糊邊界地帶，我們使用下列的 S 型歸屬函數對每一個資料樣本指派一個模糊歸屬程度

$$\mu(\Delta p_i, a, b) = \begin{cases} 0 & |\Delta p_i| \le a \\ 2\left(\dfrac{|\Delta p_i| - a}{b-a}\right)^2 & a \le |\Delta p_i| \le \dfrac{a+b}{2} \\ 1 - 2\left(\dfrac{|\Delta p_i| - b}{b-a}\right)^2 & \dfrac{a+b}{2} \le |\Delta p_i| \le b \\ 1 & |\Delta p_i| \ge b \end{cases} \tag{16}$$

其中$\triangle p_i$ 是對應第 i 筆資料樣本的公司之股價變動量，$a = 0$ 與 $b=median(|\triangle p_i|)$ 表示$|\triangle p_i|$, i=1,…,N 的中位數。根據公式(16)，股票價格變化量越大，則它屬於上漲（或下跌）類別的歸屬程度就越大，在訓練分類器的階段，這些資料樣本扮演的腳色就越重要，反之，股票價格變化量越小，則它屬於上漲（或下跌）類別的歸屬程度就越小，這些資料樣本較有可能為雜訊，在訓練分類器階段，這些資料樣本的預測誤差便可以適度地忽略。

在第一個實驗中，我們使用傳統支持向量機來分別比較使用關鍵字、全部 30 個主題模型以及透過粒子群演算法挑選具有鑑別力的主題模型作為特徵，對於食品類股、半導體類股、與電腦周邊類股的預測準確率，實驗結果如表 8 所示。實驗結果顯示使用主題模型的預測正確率，是優於以關鍵字為基礎的正確率，原因在於關鍵字屬於較低階的特徵，例如，關鍵字「上揚」可能沒有辦法鑑別股價的漲跌趨勢，因為有可能是成本上揚或是獲利上揚，反之，主題模型屬於較高階的特徵，所謂的主題（topic），就是指能夠闡述高階語義概念的關鍵字集合，例如，「成本」與「上揚」結合在一起的主題，對於股價的漲跌就有鑑別能力了，因為他揭露了成本上揚將導致獲利縮減的資訊。使用關鍵字的效果較差的另一個原因，是使用傳統關鍵字為基礎的文章向量表示法，文章向量的維度會非常高（通常是數千到上萬個），而且資料會非常稀疏，因此會受到高維度的詛咒（curse of high dimensionality）影響其預測正確率，因為在超高維度的向量空間中，距離是沒有意義的。反之，使用主題模型為基礎的文章向量表示法，文章向

量的維度可以大幅縮減（只有 30 個維度），因此不會受到高維度的詛咒的影響。此外，實驗結果顯示透過粒子群挑選主題特徵的預測正確率會高於使用全部 30 個主題，這是因為 LDA 是採取非監督式的策略，從大型語料庫中生成隱藏的主題，這些主題並不一定能夠用來鑑別股價的漲跌趨勢，而透過粒子群演算法來篩選能真正鑑別股價漲跌趨勢的主題，這樣的預測準確率能夠更高。而且透過粒子群演算法挑選重要的主題特徵後，文章向量的維度更為縮減，更不會受到高維度詛咒的影響。

表 8：傳統支持向量機使用關鍵字、全部 30 個主題、粒子群演算法挑選的主題作為特徵的預測正確率

	食品類股		半導體類股		電腦周邊類股	
	1 日	3 日	1 日	3 日	1 日	3 日
關鍵字	59.967721	61.862146	52.942134	53.677489	55.769234	54.383549
主題（全部 30 個）	62.857143	66.666667	55.463710	56.732026	58.931624	57.526882
主題（粒子群篩選）	69.523810	77.333333	60.161290	61.220044	64.145299	64.928315

在第二個實驗中，我們比較傳統支持向量機與本研究使用的模糊支持向量機，在使用粒子群挑選的主題特徵、情緒分布特徵、與結合主題與情緒特徵後，針對食品類股、半導體類股、與電腦周邊類股，在反應時間設定為一日與三日時的預測正確率，實驗結果如表 9 所示。實驗結果顯示，使用本研究所提出的模糊支持向量機的預測正確率，是優於傳統的支持向量機，而結合情緒與主題模型特徵的預測準確率，是優於單純使用情緒特徵或是單純使用主題模型特徵。更進一步分析，我們發現透過新聞文章預測股價漲跌趨勢，在食品類股能獲得不錯的效能（8 成以上的準確率），推論是因為食品類股的相關新聞多為食安風暴與黑心食品等相關消息，這些新聞會明顯地影響該公司的股票走勢，反之，對於半導體類股與電腦周邊類股，透過新聞文章預測股價趨勢的準確率較低，這是由於在半導體與電腦周邊方面，充斥著假新聞或捕風捉影的消息，例如在公佈公司財報表現不佳的消息之前，董事長率先信心喊話，在媒體上發表公司本身體質是很強健的新聞，或是在 i-phone 8 正式公布之前，發布臆測 i-phone 8 的規格與嶄新功能的新聞，或是臆測哪家公司能夠接到蘋果公司的訂單，因此，在半導體與電腦周邊的股價預測方面，準確率多為 6 成，但是，結合模糊理論的支持向量機，在整合情緒與主題特徵之後，對於半導體類股的股價預測可以到達 7 成，說明模糊理論

很適合處理充滿雜訊資料的現實世界應用問題。此外，對於食品類股與電腦周邊類股，使用主題特徵的預測正確率會優於使用情緒特徵，反之，對於半導體類股，使用情緒特徵的預測正確率會優於使用主題特徵，可見主題特徵與情緒特徵各有其擅長的領域，並沒有孰優孰劣，將兩者混合後可以得到更周詳的文章高階內容資訊。另外，我們也發現一個有趣的性質，對於模糊支持向量機的預測實驗中，在單純使用情緒特徵時，對於第 1 日股價漲跌的預測正確率，會高於對第 3 日股價漲跌的正確率，反之，在單純使用主題模型特徵時，對於第 3 日股價漲跌的預測正確率，會高於對第 1 日股價漲跌的正確率，這意味著情緒特徵是較為短暫的激情，所以對於股價預測的反應時間較為短暫，反之，主題特徵隱含著較為深遠的意涵，所以對於股價預測的反應時間較為長久。最終，結合情緒特徵與主題特徵的預測準確性，優於單純使用情緒特徵或是單純使用主題特徵，代表融合短暫激情的情緒與意義深遠的主題資訊後，能夠獲得更加優異的預測正確率。

表 9：模糊支持向量機與傳統支持向量機對使用主題特徵、情緒特徵與混合特徵對不同類股與反應時間的預測正確率

		食品類股		半導體類股		電腦周邊類股	
		1 日	3 日	1 日	3 日	1 日	3 日
傳統支持向量機	主題特徵	69.523810	77.333333	60.161290	61.220044	64.145299	64.928315
	情緒特徵	66.190476	66.666667	61.108871	61.655773	58.717949	58.494624
	主題+情緒特徵	75.714286	80.666667	69.919355	62.832244	69.139785	69.372760
本研究的模糊支持向量機	主題特徵	80.952381	84.666667	60.483871	61.590414	66.153846	67.150538
	情緒特徵	80.952381	75.333333	65.221774	62.026144	62.051282	61.817204
	主題+情緒特徵	87.857143	86.000000	71.491935	66.753813	69.318208	69.605735

接著，我們對本研究提出的方法與先前的股價預測模型進行比較，Schumaker 與 Chen（2009）使用標準的支持向量機與關鍵字的詞袋（bag of words; BOW）特徵來預測股價漲跌；Li 等（2014）使用標準的支持向量機與情緒特徵來分析新聞中所攜帶的情緒，進而預測股價的漲跌；Day 與 Lee（2016）則是使用深度學習與財經情緒詞典來預測股價的漲跌，本研究實作了 Schumaker 與 Chen（2009）、Li 等（2014）與 Day 與 Lee（2016）學者提出的方法，並且將它們應用在中文財經新聞文字探勘以預測股價漲跌的研究領域，雖然 Schumaker 與 Chen（2009）與 Li 等（2014）的方法並不是嶄新的股價預測模型，但是大多數該領域的研究都是採取傳統支持向量機搭配詞袋模型或情緒特徵的作法，而 Schumaker 與 Chen（2009）與 Li 等（2014）是其中較具代表性的作法。而 Day 與 Lee（2016）則是採用最熱門的深度學習技術來訓練多層感知器作為分類器，因此，

本研究與這些方法進行比較，以展示本研究的預測效能。除此之外，至今為止尚未有研究使用模糊支持向量機來對新聞文章進行文字探勘，以預測股價的漲跌趨勢。因此，我們接下來探討在使用主題/情緒混合特徵的情形下，對於本研究提出的具有最大邊界模糊超平面的支持向量機，與 Lin 與 Wang（2002）和 An 與 Liang（2013）學者所提出的模糊支持向量機，進行股價漲跌趨勢的預測效能的優劣比較，預測正確率顯示在表 10。如表 10 所示，考慮訓練樣本的模糊性，並且搭配高階的主題與情緒特徵，能夠明顯改善預測正確率。Day 與 Lee（2016）的方法雖然是使用最近熱門的深度學習技術，但是由於它採用較為簡單的情緒特徵，並且未考慮新聞文章樣本的模糊性質，而且本實驗的訓練樣本數目較少，不足以訓練深度多層感知器內數量龐大的權重參數，因此 Day 與 Lee（2016）方法的預測正確率並不如本研究優異。此外，本研究提出的模糊支持向量機比之前 Lin 與 Wang（2002）和 An 與 Liang（2013）學者提出的模糊支持向量機有更優異的表現，這是因為 Lin 與 Wang（2002）和 An 與 Liang（2013）的模糊支持向量機只有考慮到樣本點的模糊度，但是他們依舊使用一個明確的超平面來分割漲與跌的類別，也就是說，在 Lin 與 Wang（2002）和 An 與 Liang（2013）的方法中，漲與跌類別之間是存在一個明確的邊界，所以它們無法捕抓到漲與跌之間由於雜訊資料所導致的曖昧與模糊的特性。反之，本研究提出的模糊支持向量機使用一個具有最大邊界的模糊超平面來分割漲與跌的類別，它更能夠有效地處理充滿雜訊的新聞資料，並且捕抓到漲跌之間曖昧不明確的邊界。總結來說，不論使用主題特徵、情緒特徵或是混合特徵（反應時間 1 日或 3 日皆是），本研究所提出的具有最大邊界的模糊支持向量機皆能獲得比傳統支持向量機更優異的表現，這驗證了我們之前的假設: 模糊理論更適合於趨勢分類，因為在漲與跌的類別之間，並非存在一條明確的邊界，而是一個模糊的灰色界線，本研究結合模糊理論更能夠處理現實世界中充斥的雜訊與曖昧不明的新聞訊息。

表 10：本研究提出的模糊支持向量機與傳統模糊支持向量機使用主題 / 情緒混合特徵對不同類股與反應時間的預測正確率

	方法	特徵	食品類股		半導體類股		電腦周邊類股	
			1 日	3 日	1 日	3 日	1 日	3 日
未考慮樣本的模糊性	Schumaker & Chen (2009)	詞袋	59.9677	61.8621	52.9421	53.6775	55.7692	54.3835
	Li et al. (2014)	情緒特徵	66.1904	66.6667	61.1089	61.6558	58.7179,	58.4946
	Day & Lee (2016)	情緒特徵	72.8571	62.6667	55.4839	52.2440	57.9487	59.5161

考慮樣本的模糊性	Lin & Wang (2002)	主題+情緒特徵	83.8095	84.6667	70.0582	62.8322	69.1542	69.4256
	An & Liang (2013)	主題+情緒特徵	80.4761	80.9524	69.9194	63.7255	69.1398	69.5161
	本研究的方法	主題+情緒特徵	87.8571	86.0000	71.4919	66.7538	69.3182	69.6057

伍、結論

本研究提出使用模糊支持向量機進行從新聞文章預測公司股價的漲跌趨勢，並且結合新聞文章的情緒資訊與主題模型。為了探討本研究提出的方法在不同族群類股中的預測能力，本研究聚焦於食品類股、半導體類股與電腦周邊類股，本研究所提出的方法在食品類股的預測正確率最高為 87%，半導體類股的正確率最高為 71%，電腦周邊類股的預測正確率最高為 69%，相較於傳統支持向量機使用關鍵字為基礎的預測正確率僅有 5 成多（近似於隨機猜測），本研究所提出的方法明顯優於傳統的預測模型。相較於傳統關鍵字表示法僅能夠提供低階的文字資訊，情緒特徵與主題模型能夠提供更高階的文章內容特徵，而且能大幅縮短文章向量的維度，避免高維度的詛咒，因此能獲得更優異的預測準確率，此外，本研究發現情緒特徵通常攜帶較為短暫的激情，所以對於股價預測的反應時間較為短暫，主題特徵隱含著較為深遠的意涵，所以對於股價預測的反應時間較為長久。而結合情緒特徵與主題特徵，代表融合短暫激情的情緒與意義深遠的主題資訊，所以能夠獲得更加優異的預測正確率。此外，模糊理論非常適合於趨勢分類，因為在趨勢上漲與下跌的類別之間，並非存在一條明確的邊界，而是一個模糊的灰色界線，上漲 1%與上漲 0.01%雖然都是屬於上漲的類別，但是屬於上漲類別的程度明顯地不同，本研究證實結合模糊理論更能夠處理現實世界中充斥的雜訊資訊與曖昧不明的新聞文章訊息。

誌謝

本文接受行政院科技部專題研究計畫（MOST 106-2221-E-151-049-）之補助研究經費，順利完成此篇著作之研究工作，謹此致謝。

參考文獻

黃金蘭、林以正、謝亦泰、程威銓（2012），『中文版「語文探索與字詞計算」詞典之建立』，*中華心理學刊*，第 54 卷，第 2 期（2012/06/01），頁 185-201。

An, W. and Liang, M. (2013), 'Fuzzy support vector machine based on within-class

scatter for classification problems with outliers or noises', *Neurocomputing*, Vol. 110, pp. 101-110.

Bahrepour, M., Akbarzadeh-T, M.-R., Yaghoobi, M. and Naghibi-S, M.-B. (2011), 'An adaptive ordered fuzzy time series with application to FOREX', *Expert Systems with Applications*, Vol. 38, No.1, pp. 475-485.

Blei, D.M., Ng, A.Y. and Jordan, M.I. (2003), 'Latent dirichlet allocation', *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022.

Bollen, J., Mao, H. and Zeng, X. (2011), 'Twitter mood predicts the stock market', *Journal Computational Science*, Vol. 2, No. 1, pp. 1-8.

Day, M.-Y. and Lee, C.-C. (2016), 'Deep learning for financial sentiment analysis on finance news providers', *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 1127-1134.

Fama, E.F. (1991), 'Efficient capital markets: II', *The Journal of Finance*, Vol. 46, No. 5, pp. 1575-1617.

Fung, G.P.C., Yu, J.X. and Lam, W. (2002), 'News sensitive stock trend prediction', In: Chen MS., Yu P.S., Liu B. (eds) *Advances in Knowledge Discovery and Data Mining. (PAKDD 2002,),* pp. 481-493.

Groth, S.S. and Muntermann, J. (2011), 'An intraday market risk management approach based on textual analysis,' *Decision Support Systems*, Vol. 50, No. 4, pp. 680-691.

Hagenau, M., Liebmann, M. and Neumann, D. (2013), 'Automated news reading: Stock price prediction based on financial news using context-capturing features', *Decision Support Systems*, Vol. 55, No.3, pp. 685-697.

Hao, P.-Y. (2016), 'Support vector classification with fuzzy hyperplane', *Journal of Intelligent & Fuzzy Systems*, Vol. 30, No. 3, pp. 1431-1443.

Kang, D. and Park, Y. (2014), 'Review-based measurement of customer satisfaction in mobile service: sentiment analysis and VIKOR approach', *Expert Systems with Applications*, Vol. 41, No. 4, pp. 1041-1050.

Kennedy, J. and Eberhart, R. (1995), 'Particle swarm optimization', *Proceedings., IEEE International Conference on Neural Networks, Australia*, pp. 1942-1948.

Kennedy, J. and Eberhart, R.C. (1997), 'A discrete binary version of the particle swarm algorithm', *IEEE International Conference on Systems, Man, and Cybernetics*, Orlando USA October 12, Vol.5, pp. 4104-4108.

LeBaron, B., Arthur, W.B. and Palmer, R. (1999), 'Time series properties of an artificial stock market,' *Journal of Economic Dynamics and Control*, Vol. 23, No. 9-10, pp. 1487-1516.

Leigh, W., Purvis, R. and Ragusa, J.M. (2002), 'Forecasting the NYSE composite index with technical analysis, pattern recognizer, neural network, and genetic algorithm: a case study in romantic decision support', *Decision Support Systems*, Vol. 32, No. 4, pp. 361-377.

Li, F. (2010), 'Textual analysis of corporate disclosures: A survey of the literature', *Journal of Accounting Literature*, Vol. 29, 2010, pp. 143-165.

Li, Y.M. and Li, T.-Y. (2013), 'Deriving market intelligence from microblogs', *Decision Support Systems*, Vol. 55, pp. 206-217.

Li, X., Xie, H., Chen, L., Wang, J. and Deng, X. (2014), 'News impact on stock price return via sentiment analysis', *Knowledge-Based Systems*, Vol. 69, pp. 14-23.

Lin, C.-F. and Wang, S.-D. (2002), 'Fuzzy support vector machines', *IEEE Transactions on Neural Networks*, Vol. 13, No. 2, pp. 464-471.

Liu, B. and Zhang, L. (2012), 'A survey of opinion mining and sentiment analysis', In *Mining Text Data*, pp. 415-463, Springer.

Lu, C.J., Lee, T.S. and Chiu, C.C. (2009), 'Financial time series forecasting using independent component analysis and support vector regression', *Decision Support Systems,* Vol. 47, No.2, pp. 115-125.

Mabu, S., Hirasawa, K., Obayashi, M. and Kuremoto, T. (2013), 'Enhanced decision making mechanism of rule-based genetic network programming for creating stock trading signals', *Expert Systems with Applications*, Vol. 40, No.16, pp. 6311-6320.

Nassirtoussi, A.K., Aghabozorgi, S., Wah, T.Y. and Ngo, D.C.L. (2014), 'Text mining for market prediction: a systematic review,' *Expert Systems with Applications*, Vol. 41, No. 16, pp. 7653-7670.

Nizer, P.S.M. and Nievola, J.C. (2012), 'Predicting published news effect in the Brazilian stock market', *Expert Systems with Applications*, Vol. 39, No. 12, pp. 10674-10680.

Premanode, B. and Toumazou, C. (2013), 'Improving prediction of exchange rates using differential EMD', *Expert Systems with Applications*, Vol. 40, No.1, pp. 377-384.

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M. and Mozetič, I. (2015), 'The effects of Twitter sentiment on stock price returns,' *PLoS ONE*, Vol. 10, No. 9, e0138441, doi:10.1371/journal.pone.0138441.

Ravi, K. and Ravi, V. (2015), 'A survey on opinion mining and sentiment analysis: Tasks, approaches and applications', *Knowledge-Based System*, Vol. 89, pp. 14-46.

Rui, H., Liu, Y. and Whinston, A. (2013), 'Whose and what chatter matters? The effect of tweets on movie sales', *Decision Support Systems*, Vol. 55, No. 4, pp. 863-870.

Ryan, P. and Taffler, R.J. (2004), 'Are economically significant stock returns and trading volumes driven by firm specific news releases?', *Journal of Business Finance & Accounting*, Vol. 31, No.1-2, pp. 49-82.

Schumaker, R.P. and Chen, H. (2009), 'Textual analysis of stock market prediction using breaking financial news: the AZFin text system', *ACM Transactions on Information Systems*, Vol. 27, No. 2, [a12]. DOI: 10.1145/1462198.1462204.

Schumaker, R.P., Zhang, Y., Huang, C.-N. and Chen, H. (2012), 'Evaluating sentiment in financial news articles', *Decision Support Systems*, Vol. 53, No.3, pp. 458-464.

Sermpinis, G., Laws, J., Karathanasopoulos, A. and Dunis, C.L. (2012), 'Forecasting and trading the EUR/USD exchange rate with gene expression and psi sigma neural networks', *Expert Systems with Applications*, Vol. 39, No. 10, pp. 8865-8877.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H. and Deng, X. (2013), 'Exploiting topic based twitter sentiment for stock prediction', *In Proceedings of the 51st annual meeting of the association for computational linguistics*, Vol. 2, short papers, pp. 24-29, *The Association for Computer Linguistics*.

Shi, Y. and Eberhart, R. (1998), 'A modified particle swarm optimizer', *Proc. of the IEEE Congress on Evolutionary Computation. IEEE Service Center*, Anchorage USA, May, pp. 69-73.

Tetlock, P.C., Saar-Tsechansky, M. and Macskassy, S. (2008), 'More than words: Quantifying language to measure firms' fundamentals', *Journal of Finance*, Vol. 63, No. 3, pp. 1437-1467.

Tetlock, P.C. (2011), 'All the news that's fit to reprint: Do investors react to stale information?', *The Review of Financial Studies*, Vol. 24, No. 5, pp. 1481-1512.

Vapnik V. (1995), *The Nature of Statistical Learning Theory*, Springer-Verlag, New York.

Vu, T.T., Chang, S., Ha, Q.T. and Collier, N. (2012), 'An experiment in integrating sentiment features for tech stock prediction in Twitter', In: *Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data,* pp. 23-38.

Walczak, S. (2001), 'An empirical analysis of data requirements for financial forecasting with neural networks', *Journal of Management Information Systems*, Vol. 17, No. 4, pp. 203-222.

Weng, B., Ahmed, M.A. and Megahed, F.M. (2017), 'Stock market one-day ahead movement prediction using disparate data sources,' *Expert Systems With*

Applications, doi: 10.1016/j.eswa.2017.02.041 Vol.79, pp. 153-163.

Wu, D., Fung G.P.C., Yu, J.X. and Pan, Q. (2009), 'Stock prediction: An event-driven approach based on bursty keywords', *Frontiers Computer Science in China*, Vol. 3, No. 2, pp. 145-157.

Yu, L.-C., Wu, J.-L., Chang, P.-C. and Chu, H.-S. (2013), 'Using a contextual entropy model to expand emotion words and their intensity for the sentiment classification of stock market news', *Knowledge-Based Systems*, Vol. 41, pp. 89-97.

Yu, H., Nartea, G.V., Gan, C. and Yao, L.J. (2013), 'Predictive ability and profitability of simple technical trading rules: Recent evidence from Southeast Asian stock markets', *International Review of Economics and Finance*, Vol. 25, pp. 356-371.

Zadeh, L. A. (1965), 'Fuzzy sets', *Information and Control*, Vol. 8, pp. 338-353.