新聞情緒及技術指標於股票漲跌幅排名預測及動態投資組合最佳化之研究

林政憲
資訊工程學系
國立中央大學
jslin931@gmail.com

張嘉惠
資訊工程學系
國立中央大學
chia@csie.ncu.edu.tw

***摘要*—科技金融在人工智慧應中是一個熱門的主題，過去雖已有許多研究使用股票歷史數據及技術指標進行個股漲或跌的預測，但是要如何將預測結果結合投資組合的配置仍然是個問題。因此本研究中，我們將財經新聞與技術指標納入為漲跌預測中，以獲得更好的預測效果。我們提出以預測隔週/隔月的股票漲跌幅排名問題，以減少頻繁交易，在使用技術指標基準下， 比較有無加入新聞情緒時,對於預測股票漲跌幅排名的影響。實驗結果顯示，加入新聞情緒其函數的損失較沒加入新聞情緒時低。最後我們選擇漲跌排名預測前 K 名的股票，形成動態資組合配置，其投資報酬率比台灣 50ETF 高2.5~3.6倍；尤其是模型在加入新聞情緒時，投資報酬率可再提高 4~8%。**

Keywords—科技金融, 投資組合配置, 股票漲跌排名預測

I. 緒論

股票是金融市場中最廣為投資人所熟悉的商品，大部份投資者都希望從中獲得利益。但影響股價的面向非常的多，包含政治面向（貿易戰、政府政策、國內治安）、經濟面向（國際環境、景氣循環、貨幣匯率、利率）、公司營運面向（公司經營狀況、是否有新產品推出、增減資）、市場面向（投資者的預測心理、人為操縱）等等。影響面向太多讓股票市場變得難以預測。

有鑑於此，市場上已有證券業者，開始將金融科技導入 AI 人工智慧技術，利用大數據分析建立預測模型，應用技術面、財務面、籌碼面及非量化網路聲量、輿情等選股策略，為投資人提供智慧化金融服務，吸引投資人的下單意願。例如美國 ETF Managers Group 公司於 2017 年底，推出一檔以 AI 技術作為進出場依據的 ETF。蒐集美國股票市場上 6000 多檔股票公開披露的訊息進行分析，並使用深度學習建立價格預測模型。其模型不僅考量公司管理階層的能力，也蒐集社群媒體中的公開評論及新聞內容，將市場輿論進行整合分析，以評估股票是否具有投資價值。此檔 ETF 在首年即取得領先大盤 5%的績效。

雖然社群媒體中輿情分析有助於股價趨勢預測，然而台灣市場經濟規模較小，社群媒體討論量不大，難以每一天都有每一家上市公司的訊息，因此對於單一股票的預測幫助不大。再者，個股預測如何結合投資策略，提高投資報酬率，才是投資者最後關注的重點。在本篇研究中，我們提出以每週股票漲跌排名為預測之任務，採用個股在每日新聞的情緒，藉以抵消個股資訊不足之影響。並以每週漲跌排名前十名之個股做為投資標的。我們收集台灣股市資訊網及玩股網上，2017 至 2019 間 150 家上市公司每日相關新聞，預測每週股票漲跌排名。在此投資策略下，其投資報酬率可高於台灣 50 ETF 之表現。

II. 相關研究

有關於股票市場預測的研究非常多元，有許多不同切入的角度。我們比較關注個股價格或股價漲跌趨勢預測、新聞/新聞情緒與股價的關聯性、以及結合個股預測於投資組合的調整上面。

價格預測是機器學習在 FinTech 上的主要應用之一，如股票市場[13]、比特幣[8]、外匯市場的匯率[7]等。常用的模型包括研究中會提出自己的深度學習模型，給予模型相同的輸入，並以 MSE、RMSE 或標準差等，比較相同輸入在其他不同的模型間的預測能力。除此之外，漲跌趨勢預測也是常見的研究，[6]使用自然語言處理的技術，將新聞或是社群網路中的評論，轉換成結構化的資料，作為是模型的輸入，查看市場輿論對於股票漲跌的相關性。[10]則將社群網路中的評論進行情緒分析，將情緒當作是預測股價的一項指標。[12]則採用端對端的模型，進行 88 檔美股股價漲跌之預測。

投資組合的目的在分散風險，因此考量的不僅是個股的投資報酬率，也會著重在投資風險。[1]從台股 20 檔指數股票型基金(ETF)中挑選六檔相關係數較低的 ETF 做為投資組合，並利用投資組合權重最佳化的方法，如 Mean-Variance Optimizer （MVO）、Monte Carlo Simulation 等，將投資報酬最大化。[9]則從美國 11 個不同產業中各選取一檔 ETF 做為投資組合，比較不同權重最佳化方法的年化報酬率及夏普值。[11]蒐集美國 S&P500 成份股，應用 LR、SVR[3]進行股價預測，並挑選四檔預測最準的股票做為投資組合，再進行權重調整 ，比較累積收益及夏普值。

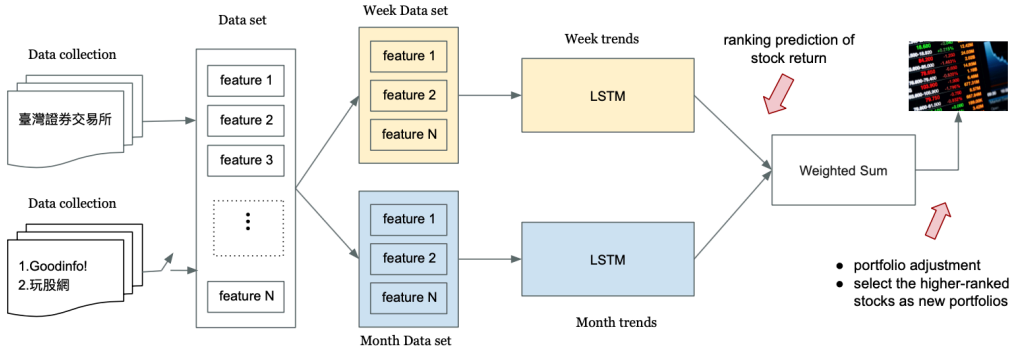

圖 1. 投資組合報酬預測流程

由於台灣社群媒體資訊有限，難以每日都有個股的消息，加入社群評論對個股預測效能影響有限，因此本篇論文以預測股票漲跌幅排名做為問題定義，應用每日個股新聞情緒分析結果，使用各新聞情緒的提及比例，以降低個股新聞資訊不足的問題。同時我們以漲幅排名前十的股票進行投資。

III. 股票漲跌幅排名預測及投資組合策略

我們以台灣 50 及中型 100 指數之成份股作為我們的觀察股，探討深度學習的方法是否能夠使用於投資標的選取。配合漲幅排名前十名的投資策略，測試在不同時空條件下，系統是否能動態的找到適合的投資組合，並能夠在投資持有的區間內，得到好的投資報酬率。

在投資組合報酬率預測中，為了要瞭解股票在短中長期的移動趨勢，我們分別訓練兩個模型，一個以週為單位，用來預測短期股價的移動趨勢；另一個以月為單位，用來預測中長期股價的移動趨勢。在操作時，參考兩個模型的預測趨勢，給予不同的權重，加權作為選取股票排名分數[5]，作為交易訊號。每週計算完分數後，即依照分數的排名變更原本手中的持股，選擇排前面的股票進行投資。透過預測股票的移動趨勢，找出強勢股票，消除掉外在大環境的干擾。其系統流程如圖 1。

在短期股價漲跌幅排名預測中，我們考慮公司短期股價的影響因素，使用新聞情緒/每日新聞情緒比例及正規化處理後的開盤價、最高價、最低價、收盤價及成交量(以下稱 OHLCV 指標)作為模型的輸入特徵，預測目標則為每週的排名分數；在中長期股價移動趨勢的預測中，我們考慮公司中長期的營運發展，與公司股價是否過高，使用月營收、乖離率及 RSV 等指標作為模型的輸入特徵，預測目標則為每月的排名分數。

A. 技術面特徵

為了進行技術面分析，我們從台灣證券交易所網上收集，2005 至 2019 間 150 家上市公司相關歷史資訊。針對週模型、及月模型分別準備技術面特徵如下。

首先，週模型以 OHLCV 指標作為模型的輸入特徵，其正規化的方式如下：。

- 開盤價/最高價/最低價之正規化

$$p^s(\mathrm{t}) = \frac{\tilde{p}^s(t)}{\tilde{p}_c^s(t)} - 1 \tag{1}$$

其中$\tilde{p}^s(t)$表示交易日 t 股票 s 的實際開盤/最高/最低價、$\tilde{p}_c^s(t)$表示交易日 t 的實際收盤價。

- 收盤價之正規化

$$p_c^s(t) = \tilde{p}_c^s(t) - \tilde{p}_c^s(t-1) \tag{2}$$

其中$\tilde{p}_c^s(t)$表示交易日 t 的股票 s 實際收盤價、$\tilde{p}_c^s(t-1)$表交易日 t 前一日的股票 s 實際收盤價。

- 成交量之正規化

$$vol^s(t) = \frac{\widetilde{vol}^s(t)}{\frac{1}{10}\sum_{t-10}^{t-1}\widetilde{vol}^s(t")} \tag{3}$$

其中$\widetilde{vol}^s(t)$表示股票 s 在交易日 t 的實際成交量、$\frac{1}{10}\sum_{t-10}^{t-1}\widetilde{vol}^s(t")$表交易日 t 前 10 日的平均成交量。

其次，月模型使用月營收、乖離率、RSV 指標作為模型的輸入特徵，其資料的準備方式如下：

- 月營收之正規化：

  透過月營收可以判斷公司是否持續賺錢，正規化方式如公式(4)。

$$m^s(t) = \frac{\widetilde{m}_c^s(t)}{\widetilde{m}_c^s(t-1)} \tag{4}$$

其中$\widetilde{m}^s(t)$表示 t 月的實際月營收。

- 乖離率指標

  乖離率指標為當日股票收盤價與移動平均線的距離，可以用來辨斷股票買賣超的程度，其公式表示如(5)。

$$BIAS_n = \frac{\tilde{p}_c^s(t) - MA_n^s}{MA_n^s} \cdot 100 \tag{5}$$

其中$BIAS_n$表示 n 日的乖離率，$\tilde{p}_c^s(t)$表示股票 s 於交易日 t 的收盤價。MA_n^s為股票 s 於 n 日的移動平均價格。

- RSV 指標

  RSV 指標可以用於判斷市場屬於超買/超賣狀態，以了解市場行情，如公式(6)。

$$RSV_n = \frac{\tilde{p}_c^s(t) - \tilde{p}_{min,n}^s}{\tilde{p}_{max,n}^s - \tilde{p}_{min,n}^s} \quad (6)$$

  其中$\tilde{p}_c^s(t)$為交易日 t 股票 s 的收盤價，$\tilde{p}_{min,n}^s$及$\tilde{p}_{max,n}^s$分為最近 n 天內股票 s 最低價及最高價。

B. 市場面特徵

為了進行市場面分析，我們收集台灣股市資訊網及玩股網上，2017 至 2019 間 150 家上市公司每日相關新聞。首先將新聞的標題進行情緒分類，把新聞情緒分類為正向、中立及負向，分別用 1、0 及-1 表示。我們手動標記 2017/01~2017/06 之間，觀察股的新聞情緒約 5000 則，作為 Bert 模型訓練及測試使用。將標記的新聞情緒分成訓練集 3000 則、開發集 1000 則及測試集 1000 則進行訓練，其訓練及測試結果如表格 1。我們使用 Bert[2][4]模型進行中文語意分類。將新聞標題序列作為 Bert 的輸入，經過 Bert 轉換後會變成另一串 Embedding 序列，並對這串序列進行處理，以完成新聞情緒的分類。

表格 1. Bert 模型訓練及測試集之預測效果

	訓練集	測試集
準確準	0.7889	0.7459
F-measure	0.7492	0.7458

剩餘其它尚未標記的新聞標題，則使用 Bert 模型進行自動標記，將所有新聞標題皆轉換成正向、中立、負向的新聞情緒。最後得到正向情緒約 18000 則、中立情緒約 19000 則、負向情緒約 9000 則。

C. 輸出目標：排名分數

由於預測目標為股票的漲跌幅排名，因此我們考量 k 個的交易日中每檔觀察股 s 在特定區間內的漲跌幅$R_k^s(t)$，再依序把這些漲跌幅進行排名(ranking)並做正規化，使其值介於 0~1 之間，其數學表示如公式(8)。

$$A_k^s(t) = \frac{\#\{R_k^i(t) | R_k^s(t) \geq R_k^i(t), 1 \leq i \leq N\} - 1}{N-1} \quad (7)$$

$$R_k^s(t) = \frac{P^s(t+k) - P^s(t)}{P^s(t)} \quad (8)$$

其中 N 為觀察股個數；$R_k^s(t)$為觀察股 s 在 k 個交易日中漲跌幅；# 函數回傳集合個數，其值介於 1~N 之間；因此經過公式(1)的計算，$A_k^s(t)$為股票 s 在 k 個交易日中，正規化後的漲跌幅排名分數，排名分數愈靠近 1 的股票代表漲幅愈大，反之愈靠近 0 的股票則代表漲幅愈小。

排名分數大於 0.5 的股票，表示其股票的漲幅，超越所有觀察股集合的平均排名分數，屬於強勢股其股價上漲機率較大；反之排名分數小於 0.5 的股票，屬於弱勢股其股價上漲機率較小。

IV. 實驗

我們將技術面及市場面的資料拆分成訓練集、開發集及測試集，其區間如表格 2。在週模型中打算預測股價短期的走勢，考量到新聞資料蒐集的時間，使用資料區間為 2017/01~2019/07；在月模型中打算預測股價中長期的走勢，將基本面的指標(如每個月只公布 1 次的月營收)納入考量，因此需要考慮的資料區間較週模型來的長，使用 2005/01~2019/07。

表格 2. 週模型及月模型之資料使用區間

	週模型	月模型
訓練集	2017/01~2018/09	2005/01~2017/06
開發集	2018/10~2018/12	2017/07~2018/12
測試集	2019/01~2019/07	2019/01~2019/07

A. 評估方法

本文採用兩種預測評估方法：RMSE(Root Mean Square Error)以及 nDCG(normalize Discounted cumulative gain)。RMSE，一般稱為均方根誤差。為預測值與實際值間的誤差平方，用來衡量預測值與實際值之間的偏差。nDCG，一般用於衡量搜索引擎算法的指標，逐條針對每個搜尋結果進行評分。評分時分數高的結果表示比分數低的結果好；在搜尋時相關度越高的結果排在前面越好。利用此概念評價我們模型在預測排名前面的股票結果是否精準。

B. 漲跌幅排名預測

我們在漲跌幅排名預測中比較週模型在不同輸入特徵時的預測效果。以 OHLCV 指標作為模型的基準模型(Baseline)，比較加入新聞情緒 (with NS)在測試集的 RMSE，其結果如表格 3。

表格 3. 週模型效能之 RMSE 比較

	Baseline	with NS
週模型	0.2663	0.2648

我們將每檔股票每期的預測分數，依照分數高低切分成不同的分數區間，橫軸為股票排名分數區間，縱軸為觀察股之排名分數累積。其觀察預測分數區間分佈，可以發現大多數的分數皆落在 0.55 的區間，如圖 2。

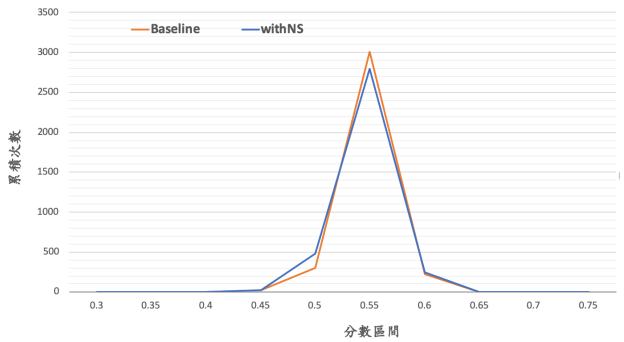

圖 2. 不同模型之分數預測區間

利用 nDCG 計算預測之前 10 及 30 名的股票，是否能夠準確挑出實際排名前 10 名或 30 名的股票，其結果如

表格 4。可以看出在預測排名前 10 名模型中，加入新聞情緒的模型效能最好，最能夠挑出實際排名前 10 名的股票；在預測排名前 30 前名的比較中則是 Baseline 的模型效能最好，最能夠挑出實際排名前 30 名的股票。

表格 4. 不同模型之 nDCG 比較

	nDCG@10	nDCG@30
Baseline	0.3176	0.2914
with NS	0.3645	0.2852

C. 投資組合策略回測比較

在預測觀察股之股票區間的漲跌幅排名後，我們首先比較週模型、月模型，以及混合模型的效能。其中混合模型依經驗法則將股票週、月權重分別設定為 0.55 與 0.45，每週計算排名分數，再選擇排名前面的 10 檔股票形成新的投資組合進行換股操作，回測其當週投資組合報酬率，將每週投資報酬率相乘即得累積報酬率。如圖 3 所示，單純使用週模型效能較月模型累積報酬率更好，混合模型也比個別單純模型效能佳。

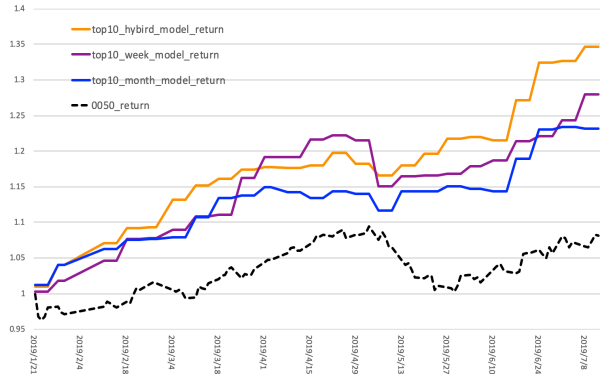

圖 3. 週模型、月模型、混合模型投資組合報酬率比較：紫色、藍色、橘色線分別為（不含新聞情緒之）週模型、月模型及混合模型，選擇排名前 10 名股票的投資組合報酬率;黑色虛線為台灣 50ETF 的投資報酬率。

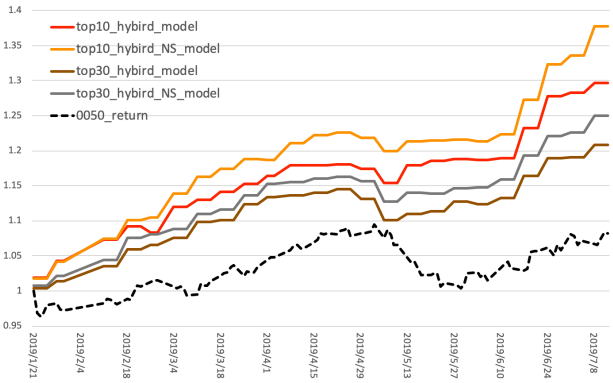

圖 4. 加入新聞情緒，以及選擇不同排名投資組合報酬率比較：橘色線、灰色線分別為加入新聞情緒且選擇排名前 10 及前 30 名股票的投資組合報酬率；紅色及咖啡色為不含新聞情緒之混合模型。

其次，我們比較加入新聞情緒，以及選擇不同排名之投資組合對於投資報酬率的影響。如圖 4 所示，在混合模型中加入新聞情緒的週模型效能比只使用 OHLCV 的週模型效能更好；同時選擇排名前 10 的投資組合報酬率，也比選擇前 30 名股票的效能好。以 7/28 最終投資報酬率結算，選擇排名前十的策略下，無論有無使用新聞情緒，其投資報酬率(37.7%, 29.7%) 比台灣 50ETF (8.2%)均高了許多。即使選擇排名前 30 的投資組合，其投資報酬率(25.0%, 20.9%)，也是大幅領先台灣 50ETF。

V. 結論與建議

在股票市場中只要是在預測效果上有一點的提升，就能更準確的提示投資者在股票買賣上的判斷，進而獲得更多的投資報酬率。實驗結果顯示，應用漲跌幅排名預測機制的投資策略，相較於台灣 50ETF，可大幅提升投資報酬率 2.5~3.6 倍(20.9/8.2≒2.5, 29.7/8.2≒3.6)；使用混合模型的回測投資報酬率效果比單純使用週模型及月模型來的好。而且在加入新聞情緒時，可進一步提高讓投資報酬率再提高約 4~8%(25-20.9≒4.1, 37.7-29.7≒8)。另外投資漲跌幅排名愈前面的股票，獲得的報酬率也愈高。在未來研究上，可以將風險因素考量進來，或以多目標最佳化方式來尋找最佳投資組合。

VI. 參考文獻

[1] Day M.-Y. and Lin J.-T., 2019. Artificial Intelligence for ETF Market Prediction and Portfolio Optimization, IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp. 1026–1033.

[2] Devlin J., Chang M.-W., Lee K. and Toutanova K., 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, Association for Computational Linguistics, pp. 4171-4186.

[3] Drucker, H., Burges, C.J.C., Kaufman, L., Smola, A. and Vapnik, V., 1997. Support vector regression machines. *Advances in Neural Information Processing Systems*, 9:155–161.

[4] GitHub, 2020. BERT retrieved from https://github.com/google-research/bert

[5] HELLSTROM T., 2000. Predicting a Rank Measure for Stock Returns, Theory of Stochastic Processes, Vol.6 (22), no.3-4, pp. 64-83.

[6] Jacobs G., Lefever E. and Hoste V., 2018. Economic Event Detection in Company-Specific News Text, Proceedings of the First Workshop on Economics and Natural Language Processing, Melbourne, Australia, pages 1–10.

[7] Khairalla M., Ning X. and Jallad N.-T., 2017. Hybrid Forecasting Scheme for Financial Time-Series Data using Neural Network and Statistical Methods, International Journal of Advanced Computer Science and Applications, pages 319–327.

[8] McNally S., Roche J. and Caton S., 2018. Predicting the Price of Bitcoin Using Machine Learning, 26th Euromicro International Conference on Parallel, Distributed, and Network-Based Processing, pages 339–343.

[9] Obeidat S., Shapiro D. and Lemay M., 2018. Adaptive Portfolio Asset Allocation Optimization with Deep Learning, International Journal on Advances in Intelligent Systems, page 25-34.

[10] Pagolu V.-S., Challa K.-N.-R., Panda G., and Majhi B., 2017. Sentiment Analysis of Twitter Data for Predicting Stock Market Movements, International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES), page 1-6.

[11] Ta V.-D., Liu C.-M. and Addis D., 2018. Prediction and Portfolio Optimization in Quantitative Trading Using Machine Learning Techniques. Proceedings of the Ninth International Symposium on Information and Communication Technology (SoICT), Pages 98–105

[12] Xu Y. and Cohen S.-B., 2018, July. Stock Movement Prediction from Tweets and Historical Prices, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pages 1970–1979.

[13] Zhang L., Aggarwal C. and Qi G.-J., 2017. Stock Price Prediction via Discovering Multi-Frequency Trading Patterns, KDD Applied Data Science Paper, pages 2141–2149.