Based On Word embedding DA-RNN Model For Stock Prediction

1 INTRODUCTION

股票預測的準確度一直以來都是各界非常關注的議題，從以往由人類來判斷股票交易點，至今日人工智慧技術日益成熟，我們可以讓機器以「人類判斷股價的依據」學習，並得出股價預測的結果。而「人類判斷股價的依據」不外乎為歷史股價，以及財經新聞資訊。

接續上所述，除了歷史股價之外，財經新聞對於股價同樣也有巨大的影響，因此如何正確判斷出各股相關新聞資料就相當重要了。以人工智慧預測股票來說，已經有數不清的研究不斷地在嘗試，眾多股價預測模型除了參考歷史股價之外，當然也同時參考了新聞資訊。

然而，在處理新聞時普遍使用關鍵字提取之方法，例如預測目標為台積電股票，即以台積電為關鍵字提取新聞做為模型的特徵，但如此真的足夠了嗎?其餘未提及關鍵字的財經新聞不會對個股股價造成影響嗎?

摩根士丹利等八大外資看好，新 iPhone 拉貨將推升蘋概股 9 月業績，點名台積電、鴻海、廣達、日月光投控、臻鼎 - KY、可成、大立光、美律、穩懋等九大指標股可望展現旺季力道，點燃台股 10 月多頭氣氛。

沙烏地阿拉伯媒體報導，美國總統川普 9 月 29 日致電沙國國王沙爾曼，討論油市穩定、以及中東與全球局勢發展，希望能確保「維持油市的供應與穩定」及「全球經濟成長」，可能對油價造成壓力。

（圖 1，台積電新聞）

記憶體大廠華邦電 (2344-TW) 今 (3) 日在南科高雄園區舉行 12 吋新廠動土典禮，該廠導入 25 奈米技術，預計 2020 年廠房興建完成，並於 2021 年投產營運，投資額約 3350 億元。華邦電董事長焦佑鈞表示，新廠選擇高雄、根留台灣，是再自然不過的事，為努力在每單位面積的土地上，創造最大價值。

華邦電南科高雄園區新廠今日動土，該廠為華邦電第 2 座 12 吋晶圓廠，佔地 25 公頃，預計 2020 年廠房興建完成，並於 2021 年投產營運，將視市場需求逐步擴增產能。華邦電規劃將新廠打造為智慧生產的 12 吋晶圓廠，並導入自行開發的 25 奈米技術，著眼利基型 DRAM 產品，滿足物聯網、智能系統與工業自動化等需求。

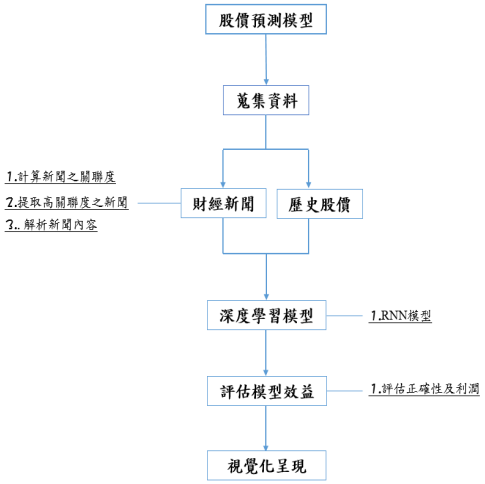
（圖 2，未提及台積電之新聞）

圖 1 及圖 2 源自於鉅亨網之財經新聞。以台積電為例，圖 1 提及台積電，無庸置疑此則新聞對於台積電之股票來說是相當重要且不可忽略的信息。圖 2 之新聞雖未提及台積電，但是其與台積電同樣有晶圓相關之業務，對於台積電來說，同產業或是競爭對手之新聞理應同樣為不可輕忽之新聞，但若是以關鍵字方法提取，將會忽略此則信息。當然使用關鍵字提取或許可以無窮盡的列舉關鍵字，但是無法保證關鍵字之完善，而因此錯失與台積電相關之重要新聞。

考量上述之情況，本研究將以「關聯式提取特徵方法」提取財經新聞，並且將要證明，在資料及其他條件一致的情況下，本研究之「關聯式提取特徵方法」相較於「關鍵字提取特徵方法」以及「僅參考歷史股價」，對股價預測模型將會有更佳準確的結果。

2 METHODS

2.1 流程概述



（圖 3，此次研究之流程圖)

圖 3 即為本研究之流程，說明如下:

1. 蒐集財經新聞及相關資料。以台積電股票(2330)為例。
2. 正確辨識與目標股票高關聯之新聞資料。本計畫預計開發出以 word embedding 之技術尋找與目標股票更為相關之新聞，以該股較詳盡且具有標的性之介紹為中心點，藉此計算出各篇財經新聞與中心點之關聯度，提取高度關聯之新聞資料做為後續處理之用。
3. 本研究同時將會比較以關聯度提取新聞之方法與傳統關鍵字提取方法以及未參考新聞資料方法之差異。
4. 提取之新聞資料轉為向量後連同歷史股價輸入深度學習

模型，預測目標為**收盤價、漲跌趨勢、買賣建議**。

5. 預測效能之評估。評估股價趨勢預測之正確性，以及買賣交易之建議及利潤。
6. 視覺化呈現預測模型所建議的最適交易時機(評估誤差及正確性)。

2.2 資料設定

- ✓ 目標股票:台積電
- ✓ 資料期間:2017/10/01-2018/11/01
- ✓ 訓練資料:2017/10/01-2018/09/30
- ✓ 最終目標:2018/10/01-2018/11/01 一個月所有開盤日之收盤價、漲跌預測及買賣建議。
- ✓ 主要使用技術:Python、Pytorch、fastText

2.3 蒐集資料-歷史股價

資料來源 Yahoo Finance。在此說明選取一年之原因，由於財經新聞之取得為一年期間之資料，因此歷史股價乃為配合新聞之資料，故選取 2017/10/1 至 2018/11/1 之資料期間。

除了使用開盤價、最高價、最低價、收盤價、調整後收盤價及成交量之股票基本資訊外，另外計算數十種技術分析指標做為輸入模型之特徵。

2.4 蒐集資料-財經新聞

財經新聞來源為鉅亨網之一年期間財經新聞，共 30304 篇。

如上所述，財經新聞的處理分為兩種:一為以關鍵字提取目標股票之新聞(台積電)，二為以關聯度之方法提取與台積電具備高關聯度之新聞做為模型之特徵。

兩種方法本研究皆是採取將文章化為向量之方法，不同的地方在於一為先以關鍵字取出新聞文章後再將之轉為向量，二為先行將文字轉為向量，再藉由計算其關聯度提出。

2.5 關聯式提取特徵方法-財經新聞

蒐集財經新聞資料之後，本研究首先使用 Gensim(fastText)將所有文章化為向量，至於選用 fastText 原因在於其極為高速的訓練速度，且在速度快的基礎上訓練的成果也相當具有水準，故本研究採用 fastText 做為將文字轉為向量之工具。轉化向量示例如下圖 4。

（圖 4，fastText 示例）

本研究以餘弦相似性(Cos(θ))計算兩向量之間的關聯度，接著將詳述做法。

$$\mathbf{a} \cdot \mathbf{b} = \|\mathbf{a}\| \, \|\mathbf{b}\| \cos\theta$$

（圖 5，歐基里德點積公式)

根據圖 5 之歐幾里得點積公式，可得兩向量之餘弦值，僅需將等號兩邊變化，即可得到所需之餘弦相似性，如圖 6。

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

（圖 6，Cos(θ)）

其中 A_i 及 B_i 分別代表兩向量 A、B 之各分量。而根據此公式計算之餘弦相似性範圍為-1 至 1，-1 意味著兩向量指向之方向正好截然相反，1 則表示兩向量是完全相同的。

而此次取出之新聞為 Cos(θ)大於 0.7 之向量，Cos(θ)值之計算由 30304 篇財經新聞向量與本研究所選取之中心文章向量計算而來，而中心文章源自於台積電官網介紹(圖 7)以及維基百科(圖 8)之台積電介紹為基準。

（圖 7，台積電官網)

(圖 8，維基百科）

此外，在新聞資料備妥之同時，還面臨了兩個問題:

1.六日及國定假日未開盤

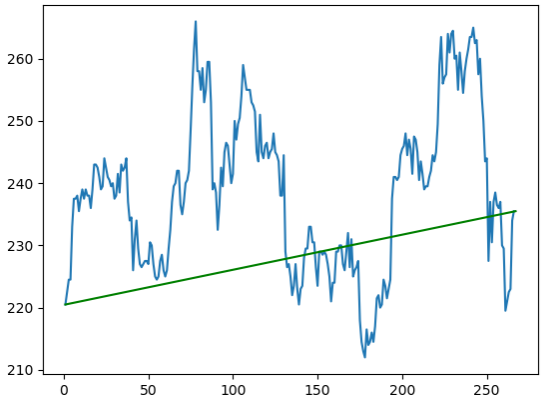
2.一天中可能會有多篇新聞

對於第一個問題，本研究會將未開盤日之新聞向量加入前一次開盤日，以此做為預測下一次開盤的參數。

第二個問題，採取的方法是將一天中所有的新聞向量做加總。

2.6 買賣建議及漲跌趨勢

買賣建議起初為提供最大利潤之買賣建議，參考圖 9，在初始點與終結點之間建構一條線，此時我們便可以根據圖 9 之線線得到每一天的對應股價，通過與每天的真實股價計算後我們可以知道每天的差值，而透過差值即可決定於差值達到多少門檻時進行交易。



（圖 9，買賣交易點決策線）

但是，經過多次試驗之後，考量最適當交易於一年期間的股價資訊下，交易次數過於稀少，導致機器學習成效相當差，故本研究於訓練資料中採用超頻繁交易，意即只要遇到一次漲跌便做交易，即使利潤僅有 0.5 元。以此方法便有足夠的交易次數供機器學習，但是若在資料量充足的情況下，建議依照開頭所述之方式做交易點之建議。

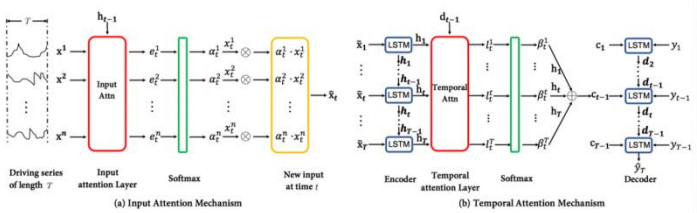
而在漲跌趨勢方面則相當單純，僅需在訓練資料集計算每日的漲跌情況給予模型學習即可。

2.7 預測模型

深度學習的模型本研究採用 **A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction** (DARNN)

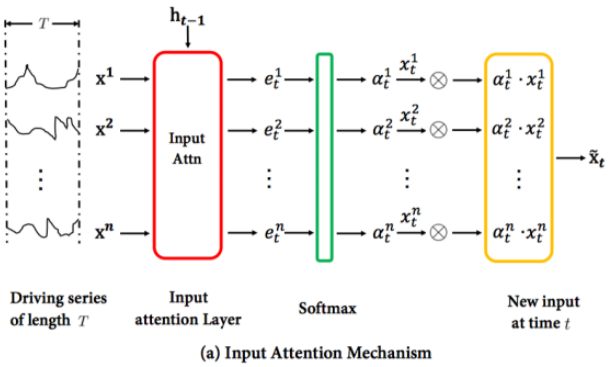
Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, Garrison W. Cottrell, A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction, IJCAI, 2017.

該篇論文所提出的 DA-RNN(雙階段注意力機制的循環神經網絡)，其包含兩個具有注意機制的 LSTM 網絡(Encoder-Decoder)，與傳統的注意機制(僅用在 Decoder 的階段)不同的是，在 Encoder 的部分引用一種新穎的輸入注意機制，而論文已證明此方法較傳統 Encoder-Decoder 模型以及普通 RNN 模型更加準確。圖 10 為該源自該論文中對於 DARNN 模型之介紹。



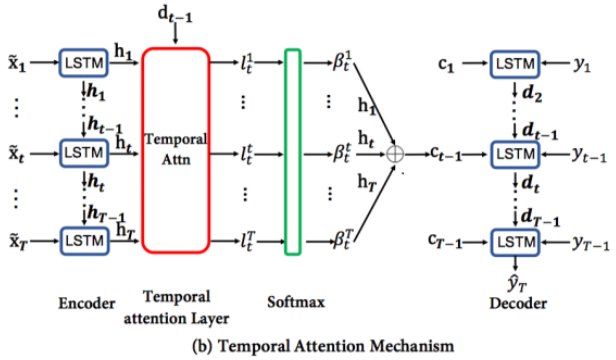
(圖 10，DARNN 概述)

請參考圖 11 及圖 12，接著將會闡述此模型於本研究中如何使用。



(圖 11，DARNN 概述-第一階段)

於該模型之第一階段(圖 11)，首先先定義一個 Window Size 作為遍歷訓練資料之大小，接著使用當前時刻的輸入 X_t 以及上一個時刻之 hidden state 來計算當前時刻之 hidden state，而接下來即是此 DARNN 之特點，作者在此引用了注意力機制，在關注完重要信息後，產生具有權重的新 X_t 做為下一個階段的輸入。以本研究來說，此注意力機能關注技術指標、股價等等的重要程度，且其會根據每天的交易情形不同而有不同的關注程度。



(b) Temporal Attention Mechanism

(圖 12，DARNN 概述-第二階段)

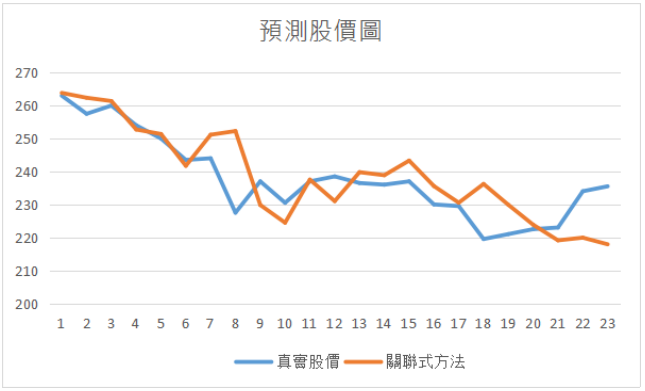
在上一階段中得到新的輸入之後，第二階段即為較傳統的注意機制，在此階段中將會對所有時刻的 h_t 取加權平均，因此可以在不同的時刻採用不同的 Context vector 。以此次研究來說，在第二階段當中將會關注過去數天的變化，且每一次預測所關注的重點皆不同。

總結來說，DARNN 模型在兩階段的關注機制皆不同，因此更能捕捉到重要的信息，且該篇論文作者在 Nasdaq100 數據集上實現了很好的預測結果，因此本研究以 DARNN 做為預測模型之使用。

3 Result

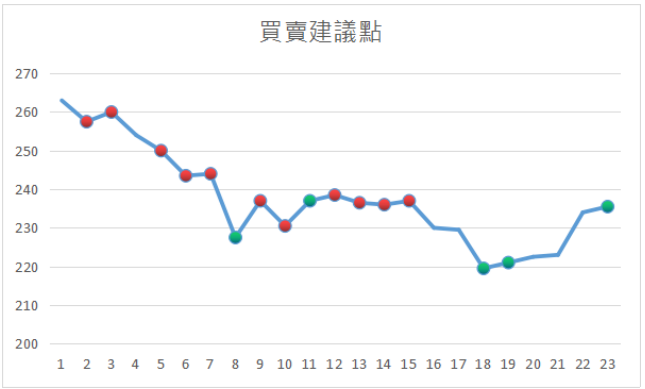
(圖 13，預測結果及真實股價)

經過 10000 Epochs 之後可得到圖 13 之預測結果及真實股價之差異，可見得在訓練資料期間成果滿意，不過在測試資料之預測上似乎就顯得沒那麼完美。當然以此圖無法準確地看出資訊，接著將會慢慢地說明成果。



（圖 14，關聯式方法收盤價價預測結果）

於圖 14 中，本研究呈現以關聯式方法提取新聞做為特徵之方法的預測成果，在第三、四、五六天之開盤日(2018-10/3、10/4/、 10/5、10/8)預測結果相當好，但在之後面臨外 資無預警大賣以及官股買超回穩之特定事件發生情況下，預測結果便差強人意。



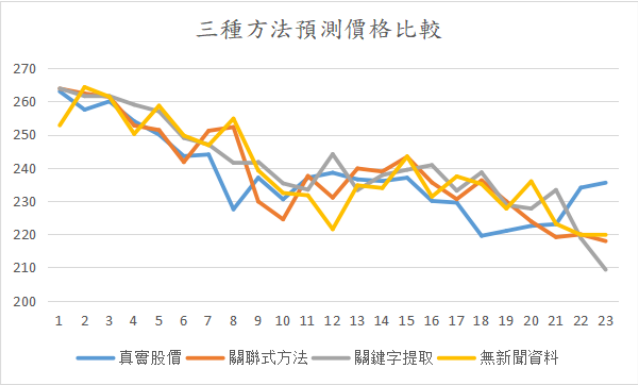
（圖 15，關聯式方法買賣點建議）

圖 15 為關聯式方法之買賣建議點預測，綠點代表的為買進建議，紅點代表賣出建議。

如 2.6 章節所述，本研究此次乃是採取超頻繁交易之作做法，因此模型之預測結果給予的買賣建議相當頻繁。遵循「於模型第一次建議買時即買進，買進後模型第一次建議賣時即賣出」的規則來看的話，此買賣建議點雖說並無給予最佳之買賣建議，但是就獲利上來說成果仍然不差。

上述僅僅呈現了關聯式提取新聞資料之方法的成果，若是沒有相比較並無法證明該方法是優於「關鍵字提取特徵方

法」以及「僅參考歷史股價」，所以將在下面做更詳細的介紹。



(圖 16，三種方法與真實股價之預測圖)

圖 16 中展示三種不同方法對於真實股價的走勢差異，可以看出在在第三、四、五六天之開盤日(2018-10/3、10/4/、10/5、10/8)中以關聯式的最為準確。在某些時間點中關鍵字提取的方式也展現了不錯的成果，但是未參考新聞資料的股價預測結果則頻頻差強人意。

僅僅觀看如此密集且複雜的圖表仍然不夠清楚，因此下方將會統整出三種不同的方法之於股價預測結果、漲跌預測正確率以及買賣建議點獲利(根據上所說之規則計算)的比較，同時必須聲明在模型的選擇上以及所有參數設定皆相同。

方法	股價平均差	趨勢預測準確度	買賣建議之獲利潤
無新聞資料	9.05	26%	None
關鍵字提取方法	7.19	26%	-16
關聯式提取方法	6.35	39%	+11

上方表格之股價平均差為預測股價與真實股價相減後取絕對值相加取平均，可得平均每日的預測與真實股價之差異，而準確度則是查看真實股價實際漲跌程度決定，買賣建議則是遵循「於模型第一次建議買時即買進，買進後模型第一次建議賣時即賣出」的規則(None 即為建議全部不交易)。

4 Conclusions

從第三章節最後的統計的表格中我們可以得知在三種結果上皆以本研究之題目「關聯式提取新聞特徵之方法」為優，其中尤以買賣建議之獲利最為突出，且無參考新聞資料的模型在各方面的結果中都表現較差，從此也可間接得知財經新聞對於股票預測模型的重要性。

從以上的結果總結可得:在股票的領域中對於目標股票之股價預測，且在參考新聞特徵的情況下，以關聯式方法提取對於預測結果有較好的成果。

就整體預測結果來看，預測的準確度還有許多進步空間，對於股票市場的一些無預警之拋售及大量買進，以及一些導致股價大幅波動的訊息，此次的成果並無法完美的捕捉，但本研究已證明「關聯式提取新聞特徵之方法」的結果為最佳，未來在準確度的精進將會是研究的主要方向，但我們不得不承認，股票市場仍是一頭難以馴服的野獸。

最後，非常感謝在此次研究中共同努力的同仁、多方提供建議的老師、論文及參考資料的撰文者，以及在 Python Taiwan、Pytorch Taiwan、Stack Overflow 中提點的各位高手及前輩們，本次的成果仍然有非常大的進步空間，而我們也將不斷地進行修正及進步，期望帶來美好的成果。

6 REFERENCES

[1] **A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction** (DARNN)
Yao Qin, Dongjin Song, Haifeng Chen, Wei Cheng, Guofei Jiang, Garrison W. Cottrell, A Dual-Stage Attention-Based Recurrent Neural Network for Time Series Prediction, IJCAI, 2017.

[2] https://github.com/Zhenye-Na/DA-RNN

[3] https://zhuanlan.zhihu.com/p/36440240

[4] https://arxiv.org/pdf/1704.02971.pdf

[5] https://github.com/facebookresearch/fastText

[6] https://fasttext.cc/

[7] https://www.finlab.tw/%E8%82%A1%E7%A5%A8%E5%85%A5%E9%96%80%E6%87%B6%E4%BA%BA%E5%8C%85/

[8] https://pytorch.org/

[9] https://ithelp.ithome.com.tw/articles/10198796

[10] https://brohrer.mcknote.com/zh-Hant/how_machine_learning_works/how_rnns_lstm_work.html

[11] https://ithelp.ithome.com.tw/articles/10193678

[12] https://morvanzhou.github.io/tutorials/machine-learning/torch/

[13] https://github.com/dzitkowskik/StockPredictionRNN

[14] Python Tainwan-Facebook

[15] Pyttorch Tainwan-Facebook