國立中正大學

資訊管理研究所

碩士論文

結合社群評論預測股價漲跌趨勢：以 S&P500 成份股為例

Predicting Stock Price Trends using Social Reviews: Evidence from the S&P 500 Firms

研究生：陳冠宇 撰

指導教授：李珮如 博士

中華民國一一一年七月

致謝

回首兩年前，期待著能在中正大學好好鍛鍊自己、充實自我，時間過得很快很短暫，很幸運在碩士期間能認識許多朋友，使我在研究所期間增廣見聞，也非常感謝我的指導教授李珮如博士，在論文指導時沒有給予太大壓力，而是讓我有很大的空間盡情發揮自己所學，並也會讓我嘗試許多自己想要探索的方向，以及當我論文進度遇到瓶頸時，教授也會不厭其煩地為我提供她的想法、不辭辛勞的為我解惑，並且教導很多的專業知識，除了專業知識外，也培養了做事態度，感謝教授在這期間給予的教導。

特別在此感謝李珮如教授、許巍嚴教授、吳帆教授在我論文提案時給予的建議，讓我的碩士論文能有更好的呈現。同時感謝 DSBDA 的彥筑、珮妤、恩安、渝婷，讓我的研究生活變得更精彩，在這裡需特別感謝許家的沛慈，在做研究的過程可以互相幫助，也感謝許家的宗儒提供不少關於神經網路的建議，另外感謝提供額外電腦的謝清福學弟，使我在深度學習的實驗中可以節省時間，有了上述同學們的幫忙，讓我論文能順利進行。

最後感謝家人的栽培，讓我能有最後一段無後顧之憂的求學生活，希望大家保持身體健康、平安喜樂，以本文獻給家人、朋友、師長。

陳冠宇 謹致

中正大學資訊管理研究所

中華民國一一一年七月

摘要

近年來社群論壇的蓬勃發展，許多人可以在網路上輕易的接收到其他民眾發表的訊息言論，而投資者受到社群資訊影響其投資決策的情形也日益普遍，這些社群資訊可能來自社群媒體、線上新聞、論壇討論區，也包含特定族群如員工論壇討論區等，其中有些投資者可能也想參考該公司員工之工作感想，許多人也可能因為接收到這些來自資訊而被吸引進入投資市場以及影響其投資決策。

過去研究顯示，股價預測已經包含了相當多不同面向的分析，可以概括成基本面分析、技術面分析、籌碼面分析，以及以新聞、社群為主軸的消息面分析等，而本研究選擇著重於消息面分析，原因在於對許多投資者而言，基本面、技術面、籌碼面均須要一些專業知識上的了解，而消息面的資訊取得容易也較容易理解，但也可能因為不同人發表的言論其主觀不同，造成投資者無法快速的做出投資決策，因此本研究希望能藉由股價走勢預測模型，輔助投資者在接收到社群之消息面資訊時，更有效的進行投資決策。

由於社群消息面資訊來源相當廣泛，加上每個投資者習慣瀏覽的社群網站皆不盡相同，以員工評論來說，本研究採用較廣為人知的 Glassdoor 作為文本資料來源，並且蒐集 Yahoo Finance!的各公司股票、技術面相關數據，以及 Macro Trends 之各公司基本面數據，並使用隨機森林(Random Forest)、極限梯度提升(Extreme Gradient Boosting)、循環神經網路(Recurrent Neural Network)、長短期記憶(Long Short-Term Memory)建立股價走勢預測模型，希望在未來能提供投資者進行投資決策時的參考。

關鍵字:股價趨勢預測、文字探勘、資料探勘、機器學習、深度學習、消息面分析

Abstract

With the boom in social forums in recent years, many people can easily receive the information and opinions from other people on the Internet, and it is becoming increasingly common for investors to be influenced by social information. Many people may also be drawn to the investment market and have their investment decisions affected by the information they receive from these sources.

Past research shows that stock price prediction has included many different aspects of analysis, which can be summarized as fundamental analysis, technical analysis, chip analysis, and news analysis including news and social media. This study focuses on news analysis because fundamental, technical, and chip analysis require some professional domain knowledge, while news information is easy to obtain and understand. Therefore, this study hopes to use the stock price prediction model to assist investors to make investment decisions more effectively when receiving news from the community.

Due to the wide range of sources of social news and the different social networking sites that each investor is accustomed to browse, this study uses Glassdoor, a well-known employee forum, and Reddit as the textual data source, gathering company's stock price and technical data from Yahoo Finance, combining with fundamentals data from Macro Trends, then used Random Forest, Extreme Gradient Boosting, Recurrent Neural Network, and Long Short-Term Memory to build a stock price prediction model. We hope to provide investors with a reference for making investment decisions in the future.

Keywords: Stock price trend prediction, Text mining, Data mining, Machine learning, Deep learning, News analysis

目錄

圖目錄

表目錄

第一章、 緒論

1.1 研究背景

股票市場是一個國家經濟的重要組成部分之一，也是公司為了籌措資金的一種方式，不僅是專業投資人，一般民眾也將其視為一種投資工具(Billah et al., 2016)，而股票價格也反應了投資人對公司未來成長的預期程度(Z. Hu et al., 2021)，學者 Smith (1937)於著名的《國富論》(The Wealth of Nations)中曾提出「公司的目標是為股東創造利潤」，而公司的營運狀況好壞將很大程度的影響公司股票價格，因此能夠準確地預測公司營運狀況對於投資人的投資決定是很重要的，特別是股票屬於高風險以及高報酬的金融商品(Chen et al., 2013)，如果能準確的投資將可以獲得高額的財務收入和降低市場風險(Kumar & Thenmozhi, 2006)。

學者 Fama (1970)曾提出著名的「效率市場假說」(Efficient Market Hypothesis [EMH])，並提出三個重要假設以說明股價的難以預測性：(1)所有投資人都是理性的，且都追求最大報酬；(2)新的市場資訊為隨機，好壞都有可能；(3)股價能反應該資產的全部資訊，且股價為隨機；但學者 Lo (2005)指出投資人有時會帶著情緒做出決策，並不會總是理性的，因此效率市場假說並不一定適用。且股票價格被多種因素所影響(Lu et al., 2021)，如學者 Roubaud and Arouri (2018)研究顯示油價、匯率和股票市場間存在相互作用。而 Huang et al. (2019)也提到股價變動會與公司業績、政府政策、通貨膨脹率等相關。以及 Kyoung-Sook and Hongjoong (2019)研究顯示股價會受國際形勢及國內外經濟環境所影響。上述研究表明，有太多因素會影響股票的價格，許多投資人因此希望能參考更多其他面向資訊以幫助投資決策。

而過去研究顯示，大多數投資者使用基本面、技術面、籌碼面來分析股價：其中使用基本面之研究主張，個別公司的營運狀況，如每股盈餘(Earnings Per Share [EPS])、本益比(Price to Earnings Ratio [PE])等與公司股價有相關性，可用來預測股價(Ballings et al., 2015; Namdari & Li, 2018; Shiva Nandhini et al., 2020; Yetis et al., 2014)。使用技

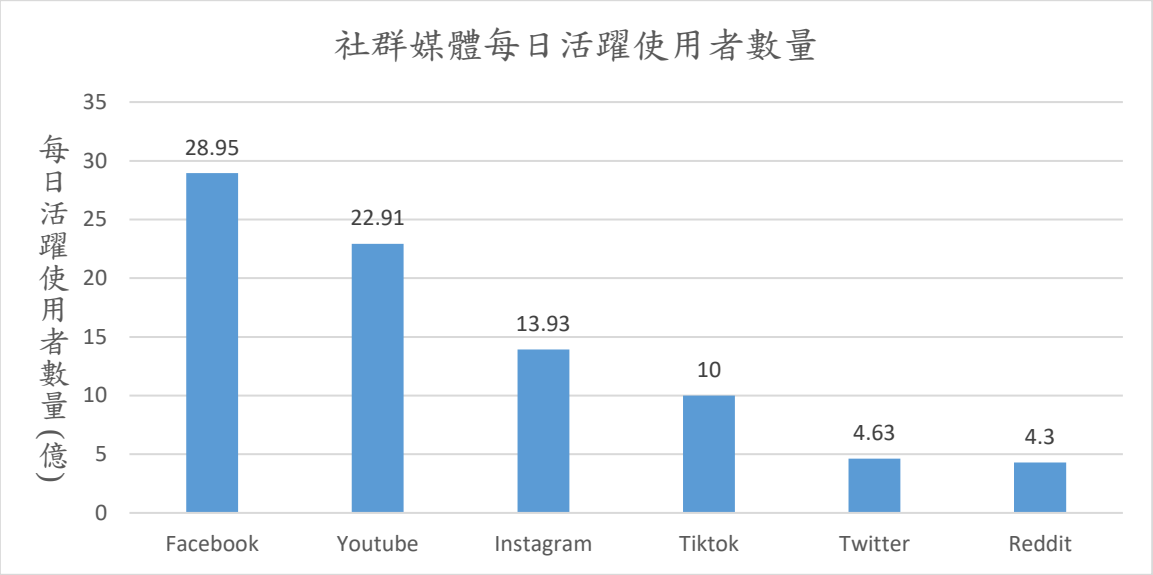
術面之研究則認為，可以透過研究市場過去的資料，如股價、交易量、日線及周線等來預測股價未來的走向(Adebiyi et al., 2012; Göçken et al., 2016; Khare et al., 2017; Mizuno et al., 1998; Picasso et al., 2019)。使用籌碼面之研究則主要為觀察大戶及法人之交易動向、融資融券比例、持股比例等為主（查欣瑜，2011；黃韻欣，2020；謝璁賦，2011）。

亦有部分研究使用消息面來預測股價，消息面認為，投資人的情緒是金融市場中的重要因子(Nofsinger, 2001)，以及投資人之情緒將影響他們的決策(Guo et al., 2017)。學者 Peng (2019)也提出投資者情緒與股價變化有正向相關性。而這方面的研究可分為如：(1)使用金融新聞文本中之公司發布消息、相關新聞等表達出的情緒反應(Li et al., 2014; Ren et al., 2018)。(2)使用社群媒體資訊(Ko & Chang, 2021; Si et al., 2013)。

自從社群媒體誕生之後，很多人會開始使用社群媒體(如 Twitter、Reddit)來與他人互動、發表言論(Jianqiang et al., 2018; Kim, 2020; Shim & Pourhomayoun, 2017)，而這些使用者發表的言論被稱為使用者生成內容(User-Generated Content, UGC)，意指由使用者所創造之內容，包含社群貼文留言、員工評論。而根據 Statista 於 2021 年公布之世界最受歡迎社群媒體排名，如圖 1 顯示，Twitter 每日活躍使用者可達 4.63 億人，而另一社群平台 Reddit 之每日活躍使用者也有 4.3 億，由此可見社群媒體的使用者之廣泛，也可以預期 UGC 資料會隱含大量的使用者情緒。而過往也有許多學者以 UGC 資料，如 Twitter 來進行情感分析並萃取其中的使用者情感並用於股價預測(Bharathi et al., 2017; Skuza & Romanowski, 2015)。

圖 1

世界最受歡迎社群媒體之每日活躍使用者數量排名



資料來源: Statista Research Department. (2021, Nov 16). *Most popular social networks worldwide as of October 2021, ranked by number of active users*. Statista. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Reddit 擁有相當多的使用者及不同的討論版，關於金融投資、股票等就有 stocks、StockMarket、investing、wallstreetbets 等數十種討論版供使用者進行討論，使用者可以透過標題、內文來表達自己對於個股、股市的想法，而其他使用者可以在下方按讚或是留言來進行討論。Teoh et al. (2019)收集 Reddit 新聞版的貼文後進行情感分析，再結合納斯達克指數之每日價格來預測股票走勢，結果顯示加上情感分析後有益於預測股票價格走勢。

圖 2

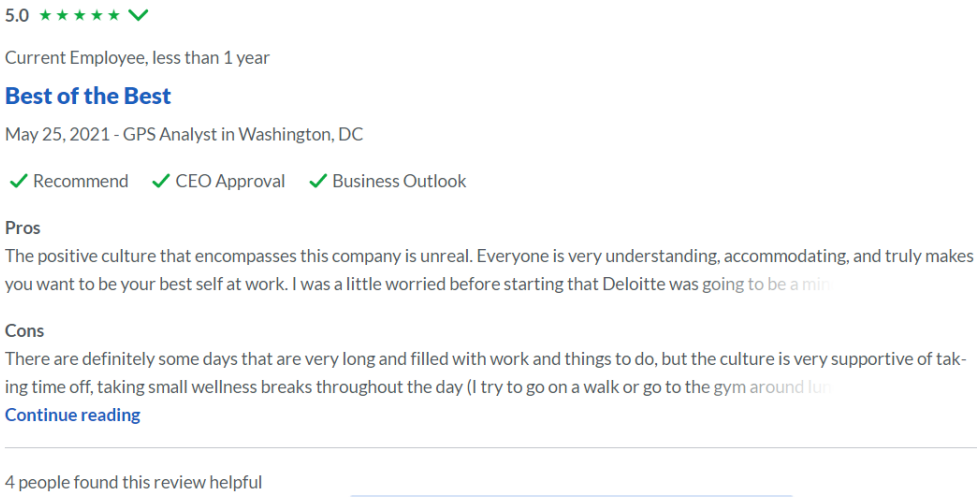
Reddit 社群評論示意圖

資料來源: ZhangtheGreat. (2021, Nov 26). *Costco (COST) - Why does it just keep running?* Reddit. https://www.reddit.com/r/stocks/comments/r2advu/costco_cost_why_does_it_just_keep_running/

社群媒體的普及同時也帶動員工評論網站(如 Glassdoor、Indeed.com)的出現，各家公司的員工可以在網站上分享自己對於公司的評價，包含公司的優缺點、對於企業執行長的評價，以及給予各項評分，是為一種新型態的 UGC 資料。Moniz and de Jong (2014)提出，將 Glassdoor 員工評論利用 General Inquirer 情感字典進行情感分析所產生之員工滿意度對於預測公司收益有益，並以此提出員工滿意度能代表公司營運狀況。

圖 3

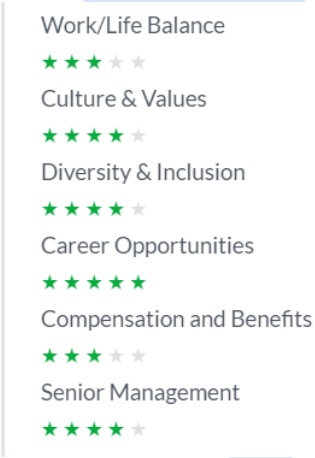
Glassdoor 員工評論示意圖



資料來源：Anonymous. (2021, May 25). *Best of the Best*. Glassdoor. https://www.glassdoor.com/Reviews/Deloitte-Reviews-E2763.htm.

圖 4

Glassdoor 員工評論評分示意圖



資料來源：Anonymous. (2021, May 25). *Best of the Best*. Glassdoor. https://www.glassdoor.com/Reviews/Deloitte-Reviews-E2763.htm.

1.2 研究動機

股價是一種時間序列的資料(Yu & Yan, 2020; Zhao & Chen, 2021)，時間序列是指資料是按照時間進行順序排列的，例如天氣、匯率也屬於時間序列的資料。在分析時間序列的資料時有兩個特點需要注意：(1)平穩性(Stationary) (2)季節性(Seasonality)。平穩性，或稱定態性，是指資料在觀察到的時間內並沒有顯著變化，並始終維持在固定範圍，也就是長期來看不會有增加或減少的趨勢，可以解釋為時間對於資料將沒有影響，因此時間序列的資料應為非平穩性。季節性則是資料會隨著觀察時間內的特定規律而產生變化，如旅遊業有旅遊淡季、旺季之分，將會影響模型的預測性能。

在早期的股價預測相關文獻中，許多學者會利用一些統計模型如 AutoRegressive Integrated Moving Average (ARIMA)進行股價預測，但 ARIMA 模型假設時間序列資料需要是平穩的，或是需要將非平穩的時間序列資料利用差分轉換為平穩時間序列資料。而且統計模型需要進行許多假設(如資料為線性關係)，可能無法捕捉到非線性的資料。(Chen et al., 2020; Earnest et al., 2005)，且也有學者提出股價其實並非平穩性的資料(Polanco-Martínez, 2019; Wang et al., 2012)。隨著近年來資料探勘及人工智慧領域的成熟，越來越多學者嘗試利用機器學習、深度學習預測股價(Adebiyi et al., 2012; Chen & Hao, 2017; Kamble, 2017; Quah, 2008)。其中許多學者發現利用深度學習比起統計模型，由於能捕捉到資料間非線性的關係，能提升股價預測準確度。Siami-Namini et al. (2018)提出利用 ARIMA 統計模型與深度學習的長短期記憶神經網路(Long Short-Term Memory [LSTM])對比，結果表明 LSTM 預測顯著提升了預測準確度。Mohan et al. (2019)利用 ARIMA 與循環神經網路(Recurrent Neural Network [RNN])在五間公司上進行比較，RNN 在五間公司上都顯著提升預測模型準確度。

隨著現今媒體及社群媒體的蓬勃發展，許多人會參考新聞，或在線上金融論壇、社群網站上討論股票的相關訊息，而這方面的股市資訊來源被稱為消息面，消息面的過去研究主要包含：財金、金融領域相關之新聞(Chen et al., 2019; Deng et al., 2011; Lee & Soo, 2017; Li et al., 2014; Schumaker & Chen, 2009)以及社群網站評論如 Twitter 的評

論、留言，或是一些線上金融股票論壇如 Stocktwits、Reddit 的貼文以及回覆(Bollen et al., 2011; Duan & Zeng, 2013; Jin et al., 2020; Ko & Chang, 2021; Lubitz, 2017; Si et al., 2013; Vu et al., 2012)。關於基本面、技術面、籌碼面之先前研究，大部分為數值類型數據，屬於結構化資料。

而消息面的文字的相關數據，則歸屬於非結構化資料。目前已有許多學者將非結構化的消息面資料，利用文字探勘，提取出如關鍵字、情感等，來做為股價預測的變數。Ranco et al. (2015)從 Twitter 推文提取文字情感，發現在 3 日內的情感極性與其股票回報的相關性是顯著的。Jin et al. (2020)收集 Stocktwits 論壇貼文，情感分析後預測蘋果公司收盤價，提出 Stocktwits 貼文有益於股價預測。Wooley et al. (2019)提取 Reddit 貼文中的情感，在預測比特幣價格上比起只使用價格數值預測，獲得更佳的準確度，並以此提出公共情緒之重要性。

而在員工滿意度、員工評論與企業營運指標(基本面指標)的相關研究中，根據學者 Schneider et al. (2003)研究顯示，員工滿意度對資產報酬率(Return On Assets [ROA])、EPS 有正向影響。Edmans (2011)研究指出員工滿意度與股東回報有相關。Huang et al. (2015)也發現，員工總評分與企業的市值有因果關係。Luo et al. (2016)也以 Glassdoor 之員工評論研究，發現員工評論產生之員工滿意度與其市值相關。Feng (2020)以 Indeed.com 網站之員工評論研究，發現員工評論評分每上升 1%，市值會上升約 0.68%。

綜合以上研究，發現過去研究使用員工滿意度預測 EPS 等企業營運指標，並證明員工滿意度與許多基本面指標等有正相關。但員工滿意度大多使用問卷調查得來，但亦有研究使用 UGC 資料如 Glassdoor 來獲得員工滿意度，因此本研究使用 UGC 資料(Glassdoor)來分別取得員工滿意度及社群使用者情感，並使用深度學習的方式來預測股價。

1.3 研究目的

故本研究將嘗試結合社群評論建立一股價漲跌趨勢預測模型，將 S&P500 成分股依照全球行業分類標準(Global Industry Classification Standard [GICS])分類，在各行業選出代表性公司，再收集代表性公司的股價資料與相關公司營運指標、技術面指標等結構化變數，結合 Glassdoor 員工評論變數、情感分析之非結構化變數，並將代表性公司分成大資本與小資本，利用機器學習、深度學習建立股價預測模型，並經由文獻整理以及實驗過程找出影響股價因子，最後透過實驗結果探討大資本與小資本的影響因子及股價間之變動關係，以提供投資者在大小資本類別間做投資決策之參考。

第二章、 文獻探討

2.1 基本面、技術面、籌碼面之股價預測

2.1.1 基本面分析

基本面分析之過去研究主要為針對個別公司財務資訊之分析，至於個別公司財務資訊主要來自於公司的財務報表(Financial Statements)，像是損益表(Income Statement)、資產負債表(Balance Sheet)、現金流量表(Cash Flow Statement)、權益變動表(Statement of changes in equity)，從中衍伸出的各項財務指標，如營業收入(Revenues)、營業毛利(Gross Profit)、EPS、ROA 等每年或每季財務報表數據，並針對這些數據進行資料探勘，以找出一些深藏在其中的有價值的資訊(Joshi & Li, 2016)。

如學者 Ballings et al. (2015)就利用資產負債表及損益表得出的本益比、流動比率(Current Ratio)、股價淨值比(Price Book Ratio [PB])、EPS、股東權益報酬率(Return On Equity [ROE])、淨利率(Profit Margin)、ROA、資產負債率(Debt Asset Ratio)等財務指標，透過支持向量機(Support Vector Machine [SVM])、羅吉斯回歸(Logistic Regression [LR])、隨機森林(Random Forest [RF])等機器學習演算法建立預測模型，來預測長期的股票價格漲跌，並最終發現 RF 具有最高的預測準度(AUC = 0.9)。

Namdari and Li (2018)使用了流動比率、存貨周轉率(Inventory Turnover)、股東權益報酬率、淨利率、資產報酬率、固定資產周轉率(Fixed Assets Turnover)、營運資金周轉率(Working capital turnover)、速動比率(Quick Ratio)、負債權益比率、資產負債率等財務指標，結合每日股價開盤價、收盤價、最高價、最低價等數據，透過多層感知機(Multi Layers Perceptron [MLP])演算法建立預測模型，來預測納斯達克成份股的長期股票漲跌，實驗過程發現對比單獨使用財務指標或是股價數據，結合所有指標後，模型顯著增加了預測準度。

本研究列出過去文獻使用之基本面財務指標，整理於表 1。

表 1

基本面分析之相關文獻

作者 (年份)	目標變數	財務指標															
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Tong-Seng Quah (2008)	長期 NASDAQ 漲跌	√	√	√			√	√									√
Emir et al. (2012)	個股股價漲跌	√					√	√		√							
Ballings et al. (2015)	長期個股漲跌	√	√	√			√	√	√	√				√	√		√
Namdari & Li (2018)	NASDAQ 成份股隔日收盤價漲跌		√			√		√	√	√	√	√	√	√	√		
Tan et al. (2019)	長期個股漲跌	√		√	√			√		√							√
Gao et al. (2020)	S&P500 隔日收盤價															√	√

註:

1. P/E:本益比、2. Current Ratio:流動比率、3. P/B:股價淨值比(Price-Book Ratio)、4. P/S:股價營收比(Price-to-Sales Ratio)、5. Inventory Turnover:存貨周轉率、6. EPS:每股盈餘、7. ROE:股東權益報酬率、8.Profit Margin:淨利率、9. ROA:資產報酬率、10.Fixed Assets Turnover:固定資產周轉率、11. Working capital turnover:營運資金周轉率、12.Quick Ratio:速動比率、13. Debt-To-Equity Ratio:負債權益比、14. Debt Asset ratio:資產負債率、15. Exchange Rate:美金匯率、16. 其他財務指標

資料來源:本研究整理

2.1.2 技術面分析

技術面分析是指透過研究過去市場上的股價歷史數據，將其量化及圖形化，使其較易被人們理解，並以此來預測股價未來走勢。技術面主要分為指標法及型態法(Chavarnakul & Enke, 2008)，指標法是利用股價歷史數據的各項數值組成的指標預測股價，而型態法則是利用如 K 線圖研究股票價格的變動。

學者 Huang et al. (2008)使用了 S&P500 每日開高收低價、以及簡單移動平均線(Moving Average [MA])、指數移動平均線(Exponential Moving Average [EMA])、指數平滑異同移動平均線(Moving Average Convergence / Divergence [MACD])、相對強弱指數(Relative Strength Index [RSI])、威廉指標(William %R)等 23 個技術指標，先將技術指標利用不同特徵選取的方法搭配 SVM，選出最佳特徵組合後，再利用 SVM、K 近鄰演算法(K-Nearest Neighbors [KNN])、反向傳播(Back-Propagation [BP])、決策樹(Decision Tree [DT])、羅吉斯回歸建立集成學習(Ensemble Learning)預測模型，最後在韓國股票指數、台灣加權股票指數測試，發現集成學習模型皆能達到最好的預測效果。

學者 Chen et al. (2020)使用了 S&P500 每日開高收低價、以及 MA、EMA、三角移動平均線(Triangular Moving Average [TMA])、RSI、布林通道(Bollinger Band)、威廉指標等 27 個技術指標，首先利用 Pearson 相關係數(Pearson correlation coefficient)找出與收盤價最顯著正相關的 14 個技術指標後，再結合自然資源價格(黃金、白銀和石油)、Google Index 上的 S&P500 搜尋熱度作為變數，並利用 MLP 結合雙向長短期記憶(Bi-directional Long Short-Term Memory [Bi-LSTM])及注意力機制(Attention Mechanism [AM])建立預測模型，最後在 S&P500、Russel 2000、Dow Jones、納斯達克四種指數上預測收盤價，發現他們提出的模型在四種指數的測試集上都有更佳的預測性能。

本研究列出過去文獻使用之技術面指標，整理於表 2。

表 2

技術面分析之相關文獻

作者 (年份)	目標變數	技術指標												
		1	2	3	4	5	6	7	8	9	10	11	12	13
Huang et al. (2008)	S&P500 股價漲跌	√	√		√	√	√	√	√		√	√		
Adebiyi et al. (2012)	個股漲跌												√	√
Göçken et al. (2016)	個股收盤價	√	√	√			√		√	√	√			√
Picasso et al. (2019)	個股漲跌	√	√		√		√			√	√	√		√
Chen et al. (2019)	個股收盤價	√			√		√				√		√	√
Chen et al. (2020)	S&P500 收盤價	√	√	√						√				

註:

1. MA:移動平均線(Moving Average)、2. EMA :指數移動平均線(Exponential Moving Average)、3. TMA:三角移動平均線(Triangular Moving Average)、4. MACD:指數平滑異同移動平均線(Moving Average Convergence / Divergence)、5. VR:成交量比率(Volume Ratio)、6. RSI:相對強弱指標(Relative Strength Index)、7. OBV:能量潮(On Balance Volume)、8. MTM :動量指標(Momentum Index)、9. BO:布林通道(Bollinger bands)、10. W%R :威廉指標 (Williams %R)、11. OSC :振盪指標 (Oscillator)、12. VO:交易量(trading Volume)、13. 其他技術指標。

資料來源:本研究整理

2.1.3 籌碼面分析

籌碼面則著重在分析主力大戶、三大法人(外資、投信、自營商)的動向，主要是因為主力大戶及三大法人對於公司持股量佔有很大比例，且大戶的股票交易量也相當多，因此可以對股票價格產生很大的影響力。所以可以透過觀察他們的股票買賣情形，進而推測未來股票漲跌趨勢。學者 Lee et al. (2004)指出主力大戶可以提前得知公司內部資訊，以做出較佳決策。以台股為例，常見的籌碼面分析變數包含三大法人買賣超張數、三大法人買賣超交易金額、三大法人買賣超變動量、券資比變動量等。

本研究列出過去文獻使用之籌碼面指標，整理於表 3。

表 3

籌碼面分析之相關文獻

作者 (年份)	籌碼面指標					
	1	2	3	4	5	6
黃祺敦 (2012)	√	√		√	√	√
范聖培 (2014)		√			√	√
蔡尚翰 (2017)	√	√	√			√
張維碩等 (2018)	√	√		√		√

註:

1.三大法人買賣超張數、2. 三大法人買賣超交易金額、3. 三大法人買賣超金額占比、4.SMR (Short Selling/Margin Buying Ratio):券資比變動量、5.成交量變動量、6.三大法人買賣超變動量

資料來源:本研究整理

但由於在美國股市中，並不會揭露法人在個別公司股份中的持股比例，也就並沒有相關的籌碼面分析變數可供使用，所以本研究不採用籌碼面分析。

2.1.4 基本面、技術面機器學習相關論文

Bustos et al. (2017)於哥倫比亞股票指數的公司中，選出 25 間不同產業各具代表性的公司，變數使用 MA、MACD、RSI、OSC 等 10 種技術指標，演算法使用 Polynomial SVM 與 MLP，並透過實驗找出最佳變數選取、以及 SVM 與 MLP 超參數(Hyper Parameter)設定優化，來預測股票收盤價的漲跌，最後得出優化後的 SVM 搭配技術面指標能有相當好的預測性能。

Gao and Chai (2018)抓取 S&P500、NASDAQ100、蘋果公司股票(AAPL)之每日開高收低，變數使用 MACD、OBV、威廉指標、MTM 等 15 種技術指標，並利用 LSTM 建立預測模型，在實驗過程中加入主成分分析(Principal components analysis [PCA])進行變數選取，於 S&P500 上進行 50、100、150、200、400 天的預測，發現短期內能有相當不錯的預測準確度，同時也在 NASDAQ100 及 AAPL 上預測，結果顯示其模型在其他資料集上，具有泛化能力。

Nelson et al. (2017)收集了巴西股市指數(Brazil Stock Exchange Index)2008 年至 2015 年的每日開高收低及交易量，以及利用 Python 3 套件 TA_Lib 產生出 180 個技術面指標，使用 LSTM 建立預測模型，預測每日股票漲跌並與 MLP、RF 做比較，最後在巴西股市指數中的 4 隻成份股上進行測試，結果發現 LSTM 優於其他預測模型(約 2%)。

Paliari et al. (2021)收集了 2014 年至 2020 年，六家澳洲公司(AX1、K2F、KMD、OKJ、ORE、RHC)的每日股票價格，變數包含開盤價、最高價、最低價、收盤價、調整後收盤價、交易量，並使用 ARIMA、XGBOOST、LSTM 三種演算法進行股票價格預測之比較，結果顯示 LSTM 在每一個資料集上的表現都優於 XGBOOST，而 ARIMA 僅在其中兩個資料集取得最佳預測表現，LSTM 可在大多數資料集取得較佳表現。

本研究列出過去文獻中利用機器學習、深度學習於基本面及技術面，整理於表 4。

表 4

機器學習及深度學習之相關文獻

作者 (年份)	資料集	方法	目標變數
Ballings et al. (2015)	bureau van Dijk(歐洲公司資料集)基本面財務指標	RF	長期個股漲跌
Bustos et al. (2017)	哥倫比亞股票指數、技術面指標	SVM	短期股票漲跌
Nelson et al. (2017)	巴西股市指數每日開高收低、技術面指標	LSTM	每日股票漲跌
Namdari & Li (2018)	NASDAQ 100 開高收低、基本面財務指標	MLP	NASDAQ 成份股隔日收盤價漲跌
Gao & Chai (2018)	S&P500、NASDAQ100、AAPL 開高收低、技術面指標	LSTM	隔日收盤價
Chen et al. (2020)	S&P 500 開高收低、技術面指標、自然資源價格、Google Index 搜尋量	MLP +Bi-LSTM+AM	S&P 500 隔日收盤價
Paliari et al. (2021)	澳洲公司資料集每日開高收低、交易量	LSTM	隔日收盤價

資料來源:本研究整理

2.2 消息面股價預測

隨著媒體傳播行業的快速發展，投資人易於取得新聞資訊，因此投資人若能正確解讀財經新聞，將有助於個人對於股價的預測準確性。

學者 Li et al. (2014)在恆生指數(HSI)中，依照其產業別商業、金融、房地產、公用事業挑選了總共 50 間公司，並配合新聞日期篩選掉 2003 年以後加入 HSI 的企業，最後使用 22 間企業的開高收低股價結合香港財華社的新聞資料，以 Loughran and McDonald 情感字典產生出之情感分析變數，利用 SVM 建立預測模型來預測股票回報，結果發現結合情感分析後，預測模型有較佳準確度(約提升 2%)。

Ren et al. (2018)使用上證 50 指數(SSE 50)中依照產業別，分別挑選了幾間公司，收集公司的收盤價、交易量、技術指標，結合新浪股票論壇、東方財富網的新聞資料產生出之情感分析變數，最後使用 SVM 建立預測模型，預測股票收盤價漲跌走勢，結果發現加入情感變數後，能提升約 18%的準確率。

近年來由於社群媒體的興起，許多學者因此開始將社群評論貼文、留言等作為新面向的消息面資料，並探討股市與社群媒體之相關研究。

Liu et al. (2017)結合了東方財富網的論壇文章之情感分析變數與股票波動性變數，並使用 RNN 建模預測股票之波動性，首先在情感分析上利用論壇文章進行建模，並製作出情感分析預測模型來產生情感分析變數，接下來在 10 間公司的股票上與波動性變數結合進行訓練及預測，最後發現加入論壇文章的情感分析後的確能有效增加預測模型準確度(提升約 4%)。

Kordonis et al. (2016)使用 Twitter 貼文結合個股股價資料進行預測，首先選定一些公司，然後利用 Twitter API 搜尋有討論這些公司的推文，經過包含斷詞、刪除停用詞等文字前處理後，再產生出每一天的情感分數，並分為今日正向分數、今日中性分數、今日負面分數，加上每日收盤價、高低價差、開盤價與收盤價差、交易量等變數，並利用樸素貝葉斯(Naïve Bayes [NB])及 SVM 建立預測模型，發現加入情感分析後能有效降低預測誤差。

Lubitz (2017)利用 Reddit 經濟討論區(economics)中，關於 S&P500 之討論貼文及留言相關變數與金融時報(Financial Times)提到的 S&P500 之新聞進行比較，首先將 Reddit、金融時報之文字進行前處理後，使用 Loughran and McDonald 情感字典來分析文字情感分數，並將同一天之情感分數以日期量化，使用 SVM、RF、NB 三種機器學習演算法進行股價預測建模，結果表明，Reddit 比起新聞分析有較佳之準確度(平均提升約 2%)。

Y. Hu et al. (2021)收集四間中國旅遊業公司(麗江、凱撒旅業、黃山旅遊、中國青年旅行社)的股價資料以及這些公司在東方財富網上的相關討論貼文作為投資人情感，同時也抓取新浪財經新聞作為新聞面情感，將文字進行前處理後，並使用卡方特徵選取測試出 700 維的單字矩陣為最佳之特徵維度，接著使用多項式樸素貝葉氏(Multinomial naive Bayes)進行情感預測模型(共分為正面、中性、負面三種情感)的建立，再將預測出的情感結合每日開高收低、5/10/20 日收盤價、交易量之移動平均等變數進行股價預測模型建立，使用 SVM、ANN、XGBOOST 三種演算法，結果表明結合新聞與社群貼文後，三種演算法都能有效提升預測表現，其中又以 XGBOOST 取得最佳預測表現。

Shaikh et al. (2021)使用 Reddit 新聞版貼文結合 Twitter 推文，Reddit 部分利用 PRAW 抓取貼文，Twitter 使用 Twitter API 抓取推文，經由去除表情符號(Emojis)、特殊符號等文字前處理後，將文字資料轉換成正面、中性、負面三種極性，再結合 Google、Apple、塔塔鋼鐵三間公司之每日股價開高收低、交易量，利用 LSTM 進行建模，結果發現結合 Reddit 新聞及 Twitter 推文後，三間公司之股價預測都有效增加其模型預測準確度。

Ko and Chang (2021)採用 PTT 股票討論區之貼文及留言與中國時報、自由時報等線上新聞，在鴻海、台積電等 6 間公司上收集每日股價資料，首先將新聞內文利用基於變換器的雙向編碼器表示技術(Bidirectional Encoder Representations from Transformers [BERT])之演算法，進行情感分析產生每日的情感分數後，結合每日股票

開高收低、交易量等數據，使用 LSTM 進行建模，結果顯示結合新聞與 PTT 貼文後，預測模型的均方誤差(Root Mean Square Error [RMSE])上升了約 12%。

綜合上列文獻後，可以發現許多學者利用新聞文章及新聞留言、Twitter 社群媒體之推文及回覆、Reddit 之討論貼文及回覆進行情感分析，首先透過情感字典或是機器學習演算法先分類出文字中的情感分數，再將其作為自變數進行預測建模，並發現情感分數對於股價有其預測力，進而提出新聞及線上論壇對於股價趨勢具有一定程度的影響力。

本研究列出過去文獻使用之消息面資料及相關技術，整理於表 5。

表 5

消息面結合文字探勘之相關文獻

作者 (年份)	資料集	方法	目標變數
Li et al. (2014)	恆生指數中公司開高收低、香港財華社新聞情感分析變數	SVM	股票回報
Ren and Wu. (2018)	收盤價、交易量、技術面指標、新浪股票論壇、東方財富網新聞產生之情感分析變數	SVM	收盤價漲跌走勢
Liu et al. (2017)	中國股市收盤價、東方財富網論壇文章情感分析變數	RNN	股票波動性
Kordonis et al.(2016)	每日開高收低、每日 Twitter 推文情感分析變數	SVM	隔日收盤價
Lubitz. (2017)	S&P500 每日開收、Reddit 與金融時報之情感分析變數	Naïve Bayes	每日開盤價與收盤價差值
Y. Hu et al. (2021)	中國旅遊公司股價、東方財富網、新浪財經新聞	XGBOOST	隔日收盤價
Shaikh et al.(2021)	每日開高收低、交易量、每日 Reddit 貼文、Twitter 推文情感分析變數	LSTM	隔日收盤價
Ko & Chang (2021)	台股開高收低、交易量、線上新聞、PTT 股票版貼文	LSTM	隔日開盤價

資料來源：本研究整理

2.3 員工評論與公司營運

Luo et al. (2016)收集 Glassdoor 各公司的員工評論，並在評論中找出如尊重、品質、獎勵等關鍵字計算詞頻，產生出九種分類為員工滿意度類別，並結合員工評論評分做為資料集，再利用迴歸分析嘗試探討不同滿意度與企業市值的相關性，最後發現員工滿意度與品質、創新、團隊合作正相關，與安全、溝通、誠信等類別呈現負相關，以及員工評論評分對企業的 Tobin’s Q 有正向的影響。

Green et al. (2019)把 Glassdoor 的員工評論評分結合公司規模、市值、ROA 等變數，迴歸分析後發現，員工對雇主的評分越高(即滿意度越高)，公司的銷售收入及淨利會有正向的成長。

Feng (2020)將 Indeed 的員工評論利用 Google Sentiment Analysis 進行情感分析，產生出員工情感分數，再結合他們在 Indeed 上的評論評分，迴歸分析後發現，員工情感分數及 Indeed 的評論評分對於公司的 EPS、市值、收入呈現正向顯著的關係。

綜合以上許多學者實證結果顯示，不同員工評論網站之員工評論、評論評分、員工滿意度對企業營運指標皆具有一定程度的相關性，因此本研究將加入員工評論、員工評分、員工滿意度作為消息面之新面向資料，並對股價趨勢進行分析預測。

本研究列出員工評論與公司營運指標相關文獻，整理於表 6。

表 6

員工評論與公司營運指標之相關文獻

作者 (年份)	資料集	方法	目標變數
Luo et al. (2016)	Glassdoor 員工評論關鍵字、評論評分	迴歸分析	Tobin’s Q
Green et al. (2019)	Glassdoor 員工評論評分、公司規模、市值、ROA	迴歸分析	銷售收入、淨利
Feng et al. (2020)	Indeed 員工評論情感分數、評論評分	迴歸分析	EPS、企業市值、收入

資料來源:本研究整理

第三章、研究方法

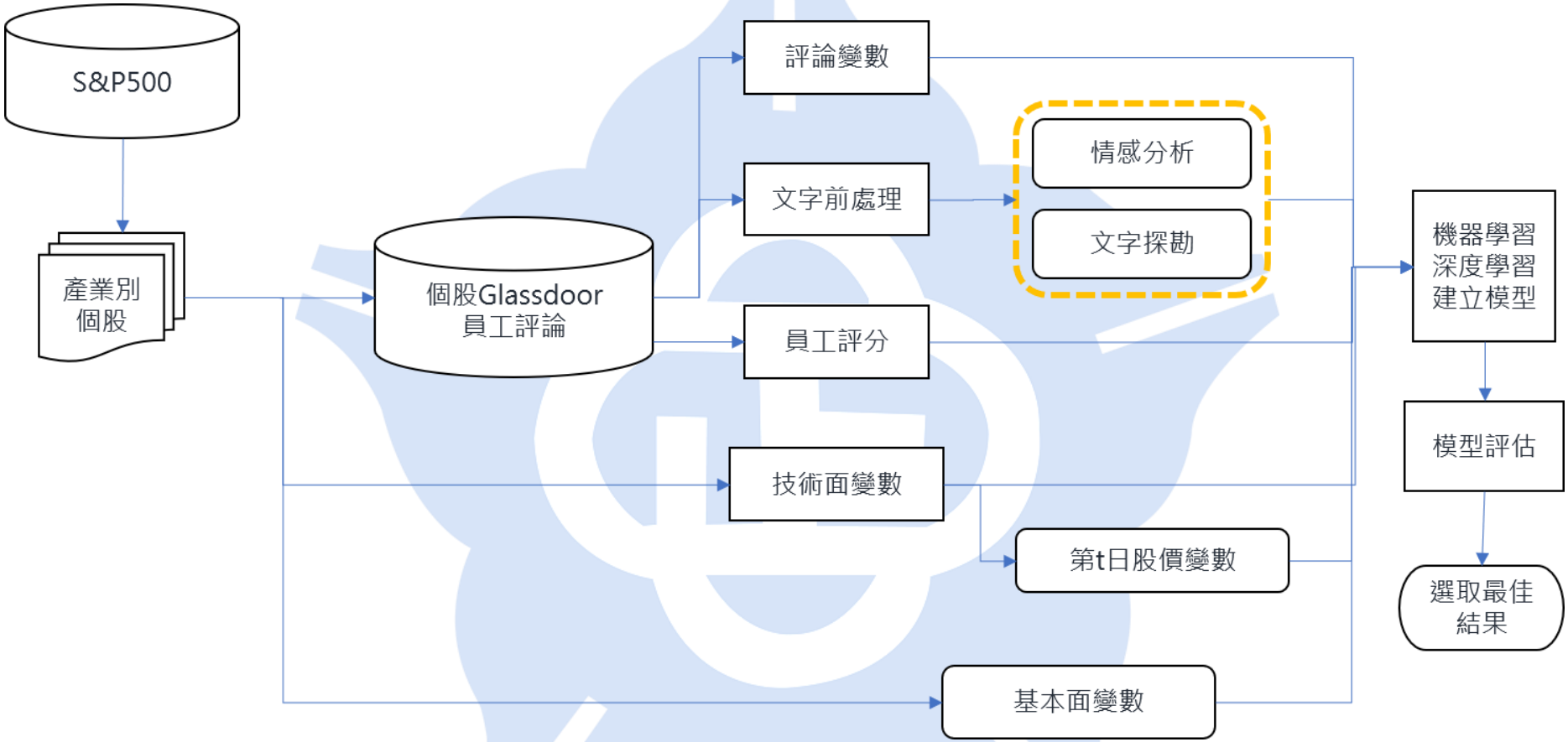
本研究首先將所有列於 S&P500 中的企業依照 GICS 產業別中的「科技」、「消費者非必須消費品」、「通訊服務」三大產業別中篩選後，再依據 S&P500 的權重值及 Glassdoor 評論數量排序，於三大產業中分別挑選一間 S&P500 權重值最大及 Glassdoor 評論數量最多之企業，另外會再選出三家 S&P500 權重值最小的企業，以作為大資本與小資本對照研究使用。接著使用 python 3 的 yfinance 套件抓取 YahooFinance!網站上提供之個股每日股價，以及在 Macrotrends 上的個股之基本面財務指標數值，結合 python 3 的 Selenium 套件建立爬蟲程式抓取 Glassdoor 網站上之個股公司員工評論、員工評分，資料時間區間收集 2014 年 1 月 1 日至 2021 年 10 月 15 日，作為本研究股價預測模型之訓練資料。

收集資料後將進行資料預處理，將 Glassdoor 評論等文字資料進行斷詞、斷句、移除特殊符號等文字前處理後，進行情感分析作為自變量，以及文字部分利用 TF-IDF 萃取出評論中的文字特徵，再結合 Glassdoor 的評論變數如員工評分、社群使用者按讚數，最後結合基本面、技術面之數值資料，接著使用 python 3 的套件 sklearn、xgboost、tensorflow-keras 進行訓練並建立股價預測模型，採用 XGBOOST、RF、RNN、LSTM，再從不同預測模型中評估比較，選出最佳預測模型進行研究探討。

本研究的研究架構如圖 5 所示，先從 Yahoo Finance!、Macrotrends、Glassdoor 蒐集資料，接著進行資料預處理，自變數涵蓋基本面、技術面、消息面，依變數為預測每日股價之一上漲預測二元分類器以及一下跌預測二元分類器，共兩個二元分類器預測模型，進而產生各資料集並建置實驗、訓練模型與評估，最後透過測試資料集進行回測，找出最佳模型。

圖 5

研究架構圖



資料來源:本研究整理

3.1 資料來源

S&P 500：Standard & Poor's 500(S&P 500)，標準普爾 500 指數，是由標準普爾道瓊指數(S&P Dow Jones Indices LLC)公司從 1957 年開始統計美國 500 支大型股票，其中包含科技、金融、能源等諸多產業。本研究將分產業別在 S&P500 中挑選個股，並採用 Yahoo Finance! API，抓取 2014 年 1 月 1 日至 2021 年 10 月 15 日的股價、技術面數據。

Macrotrends：Macrotrends 是一個提供美股上市企業歷史財務數字和圖形化統計資料的網站，並也提供使用者各式投資項目(如股票、基金、期貨等)的相關財務資訊。本研究將依照 S&P500 挑選後的個股，在 Macrotrends 上抓取 2013 年至 2021 年之個股公司的基本面資料。

Glassdoor：於 2007 年建立，提供線上員工評論的平台，員工可以在上面匿名發布自己對於公司的看法、評分等訊息。本研究將依照 S&P500 挑選後的個股，再利用網路爬蟲爬取個股公司於 2014 年 1 月 1 日至 2021 年 10 月 15 日的員工評論、員工評分等資料。

3.2 資料預處理

擴張 Dickey-Fuller 檢定

擴張 Dickey-Fuller 檢定(Augmented Dickey–Fuller [ADF])是用來檢驗時間序列資料是否存在單位根(Unit root)，其虛無假設為資料存在單位根，資料為非平穩性；對立假設則是資料不存在單位根，資料為平穩性。透過 ADF 檢定能檢驗時間序列資料是否為平穩性。收集完股價資料後，將進行 ADF 檢定，以確認股價資料是否有平穩性。如果資料呈現是平穩性，代表我們可以直接使用統計模型如 ARIMA 來建立預測模型；如果是資料呈現非平穩性，代表我們應使用機器學習與深度學習來建立預測模型會有較好的預測表現。

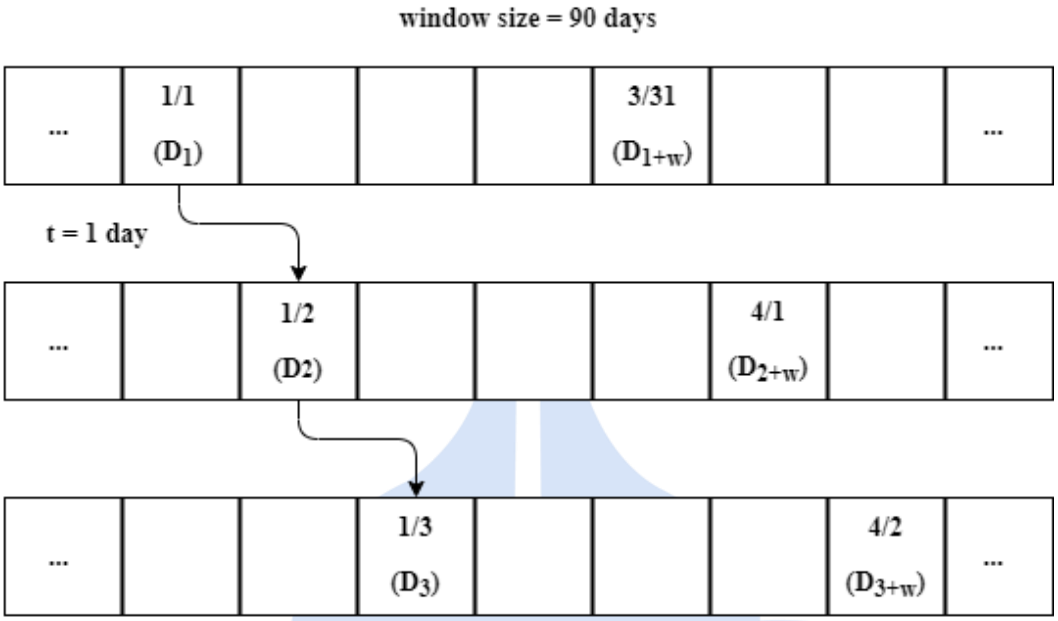
3.2.1 移動視窗

移動視窗(Rolling window)：移動視窗是一種處理時間序列資料的方式，首先設定移動視窗的大小 w($window_i$，i = 1, … ；)，w 是移動視窗選定的時間範圍，以日為單位；選定 w 後，接下來資料將依照移動視窗的大小來產生樣本，透過不斷重複的移動視窗間隔距離(t)，創造出許多時間區間。

以圖 6 為例，設定移動視窗的大小為 90 天，一次會移動 1 天(w = 90, t = 1)。

圖 6

移動視窗法操作過程



資料來源:本研究整理

3.2.2 文字預處理

在將文字資料丟進模型中進行訓練前，必須經過預處理，才能將文字轉換成模型能夠識別的型態。本研究的文字預處理包含以下四個步驟，並使用 python 套件 spacy 完成：

1. 斷句、斷詞(Tokenization)：要讓模型能夠學習文字類型的資料，首先必須將一個句子的每一個單字切分出來，讓模型能夠分別一個一個單字進行學習。

2. 去除停用詞、數字、特殊符號(Remove stop words, numbers and punctuations)：在斷詞後會發現，文字資料的數量非常多，而為了使資料中的單字量降低，本研究將一些停用詞去除，如 the、an 這些本身無特殊意義的單字；同時去除數字以及特殊符號如”*”、”%”等，以避免產生不必要的雜訊。

3. 大寫轉小寫(Capitalization)：Price 與 price 這兩個單字是相同的意思，但當模型在進行學習文字時，會誤以為是兩種不同的單字，因此為了避免上述情況，會將文本中所有的大寫字皆轉換成小寫字。

4. 詞幹還原(Lemmatization)：詞幹還原的意思是將某些經由時態、詞性變換後的單字轉換回單字原形，像是 sing、sang 與 sung 都是由 sing 轉換而成，而這些變化後的

單字跟原本的單字具有高度相似性。本研究使用 NLTK 的 WordNetLemmatizer 套件將所有的動詞做詞幹還原。

3.2.3 量化評論

在將文字變數資料經過預處理後，要結合股價資料時會發現一天有多則評論，因此將一整天的所有評論文字內容綜合起來，作為當天的代表性評論，然後再進行情感分析與 TF-IDF 計算。而評論部分也有自己的評分項目，因此也需要進行量化計算，計算方式為將該天的所有評論評分相加，再除以其數量，以得出當日平均評分。

3.2.4 標準化

在將數值資料放入深度學習模型前，必須先經過標準化(Normalization)，將資料縮放至[0,1]的區間，才能使模型進行較好的學習。學者 Singh and Singh (2020)提出不同的標準化都能有效增加模型學習的結果。本研究使用 Python3 的套件 sklearn.preprocessing 中的 MinMaxScaler 進行標準化，標準化的公式如下：

$$X_{scaled} = \frac{(X - X_{min})}{(X_{max} - X_{min})} \quad (1)$$

3.3 情感分析

情感分析(Sentiment Analysis)，是文字探勘領域的一種技術，用途是在一段文字中找出有情感意義的詞彙，並判斷一段文字表達的情感類別。這方面的分析是一個分類(Classification)問題，分為二元分類(Two-class Classification)及多元分類(Multi-class Classification)。二元分類就是將一句話中的情感簡單劃分成正面或是負面；而多元分類以英文為例，有 Sad、Fear、Anger，分別可以代表難過、恐懼、生氣等等，目前情感分析的分類方式可分為情感字典、機器學習兩種方式：

情感字典(Sentiment Dictionary)：在情感字典中，對於每個在字典中的單字會定義它的情感值分數(或稱極性)，單字的分數為浮點數值，其範圍從 1 到-1，分別代表正向情感(接近 1)或負向情感的程度(接近-1)。情感字典常見的字典有 Loughran and McDonald (2011)兩位學者提出的 Loughran and McDonald Dictionary(L&Mc)，以及 Hutto and Gilbert (2014)提出的 Valence Aware Dictionary and sEntiment Reasoner(Vader)。

L&Mc 是基於財務報表(10-K)中的單字而產生的情感字典，其中的單字大部分與金融領域相關，作者提出 L&Mc 以改善先前情感字典(哈佛字典)的單字量不足，導致先前的情感字典用於金融領域時常出現的情感錯誤分類。

Vader 則是基於社群媒體文本，經由 rule-based 而產生的情感字典，內容大部分以社群媒體上會出現的相關文字所組成，作者並於其研究中將 Vader 與真人標註、以及其他研究提出之情感字典做比較，發現 Vader 於社群媒體類型的文章中非常接近於真人標註的情感分數，作者並也有進行產品類型評論(Amazon 網站之產品評論)、電影評論、紐約時報評論等情感評分，Vader 也能於這些類型文章中，取得較以往之情感字典更佳的分數。

Vader 演算法對比先前的情感字典主要改良點在於，Vader 不只是紀錄單字的情感分數，也能夠解析(1)語句中的單字大小寫 (2)驚嘆號與問號等標點符號 (3)部分 Emoji(如「:)」表示笑臉) (4) 反向語句的情感解讀。

舉例來說，「Not bad at all」，Not、bad 兩個單字在 L&Mc 與 Vader 中都會被定義為負向的情感，而 at、all 兩單字在 L&Mc 中沒有定義其情感分數，因此該句子呈現出之情感為負面 2 分，但 Not bad 連在一起其語意表示為不差，而 Vader 中由於對於副詞、連接詞類型的單字也有加入字典，便能考慮到此種類型的單字組合，使該句子表達之情感分數能較正向，進而表達出文字中更廣泛的情感；另外，當文本中出現大寫單字、驚嘆號的時候，Vader 會加強該單字的情感分數，使整體文本能表現出更多情感。例如「Very nice!!!!!」於 L&Mc 及 Vader 上將會分別表示成「nice」、「Very nice!!!!!」，Very 以及最後的多個驚嘆號會提升 nice 本身的情感分數，其表現出來的文本情感分數會相差非常大，對比 L&Mc 可以看出 Vader 更能有效表現出文本的真實情感。

Kirlić and Orhan (2017)等人的研究也指出，Vader 標註的情感分數與真人標註之情感分數幾乎沒有差別。Li and Pan (2022)等人使用 Vader 來將華爾街日報(The Wall Street Journal)、路透社(Reuters)、消費者新聞和財經頻道(Consumer News and Business Channel)、財星(Fortune)四種不同新聞來源中的新聞標題做情感分數的標註，以作為股價預測中情感分析的分數。

機器學習(Machine Learning)：將大量單字、句子、文本預先標註好他們的情感類別，輸入到分類器、神經網路進行訓練，並建立預測模型來預測輸入的單字、句子、文本的情感類別。

本研究中的情感分析將利用情感字典中的 Vader 完成。

3.4 TF-IDF

TF-IDF(Term Frequency - Inverse Document Frequency [TF-IDF])，是一種常用於文字探勘領域的統計方法，用以評估單字對於文檔的重要程度，其中可以分為單字的頻率(Term Frequency [TF])與文檔的逆向頻率(Inverse Document Frequency [IDF])，TF 負責計算一個單字出現在一個文檔的頻率以表達一個單字對於該份文檔的表達能力，IDF 則負責計算一個單字出現在所有文檔中的頻率以表達該單字對文檔的區分能力，公式如下所述，

$$tfidf_{i,j} = tf_{i,j} \times idf_{i,j} \tag{2}$$

$$tf_{i,j} = \frac{f_{t,d}}{\sum_{t' \ni d} t', d} \tag{3}$$

$$idf_{i,j} = \log \frac{D}{f_t} \tag{4}$$

其中$tf_{i,j}$代表一個單字出現在一個文件的頻率，計算方式如公式 2，分子為該單字出現在某文檔中的次數，分母為某文檔(**d**)的總單字長度。$idf_{i,j}$代表一個單字出現在所有文檔中的頻率之倒數，分子(**D**)是總文件數，分母f_t是某個單字出現在幾個文件中，最後再取其倒數。

本研究採用 Python3 的 Scikit-learn 套件執行 TF-IDF 建立特徵詞向量，同時如果詞彙只有在少於 5%的文檔中出現(參數設置 min_df=0.05)，那麼就忽略不納入特徵詞向量中，最後將篩選出來的特徵詞向量建立成 TFIDF 矩陣。

3.5 自變數與依變數

本研究所使用的變數來自於基本面、技術面、消息面三大構面，結合每日股價之開盤價、最高價、最低價、收盤價四個自變數，以下小節將分開介紹自變數及依變數。

3.5.1 依變數

收盤價:

本研究之依變數為每日個股之收盤價之漲跌標籤，收盤價是指在當天交易結束前，最後一筆股票交易的成交價格。獲取收盤價之後，會分別建立兩個二元預測模型並將收盤價各自分為兩個類別：第一個預測模型之類別為「上漲」、「其他」，第二個預測模型之類別為「下跌」、「其他」，分類依據如下，t日收盤價為$\boldsymbol{Close_t}$，$\boldsymbol{t-1}$日收盤價為$\boldsymbol{Close_{t-1}}$：

$$\boldsymbol{P_t} = \frac{\boldsymbol{Close_t - Close_{t-1}}}{\boldsymbol{Close_{t-1}}} \times \boldsymbol{100\%} \quad (2)$$

第一個預測模型中，當$\boldsymbol{P_t}$大於1%，也就是今日的收盤價與昨日收盤價相比增加了1%以上，就將資料記為「上漲」，否則為「其他」；第二個預測模型中，$\boldsymbol{P_t}$小於-1%，則標記為「下跌」，否則為「其他」。

3.5.2 基本面自變數

本研究使用之基本面自變數將由 Macrotrends 網站上抓取每季之財務報表數值，再進行計算獲得。表 7 呈現實驗中將加入之基本面變數。

1. 本益比(PE Ratio [P/E]):

本益比是指投資人為了獲得一元的利潤，所需要投資的金額，投資人常以此比率來衡量股票是不是值得投資，代表的是目前的收盤價與每股盈餘(Earning Per Share [EPS])之間的關係，一般來說本益比的值約落在 0~20 之間。其中 $\boldsymbol{t}$ 為當天時間。

每股盈餘是衡量公司營利的重要指標，此財務比率代表的是在外流通之總股數(Shares)與每季或每年之稅後淨利(Net Income)間的關係，也可以直接解釋為每股可以賺到多少淨利，其中稅後淨利代表每季或每年之營收扣除銷貨費用、營業費用等損失後得出之金額。

$$EPS = \frac{Net\ Income}{Shares} \tag{3}$$

$$P/E = \frac{Close_t}{EPS} \tag{4}$$

2. 股價淨值比(Price Book Ratio [PBR]):

股價淨值比是指股價相對於每股淨值(Book value Per Share [BPS])的比例，表示公司目前的收盤價是淨值的幾倍，反應出公司股價的帳面價值相對股價的合理性，一般而言認為股價淨值比大於 1，表示股價較高，潛在報酬較低；股價淨值比小於 1 則是股價較低，潛在報酬較高。其中 t 為當天時間。

$$BPS = \frac{Assets - Liabilities}{Shares} \tag{5}$$

$$PBR = \frac{Close_t}{BPS} \tag{6}$$

3. 市值(Market Capitalization [CAP]):

市值即是指上市公司的在股票市場上的市場價格總值，利用收盤價乘以在外流通股數來計算，金融業常以此來衡量一間公司的規模，(Market capitalization: Pre and post COVID-19 analysis)提到，市值讓投資人能了解公司的價值及其前景，以及投資與否之決策。

$$CAP = Close_t \times Shares \tag{7}$$

4. 股票周轉率(Shares Turnover [ST]):

股票周轉率代表的是交易量與在外流通股數之間的關係，當周轉率越高就代表當日之交易量越大，屬於熱門的股票；反之則為較冷門之股票。

$$ST = \frac{Volume_t}{Shares} \tag{8}$$

表 7

基本面變數表

變數	型態
PER	數值
PBR	數值
CAP	數值
ST	數值(0~1)

資料來源:本研究整理

3.5.3 技術面自變數

以下技術面指標將利用 Python 3 套件 pandas_ta 來產生每日技術面指標之數值，表 8 呈現實驗中將加入之技術面變數。

1. 移動平均線(MA):

移動平均線又可以稱為簡單移動平均線(Simple Moving Average [SMA])，是一種算術平均線，代表過去一段時間內的平均成交價格，把一段時間內的價格相加，再除以週期頻率，本研究使用 5、15、30 日的收盤價作為 TMA 之計算。

$$MA(n)_t = \frac{1}{n}\sum_{i=t-n+1}^{t} Close_i \tag{14}$$

其中n為要計算 n 日內的 MA，t為當天時間，$Close_i$為第i日之收盤價。

2. 指數移動平均線(EMA):

指數移動平均線(Exponential Moving Average [EMA])，與移動平均線類似，都是用一段時間內的資料，來代表平均成交價格。但不同的地方在於，指數移動平均線會將時間越近的資料乘上更大的權重(α)，以此來表示越近的資料其影響力越大。至於時間

較舊的資料，其權重為每日遞減，資料時間越舊權重則越小。本研究使用 12 日與 26 日的收盤價作為 EMA 之計算。

$$\boldsymbol{\alpha = \frac{2}{N+1}}\ \boldsymbol{，EMA(n)_t = EMA(n)_{t-1} + \alpha \times (Close_i - EMA(n)_{t-1})} \qquad (15)$$

其中$\boldsymbol{n}$為移動時間長度，$\boldsymbol{t}$為當天時間，$\boldsymbol{Close_i}$為第$\boldsymbol{i}$日之收盤價，$\boldsymbol{\alpha}$為權重，$\boldsymbol{EMA(n)_t}$為第 $\boldsymbol{t}$ 日的指數移動平均線值。

3. 三角移動平均線(TMA):

三角移動平均線(Triangular Moving Average [TMA])，是對移動平均線做二次平均，使其更加平滑，也就是說，TMA 就是 MA 的平均價格。本研究使用 10 日的 MA 作為 TMA 之計算。

$$\boldsymbol{TMA(n)_t = \frac{1}{n} \sum_{i=t-n+1}^{t} MA(n)_t} \qquad (16)$$

其中$\boldsymbol{n}$為要計算 $\boldsymbol{n}$ 日內的 $\boldsymbol{TMA}$，$\boldsymbol{t}$為當天時間，$\boldsymbol{MA(n)_t}$為第 $\boldsymbol{t}$ 日的移動平均線值。

4. 指數平滑異同移動平均線(MACD):

指數平滑異同移動平均線(Moving Average Convergence / Divergence [MACD])是計算 EMA 之間的離差程度，其中有分快線(DIF)、慢線(DEA)、MACD 線，快線在投資界常用的計算方式是計算 12 天的 EMA 與 26 天的 EMA 差值；慢線則是利用快線在 9 天內的值，再進行 EMA 計算；MACD 線就是使用快線值減去慢線值。

$$\boldsymbol{DIF_t = EMA(12)_t - EMA(26)_t} \qquad (17)$$

$$\boldsymbol{DEA_t = EMA[DIF(9)_t]} \qquad (18)$$

$$\boldsymbol{MACD_t = DIF_t - DEA_t} \qquad (19)$$

其中$\boldsymbol{t}$為當天時間，$\boldsymbol{DIF_t}$是 12 天的 EMA 減去 26 天的 EMA 差值，$\boldsymbol{DEA_t}$則是將 9 天的快線值再利用 EMA 計算，$\boldsymbol{MACD_t}$為 $\boldsymbol{t}$ 日的 DIF 減去 DEA。

5. 相對強弱指標(RSI):

相對強弱指標(Relative Strength Index [RSI])是在技術分析中常用的指標，它考慮了股價的漲跌天數、漲跌幅度，因此可以用來測量價格動向的快慢和變化，本研究使用 14 天收盤價進行 RSI 計算。

$$\boldsymbol{RS(n)_t = \frac{\sum_{i=t-n+1}^{t} Up}{\sum_{i=-n+1}^{t} Down}} \text{ ，} \boldsymbol{RSI(n)_t = 100 \times \frac{RS}{1+RS}} \quad (20)$$

其中$\boldsymbol{t}$為當天時間，Up、Down 表示當天收盤價為上漲或下跌，$\boldsymbol{n}$ 為要計算 $\boldsymbol{n}$ 日內的上漲、下跌。

6. 動量指標(Momentum):

動量指標(Momentum [MOM])是衡量價格走勢變動量的指標，它考慮了最近的收盤價與 N 天前的收盤價，來看今天與 N 天前的收盤價之間的差距。投資界常用的動量指標有 5 天、10 天等，本研究使用 10 天收盤價進行 MOM 計算。

$$\boldsymbol{MOM(n)_t = Close_t - Close_{t-n}} \quad (21)$$

其中$\boldsymbol{t}$為當天時間，$\boldsymbol{n}$ 為要計算 $\boldsymbol{n}$ 日內的動量指標。

7. 布林通道(BB):

布林通道(Bollinger Bands [BB])是利用 MA 與標準差的概念，來畫出數條能分析市場股價的最高值最低值的線，其中有布林通道最高線(BBU)、布林通道最低線(BBL)、布林通道中心線(BB)、布林通道平均線(BBP)、布林通道寬度線(BBB)，並以這些線來看出股價變動的幅度，首先求出中心線，就可以利用中心線推出其他最高最低線的值，常用的布林通道 MA 為 20 天，本研究使用 20 天 MA 進行布林通道計算。

$$\boldsymbol{SD = \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{x})} \quad (22)$$

$$\boldsymbol{BB_t = MA(20)_t} \quad (23)$$

$$\boldsymbol{BBU_t = BB_t + 2 \times (SD)} \quad (24)$$

$$\boldsymbol{BBL_t = BB_t - 2 \times (SD)} \tag{25}$$

$$\boldsymbol{BBP_t = (Close_t - BBL_t) \div (BBU_t - BBL_t)} \tag{26}$$

$$\boldsymbol{BBB_t = 100 \times (BBU_t - BBL_t) \div BB_t} \tag{27}$$

其中$\boldsymbol{N}$為資料數量，$\boldsymbol{X_i}$為第$\boldsymbol{i}$個資料，$\boldsymbol{\bar{x}}$是資料的平均值，$\boldsymbol{t}$為當天時間。

8. 威廉指標(William's %R):

威廉指標(William's %R)是用今天的收盤價，以及一段時間內股價的最高價(High price)及最低價，來衡量這段時間內的股票市場是否有超買或超賣的現象，為提供投資人市場趨勢的反轉指標，常用的威廉指標區間為 14 天。

$$\boldsymbol{William(n)_t = \frac{(High\ price(n)_t - Close_t)}{(High\ price(n)_t - Low\ price(n)_t)}} \tag{28}$$

其中$\boldsymbol{t}$為當天時間，$\boldsymbol{n}$ 為要計算 $\boldsymbol{n}$ 日內的威廉指標、最高價、最低價。

9. KD 隨機指標(Stochastic Oscillator):

KD 隨機指標(Stochastic Oscillator [OSC])為動量指標的另一種表現方式，它考慮了一定期間的最高價與最低價為基準，來判斷收盤價的水準，其中會再分為%K 以及 Slow %D 線。

$$\boldsymbol{STOCH = 100 \times \frac{Close_t - LL}{HH - LL}} \tag{29}$$

$$\boldsymbol{\%K_t = MA(3)_t} \tag{30}$$

$$\boldsymbol{Slow\%D_t = MA(3)_t} \tag{31}$$

其中$\boldsymbol{t}$為當天時間，$\boldsymbol{n}$ 為要計算 $\boldsymbol{n}$ 日內的收盤價，$\boldsymbol{LL}$表示過去 n 日內的最低價，$\boldsymbol{HH}$表示過去 n 日內的最高價，%$\boldsymbol{K}$的 MA 是利用計算出的$\boldsymbol{STOCH}$數值得出，而非其他技術指標使用之收盤價，而$\boldsymbol{Slow\%D}$之 MA 是利用計算出的%$\boldsymbol{K}$數值得出，而非使用$\boldsymbol{STOCH}$計算。

10. 能量潮(OBV):

能量潮(On-Balanced Volume [OBV])表達了交易量(Volumes [V])的重要性，並考慮了今日與昨日的收盤價，如果今日收盤價大於昨日收盤價，則能量潮會加上交易量；如果今日收盤價小於昨日收盤價，則能量潮會減去交易量。

$$\boldsymbol{OBV_t = OBV_{t-1} + V_t \quad IF\ Close_t > Close_{t-1}} \tag{30}$$

$$\boldsymbol{OBV_t = OBV_{t-1} - V_t \quad IF\ Close_t < Close_{t-1}} \tag{31}$$

其中$\boldsymbol{t}$為當天時間。

表 8

技術面變數表

變數	型態
MA_5	數值
MA_15	數值
MA_30	數值
EMA_12	數值
EMA_26	數值
TMA	數值
DIF	數值
DEA	數值
MACD	數值
RSI	數值
MOM	數值
BB	數值
BBU	數值
BBL	數值
BBP	數值

表 8

技術面變數表

變數	型態
BBB	數值
William’s %R	數值
%K	數值
Slow%D	數值
OBV	數值

資料來源:本研究整理

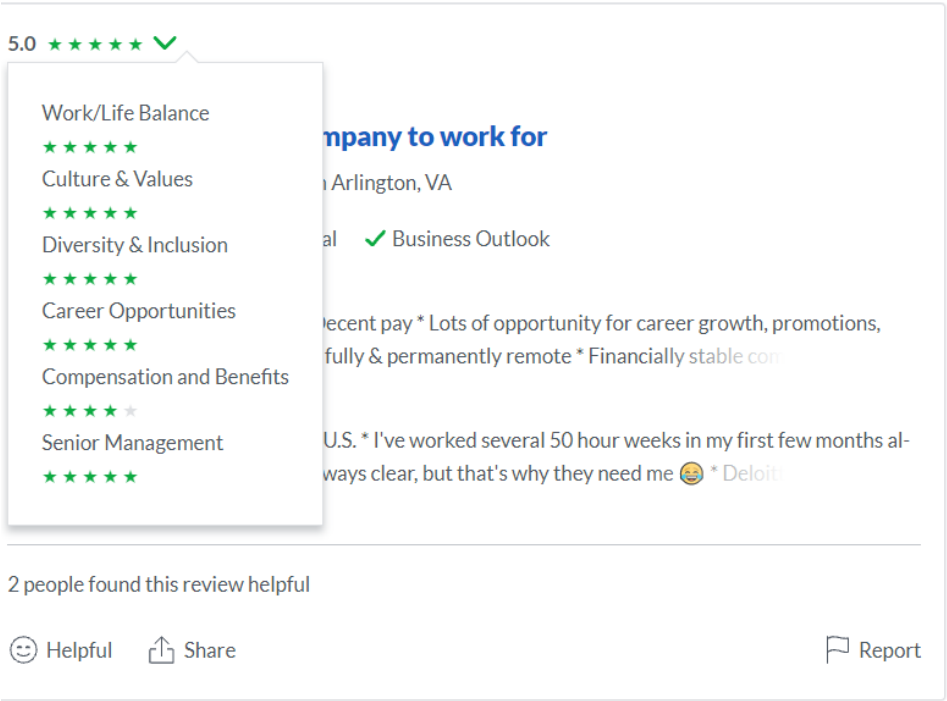
3.5.4 消息面自變數

Glassdoor 員工評論

根據 Glassdoor 上的員工評論資訊如圖 7 及圖 8，主要可以由員工評分、員工評論兩大部分組成，本研究蒐集的員工評分項目包含評論總分、工作生活平衡分數等十大項目評分，如表 9 所示。

圖 7

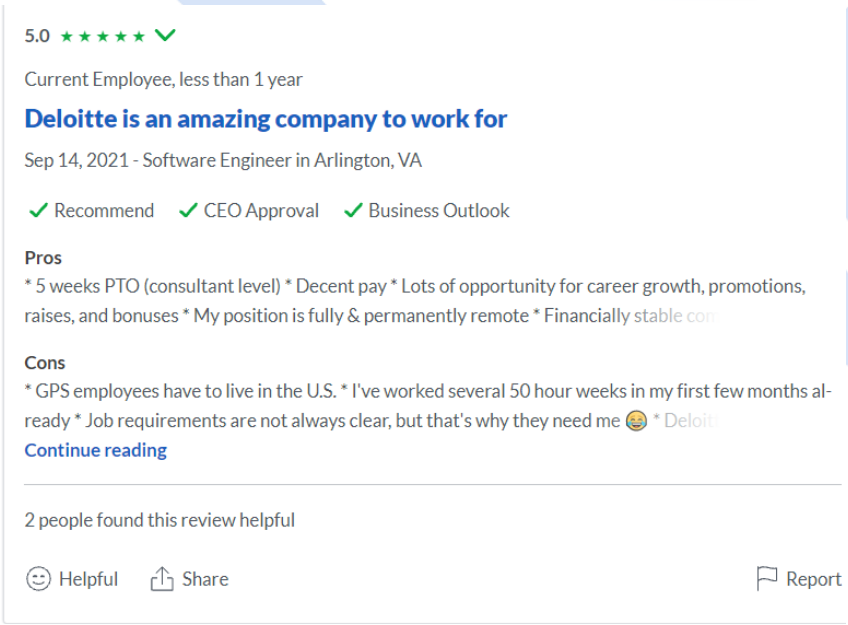
Glassdoor 評論資訊圖



資料來源: Anonymous. (2021, Sep 14). *Deloitte is an amazing company to work for*. Glassdoor. https://www.glassdoor.com/Reviews/Deloitte-Reviews-E2763.htm.

圖 8

Glassdoor 評論資訊圖



資料來源: Anonymous. (2021, Sep 14). *Deloitte is an amazing company to work for*. Glassdoor. https://www.glassdoor.com/Reviews/Deloitte-Reviews-E2763.htm.

表 9

Glassdoor 評分變數表

變數	定義	型態
Overall_Rating	員工綜合評分	數值 (1~5)
Work-Life Balance	員工對該公司工作生活平衡之認同程度	數值 (1~5)
Culture & Values	員工對該公司企業文化及價值觀之認同程度	數值 (1~5)
Diversity & Inclusion	員工對該公司多元性別、不同文化包容之認同程度	數值 (1~5)
Career Opportunities	員工對該公司員工職業生涯發展之認同程度	數值 (1~5)
Compensation & Benefits	員工對該公司員工薪資及福利的給予之認同程度	數值 (1~5)
Senior Management	員工對該公司高階管理人員的管理方式之認同程度	數值 (1~5)
Recommendation	員工推薦這間公司的程度	數值 (0~3)
CEO Approval	員工對該公司目前 CEO 及其帶領方式之認同程度	數值 (0~3)
Business Outlook	員工對於該公司的未來前景之看好程度	數值 (0~3)

資料來源:本研究整理

本研究蒐集的員工評論項目則包含評論日期，以及評論標題、優點缺點等文字經情感分析後產生之情感分數，以及優缺點透過 TF-IDF 計算之分數，如表 10 所示。

表 10

Glassdoor 評論變數表

變數	定義	型態
Review_Date	評論日期	日期
Review Title_Sentiment	員工評論標題之情感分數	數值(-1~1)
Pros_Sentiment	員工評論優點之情感分數	數值(-1~1)
Cons_Sentiment	員工評論缺點之情感分數	數值(-1~1)
Pros_TF-IDF	員工評論優點之 TF-IDF 分數	數值(0~1)
Cons_TF-IDF	員工評論缺點之 TF-IDF 分數	數值(0~1)

資料來源:本研究整理

3.6 資料探勘分析相關技術

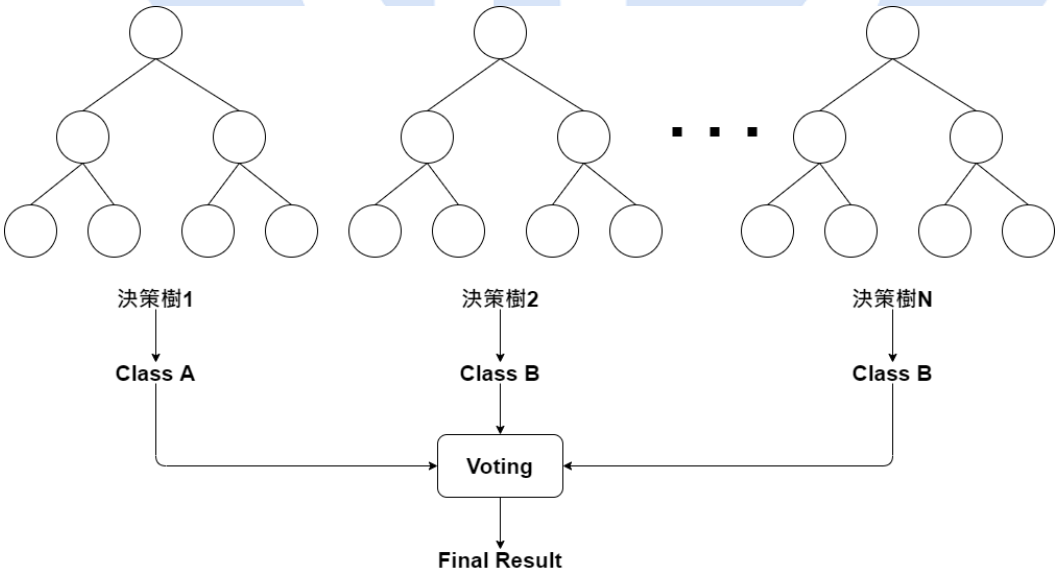
本研究採用 XGBOOST、RF、RNN、LSTM 等四種演算法，並將以 sklearn、xgboost 以及 TensorFlow 的 keras 套件實作四種演算法模型，下列將個別說明四種演算法：

3.6.1 隨機森林（RF）

隨機森林是一種集合學習(Ensemble Learning)的方法，由 Ho (1995)提出，她利用決策樹演算法的概念加以改進。隨機森林演算法首先對於資料以隨機取樣的方式選取樣本及特徵，建立出許多不同的決策樹，這些決策樹分別就代表了不一樣的特徵及樣本，而單一顆決策樹的預測容易有過度擬合(Over-Fitting)的問題(Bramer, 2007)，因此將眾多決策樹聚集起來，稱為「隨機森林」，接下來統計所有決策樹對於一筆資料的預測類別後，以所有決策樹分類之眾數來決定這筆資料的最終分類類別，它的好處是透過集合許多決策樹後進行投票，能讓模型不容易發生過度擬合的問題。

圖 9

隨機森林



資料來源:本研究整理

3.6.2 極限梯度提升（XGBOOST）

eXtreme Gradient Boosting，又稱 XGBoost，由 Chen and Guestrin (2016)提出，它以梯度提升決策樹(Gradient Boosted Decision Tree, GBDT)為基礎並加以改良。由於該演算法利用貪心法的概念，在每次樹的每層建構過程中嘗試優化目標函式的最大增益(Gini)，故該演算法會針對每一棵樹之中的每個葉子結果去計算分數，接下來會透過增量訓練(Additive Training)的方式以尋找並定義最佳的目標函式，每一次保留上一次構建成的模型不變，並且加入一個新的函數至模型中，換言之每次建立模型會於上次的樹中增加一顆樹，以修復上一顆樹的不足，有助於提升目標函數，再對誤差函式(Loss function)加入懲罰項以避免 Overfitting，而其懲罰項即是套用 Ridge Regression 與 Least Absolute Shrinkage and Selection Operator (LASSO)來將每個錯誤的葉子結果去做懲罰，再根據上述的規則並重新定義每棵樹之分數。

3.6.3 循環神經網路（RNN）

循環神經網路的雛形是由 Hopfield (1982)所提出，傳統的神經網路架構是由輸入層、隱藏層、輸出層所組成。而 RNN 在隱藏層上加入了「儲存單元」，它透過將前一次訓練的輸出結果之權重矩陣(Weight Matrix)或是前一次訓練的隱藏層的權重矩陣記憶起來，並加入下一次訓練的隱藏層中一起訓練。目前 RNN 主流為兩種網路架構，一種是學者 Jordan (1997)提出的 Jordan Network，會儲存每次訓練的輸出結果；另一種則是學者 Elman (1990)提出的 Elman Network，會儲存前一次隱藏層的參數向量。而 Keras 的 RNN 實現方式是 Jordan Network，也就是說，目前輸出的值不但受到這一次輸入的值影響，也會受到前一次訓練的結果影響。透過這種方式，達到下一次訓練的資料能夠參考前幾次訓練的結果。RNN 公式如下：

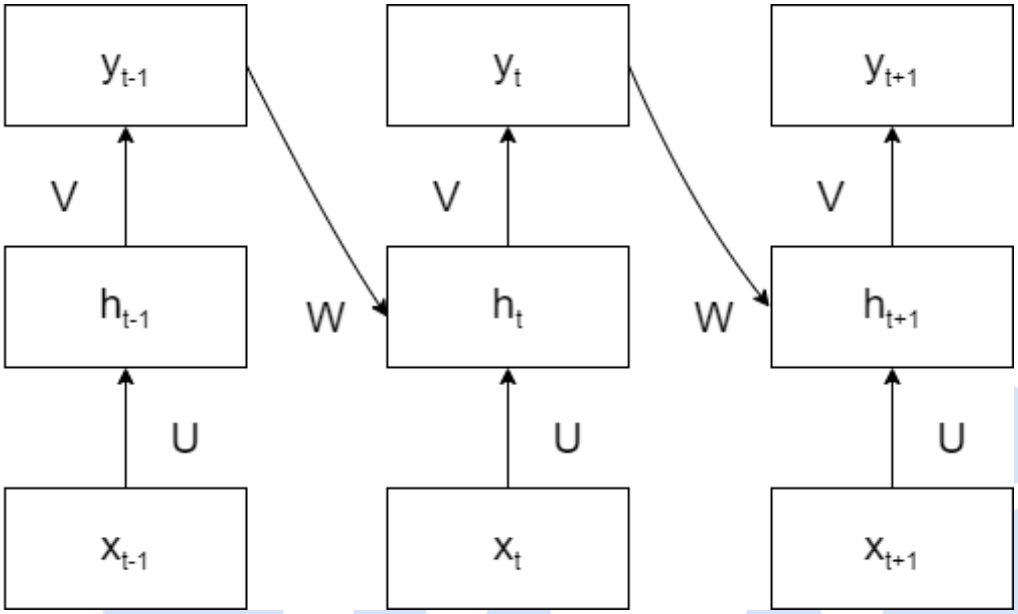
$$\boldsymbol{Y_t} = \boldsymbol{g}(\boldsymbol{V} \times \boldsymbol{h_t} + \boldsymbol{c}) \tag{32}$$

$$\boldsymbol{h_t} = \boldsymbol{f}(\boldsymbol{U} \times \boldsymbol{X_t} + \boldsymbol{W} \times \boldsymbol{Y_{t-1}} + \boldsymbol{b}) \tag{33}$$

其中 t 為當時的時間，Y_t 是 t 時間的輸出向量，X_t 是 t 時間的輸入向量，h_t 是 t 時間的隱藏層，U、W、V 為權重矩陣，b、c 為偏差值(bias)，g、f 為激活函數(Activation Functions)。

圖 10

循環神經網路



資料來源:本研究整理

3.6.4 長短期記憶神經網路（LSTM）

隨著訓練時間變長，以 RNN 訓練資料會產生梯度消失的問題，導致循環神經網路將無法處理長期的時間關聯。因此，Hochreiter and Schmidhuber (1997)根據循環神經網路改良而提出了 LSTM，不僅改善了梯度消失問題，也將儲存單元改進成「記憶單元」。LSTM 將記憶單元改成由輸入門(Input gate)、遺忘門(Forgotten gate)、神經元(Neurons)、輸出門(Output gate)四種結構組成，透過這幾個門來決定要儲存、遺忘資訊以及何時需要儲存、遺忘資訊。

$$i_t = \sigma_i(W_i X_t + U_i c_{t-1} + b_i) \tag{34}$$

$$f_t = \sigma_f(W_f X_t + U_f c_{t-1} + b_f) \tag{35}$$

$$o_t = \sigma_o(W_o X_t + U_o c_{t-1} + b_o) \tag{36}$$

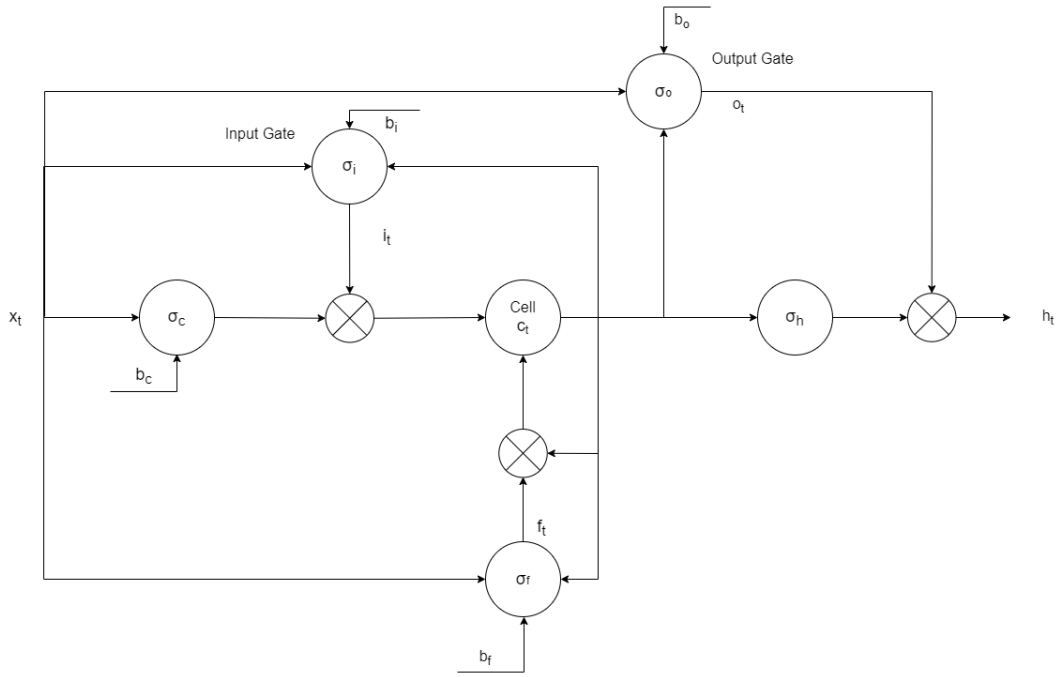
$$c_t = f_t \times c_{t-1} + i_t \times \sigma_c(W_c X_t + b_c) \quad (37)$$

$$h_t = o_t \times \sigma_h(c_t) \quad (38)$$

其中 t 為當時的時間，X_t是 t 時間的輸入向量，i_t是 t 時間的輸入門向量，h_t是 t 時間的輸出向量，f_t是 t 時間的遺忘門向量，c_t儲存了神經元，U、W 為權重矩陣，b 為偏差值，σ為激活函數。

圖 11

長短期記憶神經網路



資料來源:本研究整理

3.7 實驗建構

本研究嘗試使用 Python 3 套件 Scikit-Learn、xgboost 套件建立機器學習預測模型，神經網路部分使用 TensorFlow 及 keras 建立模型並進行結果評估以及比較各模型之表現，預測模型將使用 XGB、RF、RNN、LSTM 四種演算法，主要可分成兩大類：(1)機器學習：XGB、RF (2)深度學習：RNN、LSTM。透過以上演算法建立股價預測漲跌預測模型。實驗建構共分為三組，分別為不同公司資料集透過不同演算法以及不同的特徵變數所建立之預測模型，並選取出最佳預測模型作為研究探討。另外考量到由於資料集時間包含嚴重特殊傳染性肺炎 COVID-19（Coronavirus Disease-2019），因此先將資料集切分成四種資料集並於實驗一中進行實驗，並於圖 12 中顯示：

(1)NoCovid 資料集：資料時間範圍取為 2014/01/01~2019/12/31，其中訓練集為 2014/01/01~2018/12/31，測試集為 2019/01/01~2019/12/31。

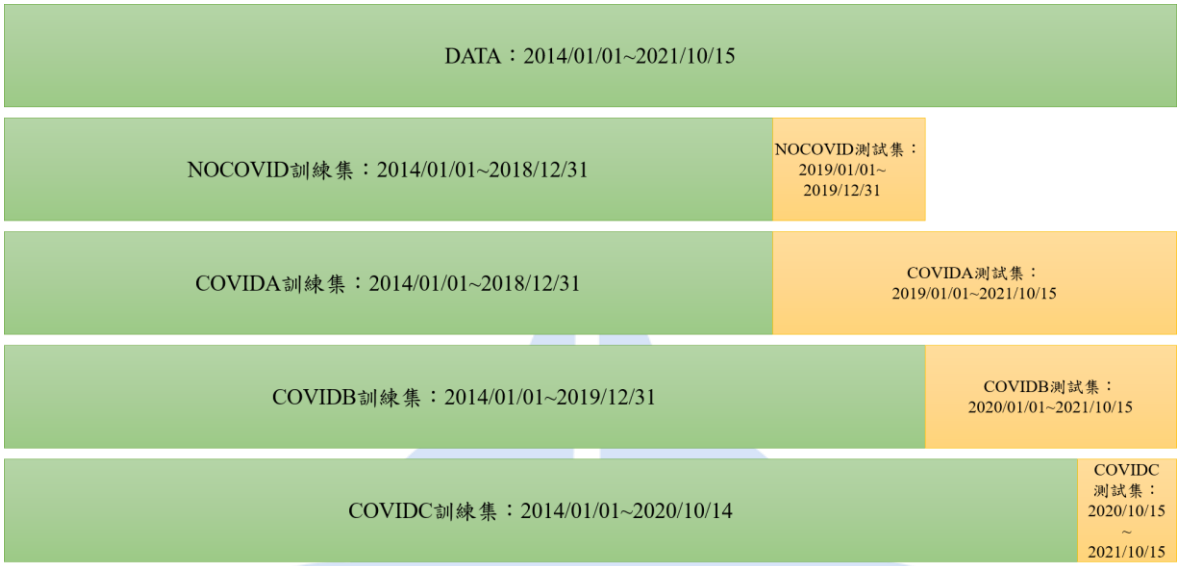
(2)Covid_A 資料集：資料時間範圍取為 2014/01/01~2021/10/15，其中訓練集為 2014/01/01~2018/12/31，測試集為 2019/01/01~2021/10/15。

(3)Covid_B 資料集：資料時間範圍取為 2014/01/01~2021/10/15，其中訓練集為 2014/01/01~2019/12/31，測試集為 2020/01/01~2021/10/15。

(4)Covid_C 資料集：資料時間範圍取為 2014/01/01~2021/10/15，其中訓練集為 2014/01/01~2020/10/14，測試集為 2020/10/15~2021/10/15。

圖 12

資料集切分示意圖



資料來源:本研究整理

透過實驗一產生之最佳資料集將用於實驗二中建立預測模型，此外本研究將使用移動視窗法產生時間序列的資料，並將分為二組實驗設計：

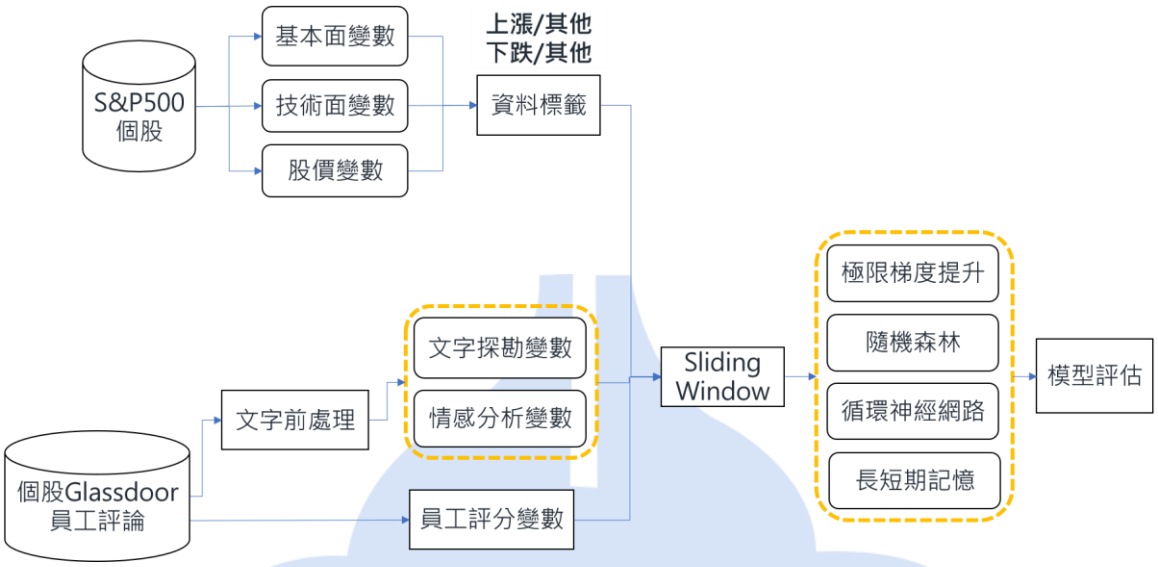
(1)實驗一(基本面、技術面)：透過使用基本面指標、技術面指標結合每日股價之開高收低資料，單純使用過去相關研究使用之指標，作為股價預測模型資料，此為本研究比較之基準線(baseline)，並探討最佳移動視窗之大小與最佳資料集，實驗一架構如圖 14 所示。

(2)實驗二(基本面、技術面、Glassdoor 評分、Glassdoor 評論變數)：使用實驗一之最佳資料集，變數涵蓋使用基本面指標、技術面指標、每日股價之開高收低，並加入 Glassdoor 員工評論評分、員工評論變數，實驗二架構如圖 15 所示。

研究最終將以機器學習與深度學習共 4 種模型執行結果來進行效能評估以及探討研究結果，最後選取出最佳的預測模型作為研究探討，實驗設計總架構如圖 13 所示。

圖 13

實驗設計架構圖



資料來源:本研究整理

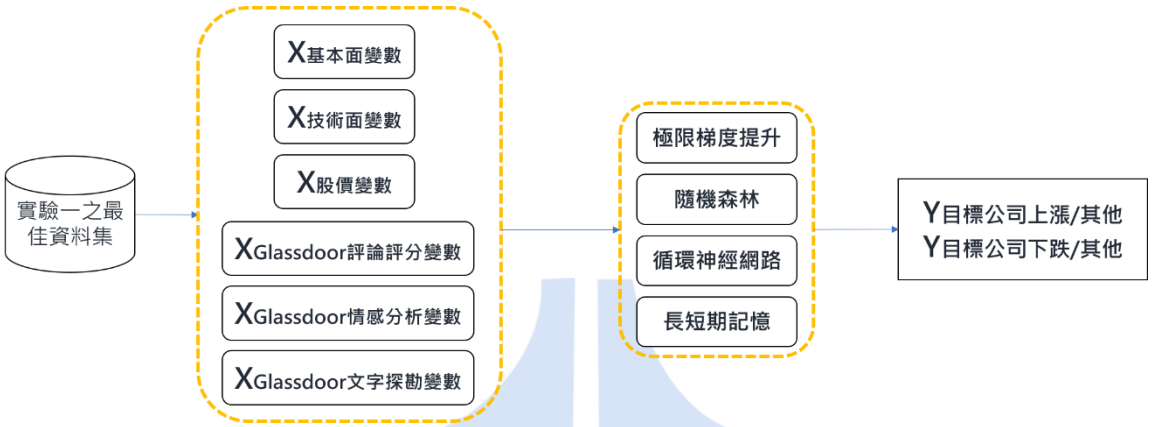
圖 14

實驗一：股價相關變數預測模型

資料來源:本研究整理

圖 15

實驗二：員工評論股價預測模型



資料來源:本研究整理

3.8 評估指標

由實驗過程產生的預測模型輸出值會建立兩個預測模型，上漲與下跌，其中上漲及下跌各自又分為兩種類別，上漲預測模型為(1)上漲；(2)其他。下跌預測模型則為(1)下跌；(2)其他，皆屬於類別型態。本研究屬於二元類別的分類預測問題，模型評估部分採用混淆矩陣(Confusion Matrix)，從中產生精確度(Precision)、召回率(Recall)、F1-Score(綜合考量 Precision 及 Recall)共三種指標作為預測模型之效能評估。如下表 11 以上漲類別之預測模型解釋。

表 11

混淆矩陣

		實際	
		股價上漲	其他
預測	股價上漲	***TP***	***FP***
	其他	***FN***	***TN***

下面公式將以上漲資料舉例，如何以上述之混淆矩陣計算 Precision、Recall、F1-Score。

- 精準度(Precision)：

表示在預測為上漲的情形時，有多少比例的資料是正確判斷的，值的範圍為 0~1 之間，越接近 1 表示模型預測精準度越高，上漲類別的 Precision 如以下公式。

$$\boldsymbol{Precision} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FP}} \tag{40}$$

- 召回率(Recall)：

表示在答案為上漲的情形時，有多少比例的資料是正確判斷的，值的範圍為 0~1 之間，越接近 1 表示模型預測召回率越高，上漲類別的 Recall 如以下公式。

$$\boldsymbol{Recall} = \frac{\boldsymbol{TP}}{\boldsymbol{TP} + \boldsymbol{FN}} \tag{41}$$

- F1-Score：

表示 Precision 及 Recall 之調和平均數，值的範圍為 0~1 之間，越接近 1 表示模型 F1-Score 越高，上漲類別的 F1-Score 如以下公式。

$$\boldsymbol{F1-Score} = \mathbf{2} \times \frac{\boldsymbol{Precision} \times \boldsymbol{Recall}}{\boldsymbol{Precision} + \boldsymbol{Recall}} \tag{42}$$

第四章、 實驗設計與評估

4.1 資料集描述

在經由收集整理資料後，最後選取出的 S&P500 權重大之個股為 GOOG、AAPL、AMZN，分別代表通訊服務、科技、消費者非必需品類別，權重小之個股為 LUMN、NLSN、GPS，分別代表通訊服務、科技、消費者非必需品類別。

而本論文使用之資料集可分為結構化與非結構化，詳細描述如下：

(一) 結構化資料集：

結構化資料集包含股價變數、基本面變數、技術面變數，其中股價變數來源為 Yahoo Finance 網站提供之歷史每日股價，利用 python 的 yfinance 套件存取 API 取得資料，資料收集時間為 2014 年 1 月 1 日起至 2021 年 10 月 15 日；基本面變數來源為 Macrotrends 網站，透過手動搜尋並整理 Macrotrends 網站上提供之各公司財報數值 (EPS、權益總額)，另外基本面變數中，每股盈餘、每股淨值等變數為財務報表所產生，它以季為時間單位進行發佈，根據 Marcotrends 網站上的資料顯示，大部分公司的財務指標變動日期為 12/31、3/31、6/30、9/30，因此本研究將財務指標加入時，皆會以這四個日期之財務指標數值為基準，再往後之日期補上相同的財務指標數值，直到下一個財務報表時的日期公布則補上新的財務指標，舉例為：3/31 時 AAPL 的 PER 欄位為新的數值，4/01 時則沿用相同數值，直到 6/30 時 PER 欄位數值為新發佈的數值；技術面變數則利用 pandas_ta 套件計算。收集完成後資料筆數為 1962 筆。

(二) 非結構化資料集：

非結構化資料集為 Glassdoor，資料收集時間為 2014 年 1 月 1 日起至 2021 年 10 月 15 日止，各公司(GOOG、AAPL、AMAZ、LUMN、NLSN、GPS)之 Glassdoor 評論總筆數分別為 18460、22770、100687、5000、5460、7140 筆員工評論。在收集完員工後，由於要量化為每天交易資料的呈現(每日情感分數)，因此會將同一天的評論結合再進行情感分數計算，最後整理完的評論有可能不會與股價資料為同一天(評論

有可能發佈在週末，而股價只發佈在週一至週五)，產生遺漏值的部分將由前一天的評論補上。(如 2021/04/29(四)當天剛好沒有 Glassdoor 評論，使用 2021/04/28(三)的情感分數補上)。

而在收集完資料後，首先要進行 ADF 檢定以確定各公司之股價資料是否具有平穩性，利用 python 套件 statsmodels 中的 adfuller 進行 ADF 計算，表 12 為各公司之收盤價資料以及其 ADF 檢定值與 t 值之比較。

各公司之 ADF 檢定假設如下:

H_0: 公司收盤價資料存在單位根

H_1: 公司收盤價資料不存在單位根

表 12

ADF 檢定

公司	ADF 檢定值	α = 0.01	α = 0.05	α = 0.1
GOOG	3.266	-3.433	-2.863	-2.567
AAPL	1.710			
AMAZ	0.985			
LUMN	-1.625			
NLSN	-0.727			
GPS	-2.040			

資料來源:本研究整理

根據 ADF 計算結果顯示，各公司之 ADF 檢定值都大於顯著水準 1%、5%及 10%之 t 值，因此沒有足夠證據證明且不拒絕 ADF 的虛無假設(所有資料都存在單位根)，亦即推斷所有公司之收盤價資料皆為非平穩性。

4.2 資料結果與評估

首先將所有資料集的時間範圍取為 2014/01/01~2021/10/15，因此資料集共會有 1962 筆資料，另外考量到由於資料集時間包含嚴重特殊傳染性肺炎 COVID-19（Coronavirus Disease-2019），因此先將資料集切分成四種資料集並將於實驗一中進行實驗，資料集切分已說明於章節 3.7 實驗建構中。

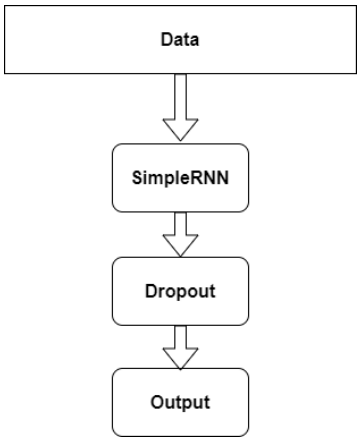
其中 NoCovid 資料集表示訓練集與測試集皆不涵蓋 COVID-19 影響期間；Covid_A 資料集表示測試資料集涵蓋正常股價波動期間及 COVID-19 影響期間；Covid_B 資料集表示測試資料集涵蓋 COVID-19 影響期間；Covid_C 資料集表示訓練資料集涵蓋前期之 COVID-19 影響期間，測試資料集則涵蓋後期之 COVID-19 影響期間；以上訓練集依序有 1258、1258、1510、1709 筆，測試集會有 252、704、452、253 筆，再進行 N 天的移動視窗處理(Window_size=N)，最終進入模型訓練的資料即會是 1258-N、1510-N、1709-N 筆。而本研究所有實驗之深度學習預測模型皆會重複跑 5 次，並取 5 次結果之平均以確認模型之穩定性。

而實驗中也會將六間公司之模型表現，依照權重大小分為兩組，以 GOOG、AAPL、AMZN 作為大資本公司代表，LUMN、NLSN、GPS 作為小資本公司代表，以探討基本面、技術面、文本資料等的加入，對於預測大小資本公司之股價間有無影響。

實驗中之 XGB 與 RF 均遵照預設值 n_estimators：100，而神經網路(RNN 與 LSTM)相關設定為：一層 RNN、兩層 LSTM，神經元分別為 64、128，激活函數使用 Tanh，中間為避免過擬合(Overfitting)問題會採取 Dropout Layer()，Epoch：500，Batch_size：8，Optimizer：adam，Early_stop：20。如圖 16、圖 17 所示。

圖 16

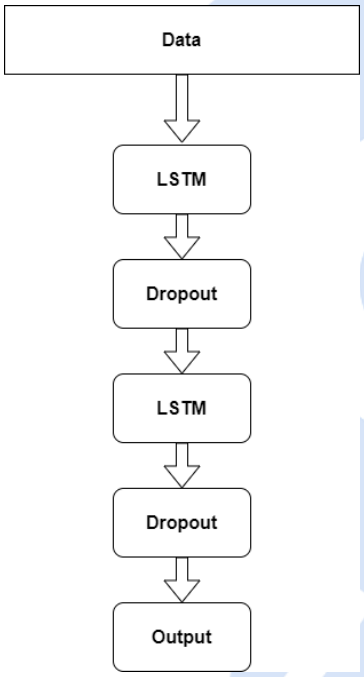
SimpleRNN 網路架構



資料來源:本研究整理

圖 17

LSTM 網路架構



資料來源:本研究整理

4.2.1 實驗一

實驗一利用先前研究之股價相關變數(每日開高低交易量、基本面變數、技術面變數)，共有 28 個變數，使用不同大小的移動視窗，再分別透過四種不同的演算法，分別為 XGBOOST、RF、RNN、LSTM 建立預測模型，並記錄其在測試集上之 F1-Score，並先計算 NoCovid、Covid_A、Covid_B、Covid_C 何者為最佳之資料集，再透

過最佳資料集之最佳 window_size，來決定後續實驗中移動視窗的大小。以 GOOG-window size 5 為例，首先會以 window size 5 建立 GOOG 的 XGBOOST、RF、RNN、LSTM 模型，然後進行預測並記錄四種演算法於上漲模型、下跌模型的 F1-Score，並在 Macro-Average 之 F1 上進行比較，如公式 43 所示。

$$\boldsymbol{Macro-Average\_F1-Score}=\frac{\boldsymbol{UP_{F1}+DOWN_{F1}}}{\boldsymbol{2}} \quad (43)$$

其中$\boldsymbol{UP_{F1}}$代表上漲模型之 F1-Score，代表$\boldsymbol{DOWN_{F1}}$下跌模型之 F1-Score。

表格中紅字呈現為該資料集之最佳演算法，粗體呈現為該演算法之最佳資料集為哪一個資料集。而大小資本表示大資本的三間公司(GOOG、AAPL、AMZN)之 F1-Score 會相加除以 3，以反應大資本公司之股價預測模型預測準確度，小資本的三間公司(LUMN、NLSN、GPS)也與大資本資料集進行相同方式計算。表 13 至 20 為實驗一之各資料集不同滑動視窗大小結果呈現。

表 13

實驗一：window_size 5 之各資料集 Macro-Average *F1 上漲結果*

大小資本	演算法	NOCOVID_F1	COVIDA_F1	COVIDB_F1	COVIDC_F1
大資本	XGB	0.477	0.470	**0.503**	0.497
	RF	**0.507**	0.487	0.500	0.503
	RNN	0.499	0.493	**0.502**	0.494
	LSTM	0.501	**0.519**	0.518	0.509
小資本	XGB	0.470	0.450	0.453	**0.490**
	RF	0.527	0.503	0.493	**0.533**
	RNN	**0.538**	0.513	0.502	0.485
	LSTM	**0.505**	**0.505**	0.495	0.480

資料來源:本研究整理

註：表 13 至表 20 中紅字呈現為該資料集之最佳演算法，粗體呈現為該演算法之最佳資料集為哪一個資料集，紅粗體呈現則為同時符合前兩者敘述。

表 14

實驗一：window_size 5 之各資料集 Macro-Average *F1 下跌結果*

大小資本	演算法	NOCOVID_F1	COVIDA_F1	COVIDB_F1	COVIDC_F1
大資本	XGB	0.480	0.477	0.457	**0.490**
	RF	0.503	0.490	0.513	**0.523**
	RNN	**0.501**	0.493	0.490	0.475
	LSTM	0.505	**0.509**	0.496	0.500
小資本	XGB	0.453	0.427	0.443	**0.487**
	RF	0.493	0.467	0.483	**0.537**
	RNN	0.492	0.487	0.493	**0.505**
	LSTM	0.497	0.481	0.478	**0.506**

資料來源:本研究整理

從表 13、表 14 中可以得知股價特徵在滑動視窗為 5 時，不同資料集上各個演算法建立上漲、下跌預測模型的結果。

在上漲模型中(表 13)，大資本於 NOCOVID 資料集以 RF 有最佳表現(0.507)；COVID_A、COVID_B、COVID_C 三個資料集都是 LSTM 有最佳預測表現(0.519、0.518、0.509)。小資本於 NOCOVID、COVID_A、COVID_B 三個資料集以 RNN 有最佳表現(0.538、0.513、0.502)；COVID_C 資料集則是 RF 達到最佳預測(0.533)。

下跌模型中(表 14)，大資本於 NOCOVID、COVID_A 資料集以 LSTM 為最佳預測模型(0.505、0.509)；COVID_B、COVID_C 資料集則是 RF 有最佳預測表現(0.513、0.523)。小資本於 NOCOVID 資料集以 LSTM 達到最佳預測(0.497)；COVID_A、COVID_B 以 RNN 有最佳表現(0.487、0.493)；COVID_C 資料集則是 RF 達到最佳預測(0.537)。

表 15

實驗一：window_size 10 之各資料集 Macro-Average *F1 上漲結果*

大小資本	演算法	NOCOVID_F1	COVIDA_F1	COVIDB_F1	COVIDC_F1
大資本	XGB	0.487	0.483	0.477	**0.493**
	RF	0.500	0.507	0.500	**0.537**
	RNN	0.513	0.511	**0.529**	0.519
	LSTM	0.496	0.503	0.502	**0.541**
小資本	XGB	**0.530**	0.457	0.453	0.463
	RF	**0.517**	0.503	0.507	0.507
	RNN	**0.508**	0.493	0.501	0.490
	LSTM	**0.528**	0.506	0.514	0.484

資料來源:本研究整理

表 16

實驗一：window_size 10 之各資料集 Macro-Average *F1 下跌結果*

大小資本	演算法	NOCOVID_F1	COVIDA_F1	COVIDB_F1	COVIDC_F1
大資本	XGB	0.470	0.473	0.460	**0.487**
	RF	0.503	0.483	0.497	**0.513**
	RNN	**0.503**	0.492	0.501	0.487
	LSTM	**0.505**	0.497	0.497	0.490
小資本	XGB	0.440	0.437	0.443	**0.500**
	RF	0.480	0.477	0.463	**0.497**
	RNN	0.497	0.497	0.495	**0.504**
	LSTM	0.491	0.492	0.452	**0.494**

資料來源:本研究整理

從表 15、表 16 中可以得知股價特徵在滑動視窗為 10 時，不同資料集上各個演算法建立上漲、下跌預測模型的結果。

在上漲模型中(表 15)，大資本於 NOCOVID、COVID_A、COVID_B 資料集以 RNN 有最佳表現(0.513、0.511、0.529)；COVID_C 以 LSTM 為最佳預測模型(0.541)。小資本於 NOCOVID、COVID_A、COVID_B 資料集皆是 LSTM 為最佳預測模型(0.528、0.506、0.514)；COVID_C 以 RF 為最佳預測模型(0.507)。

下跌模型中(表 16)，大資本於 NOCOVID、COVID_A 資料集以 LSTM 有最佳表現(0.505、0.497)；COVID_B 以 RNN 為最佳預測模型(0.501)；COVID_C 則是 RF 有最佳預測表現(0.513)。小資本於四個資料集皆是 RNN 為最佳預測模型(0.497、0.497、0.495、0.504)。

表 17

實驗一：window_size 15 之各資料集 Macro-Average *F1 上漲結果*

大小資本	演算法	NOCOVID_F1	COVIDA_F1	COVIDB_F1	COVIDC_F1
大資本	XGB	0.487	0.490	0.487	**0.553**
	RF	**0.513**	0.507	0.477	0.470
	RNN	0.501	0.509	0.507	**0.533**
	LSTM	0.511	0.506	0.508	**0.534**
小資本	XGB	**0.513**	0.463	0.457	0.453
	RF	**0.510**	0.507	0.497	**0.510**
	RNN	**0.515**	0.495	0.498	0.466
	LSTM	**0.525**	0.500	0.513	0.485

資料來源:本研究整理

表 18

實驗一：window_size 15 之各資料集 Macro-Average *F1 下跌結果*

大小資本	演算法	NOCOVID_F1	COVIDA_F1	COVIDB_F1	COVIDC_F1
大資本	XGB	0.467	0.457	0.460	**0.480**

大小資本	演算法	NOCOVID_F1	COVIDA_F1	COVIDB_F1	COVIDC_F1
	RF	0.507	0.517	0.503	**0.520**
	RNN	0.500	0.499	**0.517**	0.483
	LSTM	0.491	0.499	**0.504**	0.495
小資本	XGB	0.463	0.447	0.443	**0.480**
	RF	0.480	0.497	**0.507**	0.490
	RNN	0.489	0.496	**0.502**	0.477
	LSTM	0.482	0.497	0.474	**0.514**

資料來源:本研究整理

從表 17、表 18 中可以得知股價特徵在滑動視窗為 15 時，不同資料集上各個演算法建立上漲、下跌預測模型的結果。

在上漲模型中(表 17)，大資本於 NOCOVID 以 RF 有最佳表現(0.513)；COVID_A 是 RNN 為最佳預測模型(0.509)；COVID_B 以 LSTM 為最佳預測模型(0.508)；COVID_C 是 XGB 為最佳預測模型(0.553)。小資本於 NOCOVID、COVID_B 資料集是 LSTM 為最佳預測模型(0.525、0.513)；COVID_A、COVID_C 以 RF 為最佳預測模型(0.507、0.510)。

下跌模型中(表 18)，大資本於 NOCOVID、COVID_A、COVID_C 以 RF 有最佳表現(0.507、0.517、0.520)；COVID_B 以 RNN 為最佳預測模型(0.517)。小資本於 NOCOVID 資料集是 RNN 為最佳預測模型(0.489)；COVID_A、COVID_B 以 RF 為最佳預測模型(0.497、0.507)，而 COVID_A 之 LSTM 也有最佳表現(0.497)；COVID_C 則是 LSTM 為最佳預測模型(0.514)。

表 19

實驗一：window_size 20 之各資料集 Macro-Average *F1 上漲結果*

大小資本	演算法	NOCOVID_F1	COVIDA_F1	COVIDB_F1	COVIDC_F1
大資本	XGB	0.490	0.493	0.490	**0.530**

大小資本	演算法	NOCOVID_F1	COVIDA_F1	COVIDB_F1	COVIDC_F1
	RF	0.513	0.503	**0.533**	0.507
	RNN	0.479	0.509	0.509	**0.517**
	LSTM	**0.525**	0.502	0.508	0.524
小資本	XGB	**0.487**	0.477	0.450	0.447
	RF	**0.540**	0.493	0.497	0.497
	RNN	**0.525**	0.499	0.496	0.474
	LSTM	**0.527**	0.514	0.522	0.514

資料來源:本研究整理

表 20

實驗一：window_size 20 之各資料集 Macro-Average *F1 下跌結果*

大小資本	演算法	NOCOVID_F1	COVIDA_F1	COVIDB_F1	COVIDC_F1
大資本	XGB	**0.467**	0.450	0.460	0.460
	RF	0.480	0.513	0.483	**0.520**
	RNN	0.502	0.494	**0.507**	0.487
	LSTM	**0.503**	0.491	0.495	**0.503**
小資本	XGB	0.463	0.420	0.420	**0.480**
	RF	0.480	0.487	0.467	**0.503**
	RNN	0.484	**0.492**	0.489	0.478
	LSTM	0.472	0.474	0.470	**0.498**

資料來源:本研究整理

從表 19、表 20 中可以得知股價特徵在滑動視窗為 20 時，不同資料集上各個演算法建立上漲、下跌預測模型的結果。

在上漲模型中(表 19)，大資本於 NOCOVID 以 LSTM 有最佳表現(0.525)；COVID_A 是 RNN 為最佳預測模型(0.509)；COVID_B 以 RF 為最佳預測模型(0.533)；

COVID_C 是 XGB 為最佳預測模型(0.530)。小資本於 NOCOVID 資料集是 RF 為最佳預測模型(0.540)；COVID_A、COVID_B、COVID_C 以 LSTM 為最佳預測模型(0.514、0.522、0.514)。

下跌模型中(表 20)，大資本於 NOCOVID 以 LSTM 有最佳表現(0.503)；COVID_A、COVID_C 以 RNN 為最佳預測模型(0.513、0.520)；COVID_B 則是 RNN 表現最好(0.507)；小資本於 NOCOVID、COVID_A、COVID_B 資料集是 RNN 為最佳預測模型(0.484、0.492、0.489)；COVID_C 則是 RF 為最佳預測模型(0.503)。

在決定最佳 window_size 之前，首先先統整計算表 13 至表 20 中，各演算法最佳表現位於哪一個資料集(粗體標示於表 13 至表 20)，經過計算後發現 NOCOVID、COVID_A、COVID_B、COVID_C 上漲模型之最佳演算法數量依序為 17、2、4、11；而下跌模型之最佳演算法數量依序為 4、2、5、21，也代表大部分的演算法最佳表現均位於 NOCOVID、COVIDC 兩個資料集中，因此 NOCOVID 以及 COVID_C 會是較佳的資料集，而實驗二將利用前兩者資料集進行實驗。而 COVID_A、COVID_B 資料集預測狀況不佳推測有兩點原因：(1) COVIDA 及 COVIDB 訓練資料相對 COVID_C 來說較少；(2) COVIDA、COVIDB、COVIDC 測試集都包含 COVID 影響期間之資料，但是 COVIDC 的訓練集中有包含 COVID 影響期間之資料，使其可學習到波動較大的資料區間。表 21 呈現最佳 Window_size 計算。

表 21

實驗一：最佳 Window_size 計算

資本	演算法	NOCOVID_F1	COVIDA_F1	COVIDB_F1	COVIDC_F1
大資本	XGB	20	20	5	15
	RF	15,20	10,15	20	10
	RNN	10	10	10	15
	LSTM	20	5	5	10
小資本	XGB	10	20	15	5

資本	演算法	NOCOVID_F1	COVIDA_F1	COVIDB_F1	COVIDC_F1
	RF	20	15	10	5
	RNN	5	5	5	10
	LSTM	10	20	20	20

資料來源:本研究整理

表 21 計算演算法之最佳 Window_size 於各資料集之出現次數，舉例如大資本的 XGB 於 NOCOVID 資料集中，最佳數值為 0.490，出現於 Window_size 20 中，表 22 中的大資本 XGB 就填入「20」。

透過表 13 至表 20 結果得知，NOCOVID、COVIDC 為最佳資料集，因此計算表 21 中各個 Window_size 出現的次數就能得知最佳 Window_size 應該如何訂定，其中統整 NOCOVID 資料集之大小資本 Window_size 數量分布，從 Window_size=5 至 Window_size=20，NOCOVID 資料集之大小資本 Window_size=5 出現 1 次，Window_size=10 出現 3 次，Window_size=15 出現 1 次，Window_size=20 出現 4 次；COVIDC 資料集之大小資本 Window_size 數量分布，從 Window_size=5 至 Window_size=20，COVIDC 資料集之大小資本 Window_size=5 出現 2 次，Window_size=10 出現 3 次，Window_size=15 出現 2 次，Window_size=20 出現 1 次。同時考慮兩個資料集後，得出 Window_size=5 至 Window_size=20 之數量分布依序為 3、6、3、5。

根據以上結果，Window_size 即訂定為 10，因為 Window_size=10 於兩個資料集中出現最多的次數(6 次)，因此實驗二將使用 Window_size 10 的 NOCOVID、COVID_C 進行實驗。

而從實驗一可得出以下結論：(1)在股市沒有重大震盪時期(NOCOVID)建立股價預測模型能有較好的效果。(2)在股市重大震盪時期(COVID_A、COVID_B、COVID_C)建立股價預測模型，訓練資料應包含震盪時期之資料才能幫助模型訓練出較好的預測效果。

4.2.2 實驗二

實驗二在實驗一上，再加入 Glassdoor 員工評論之評分、TFIDF 文字特徵、員工評論變數與情感分析變數，作為股價預測模型資料。表 23 至表 26 中依序列出各資料集於上漲、下跌預測中，加入 Glassdoor 評論特徵、文字特徵及情感分析特徵在不同機器學習與深度學習分類器中的結果。TFIDF 相關的設定為 min_df=0.05，表示忽略詞彙頻率低於 5%的單字，最後得出的 Glassdoor 總文檔單字量如下表 22 所示，實驗二將由各資料集 min_df 後得出的單字數量建立文字特徵後進行實驗。

表 22

實驗二：Glassdoor - TFIDF 萃取文字整理

Dataset	Pros 單字量	Cons 單字量	min_df 後 Pros 單字量	min_df 後 Cons 單字量
GOOG	3,000	4,296	52	55
AAPL	2,033	3,122	27	24
AMAZ	3,759	6,024	65	86
LUMN	1,911	3,506	22	31
NLSN	2,530	3,939	52	55
GPS	1,900	3,078	42	38

資料來源:本研究整理

表 23

實驗二：NOCOVID 上漲類別結果整理

資料集	演算法	Precision	Recall	F1
大資本	XGB	0.503	0.433	0.447
	RF	0.520	0.537	0.487
	RNN	0.497	0.500	0.492
	LSTM	0.509	0.509	0.502
小資本	XGB	0.510	0.543	0.467

	RF	0.543	0.580	0.530
	RNN	0.520	0.525	0.518
	LSTM	0.506	0.512	0.502

資料來源:本研究整理

表 24

實驗二：NOCOVID 下跌類別結果整理

資料集	演算法	Precision	Recall	F1
大資本	XGB	0.503	0.453	0.463
	RF	0.500	0.470	0.457
	RNN	0.492	0.491	0.489
	LSTM	0.505	0.506	0.504
小資本	XGB	0.493	0.447	0.427
	RF	0.490	0.433	0.430
	RNN	0.524	0.528	0.519
	LSTM	0.501	0.502	0.495

資料來源:本研究整理

在實驗二之 NOCOVID 上漲預測模型中(表 23)，其中大資本 RF 有最好的 Precision(0.520)以及 Recall(0.537)，而 LSTM 有最佳的 F1-Score(0.502)。小資本中 RF 則是於 Precision、Recall 以及 F1 均取得最佳預測表現(0.543、0.580、0.530)。

在實驗二之 NOCOVID 下跌預測模型中(表 24)，其中大資本 LSTM 有最好的 Precision(0.505)、Recall(0.506)以及 F1(0.504)。小資本中則是 RNN 於 Precision、Recall 以及 F1 均取得最佳預測表現(0.524、0.528、0.519)。

表 25

實驗二：COVID_C 上漲類別結果整理

資料集	演算法	Precision	Recall	F1
大資本	XGB	0.520	0.567	0.483
	RF	0.497	0.503	0.490
	RNN	0.507	0.507	0.506
	LSTM	0.512	0.514	0.508
小資本	XGB	0.503	0.527	0.460
	RF	0.513	0.537	0.480
	RNN	0.508	0.509	0.502
	LSTM	0.511	0.519	0.501

資料來源:本研究整理

表 26

實驗二：COVID_C 下跌類別結果整理

資料集	演算法	Precision	Recall	F1
大資本	XGB	0.497	0.483	0.457
	RF	0.530	0.527	0.520
	RNN	0.526	0.514	0.505
	LSTM	0.510	0.507	0.497
小資本	XGB	0.497	0.487	0.450
	RF	0.493	0.490	0.480
	RNN	0.506	0.507	0.505
	LSTM	0.486	0.485	0.484

資料來源:本研究整理

在實驗二之 COVID_C 上漲預測模型中(表 25)，其中大資本 XGB 有最好的 Precision(0.520)以及 Recall(0.567)，而 LSTM 有最佳的 F1-Score(0.508)。小資本中 RF

則是於 Precision(0.513)、Recall(0.537)有最佳的預測表現，而 RNN 於 F1 取得最佳預測表現(0.502)。

在實驗二之 COVID_C 下跌預測模型中(表 26)，其中大資本 RF 有最好的 Precision(0.530)、Recall(0.527)以及 F1(0.520)。小資本中則是 RNN 於 Precision、Recall 以及 F1 均取得最佳預測表現(0.506、0.507、0.505)。

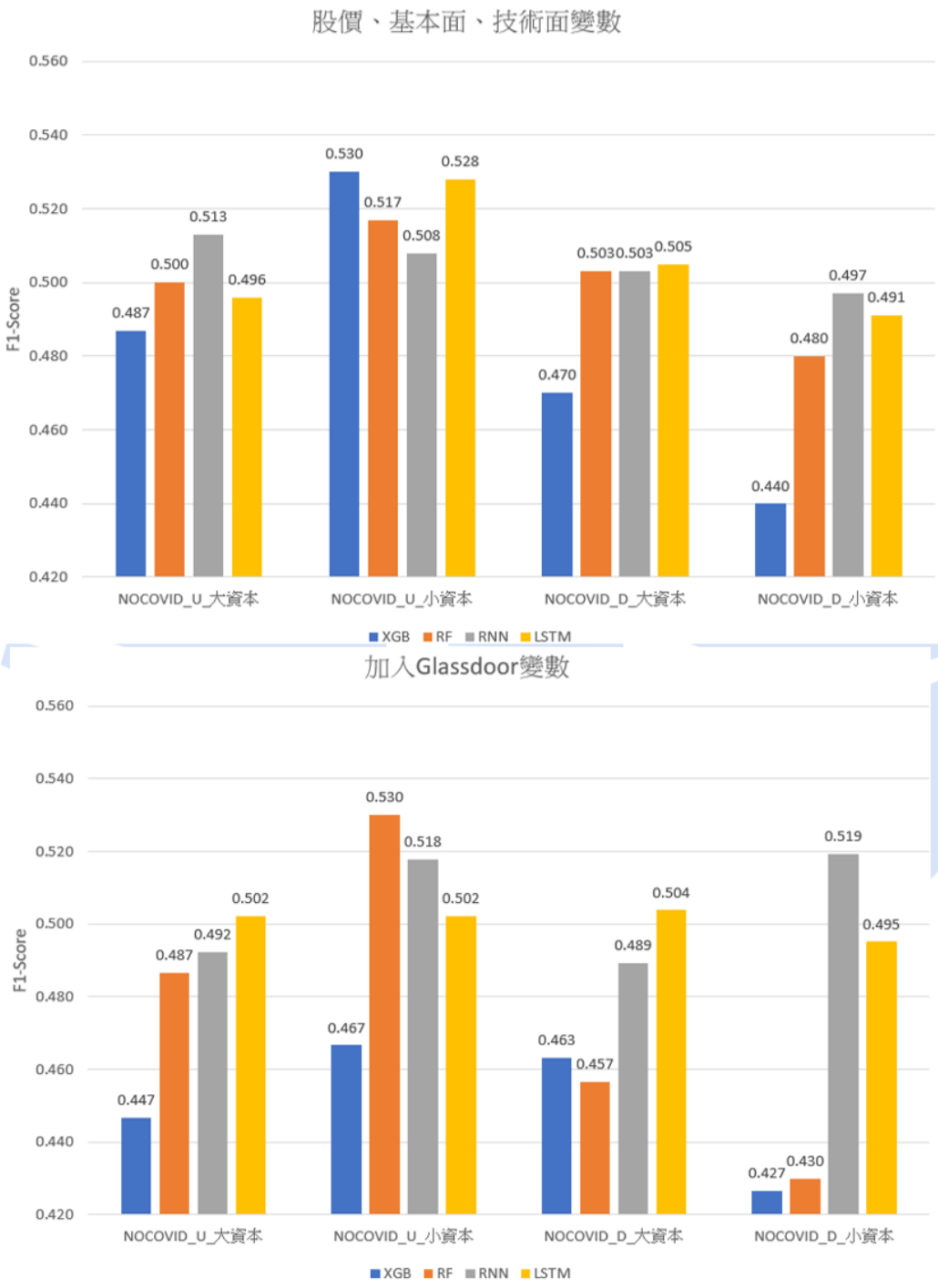
4.2.3 實驗結果比較

實驗一與實驗二

實驗一透過不同資料集(NoCovid、Covid_A、Covid_B、Covid_C)以及不同 Window_size，利用股價、基本面、技術面特徵，在 XGB、RF、RNN、LSTM 四種演算法上建立預測模型並嘗試找出最佳資料集與最佳 Window_size。實驗結果為 NoCovid、Covid_C 兩種資料集為最佳資料集，Window_size 則是 10 可達到平均最佳預測表現。而實驗二中則於實驗一之基礎上加入 Glassdoor 之員工評論特徵、情感特徵等變數。實際預測比較結果如下圖 18、圖 19 所示。

圖 18

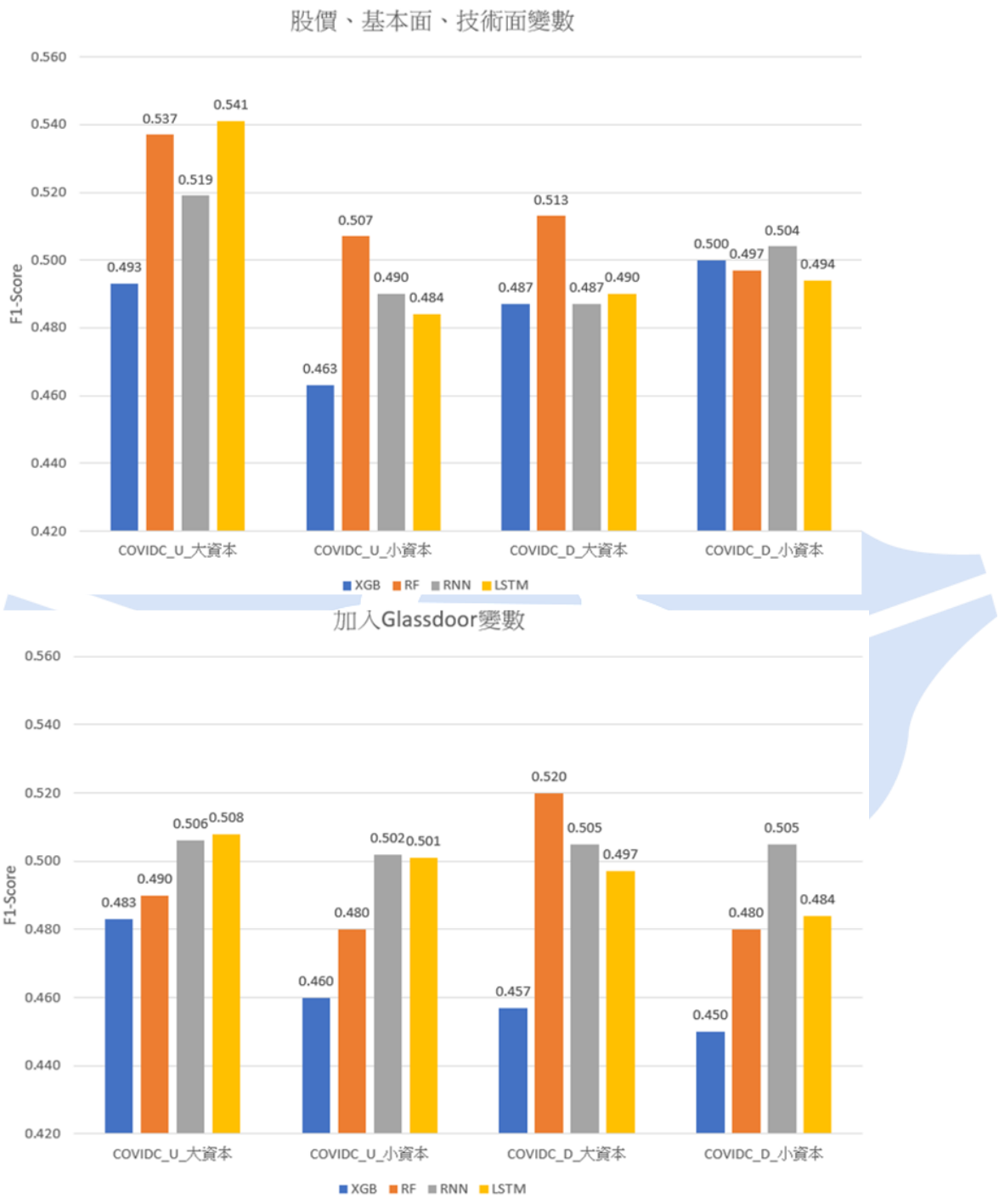
實驗一與實驗二之 NOCOVID 於 F1-Score 表現對比整理



資料來源:本研究整理

圖 19

實驗一與實驗二之 COVIDC 於 F1-Score 表現對比整理



資料來源:本研究整理

由上述實驗結果可知，NOCOVID 資料集在加入 Glassdoor 特徵後(圖 18)，NOCOVID_U 大資本中僅有 LSTM 提升預測性能(0.496 提升至 0.502)；NOCOVID_U 小資本中則是 RF、RNN 提升預測性能(RF 由 0.517 提升至 0.530，RNN 由 0.508 提升

至 0.518)；NOCOVID_D 大資本中則是所有演算法均下降預測性能，僅 LSTM 有非常接近的預測表現(0.505 下降至 0.504)；NOCOVID_D 小資本中則是 RNN、LSTM 提升預測性能(RNN 由 0.497 提升至 0.519，LSTM 由 0.491 提升至 0.495)。

而 COVIDC 資料集在加入 Glassdoor 特徵後(圖 19)，COVIDC_U 大資本中所有演算法均下降預測性能；COVIDC_U 小資本中則是 RNN、LSTM 提升預測性能(RNN 由 0.490 提升至 0.502，LSTM 由 0.484 提升至 0.501)，而 XGB 僅些微下降預測性能(0.463 下降至 0.460)；COVIDC_D 大資本中 RF、RNN、LSTM 皆提升預測性能(RF 由 0.513 提升至 0.520，RNN 由 0.487 提升至 0.505，LSTM 由 0.490 提升至 0.497)；COVIDC_D 小資本中則是僅有 RNN 提升些微預測性能(0.504 提升至 0.505)。

以大小資本面來說，實驗一中 NOCOVID 上漲資料集小資本之平均預測準確度較大資本來的高(0.499, 0.5208)，NOCOVID 下跌資料集則是大資本之平均預測準確度較高(0.4953, 0.477)；COVIDC 上漲資料集大資本之平均預測準確度較小資本來的高(0.5225, 0.486)，COVIDC 下跌資料集則是小資本之預測準確度些微較高(0.4943, 0.4988)；而實驗二 NOCOVID 上漲資料集小資本之平均預測準確度較高(0.482, 0.5043)，NOCOVID 下跌資料集則是大資本之平均預測準確度較高(0.4783, 0.4678)；COVIDC 上漲資料集大資本之平均預測準確度較小資本來的高(0.4968, 0.4858)，COVIDC 下跌資料集則是大資本之預測準確度較高(0.4948, 0.4798)。

而由預測模型面而言，發現機器學習模型如 XGB 及 RF 之準確度範圍相當大，可得出深度學習模型相較於機器學習模型較具有穩健性。

與我們相似的研究有 Ren et al. (2018)，他們使用上證 50 指數(SSE 50)中依照產業別，分別挑選了幾間公司，收集公司的收盤價、交易量、技術指標，結合新浪股票論壇、東方財富網的新聞資料產生出之情感分析變數，最後使用 SVM 建立預測模型，預測股票收盤價漲跌走勢，結果提出加入情感變數能有效提升預測準確度。

Vargas et al. (2017)收集了 2006/10/02 至 2013/11/21 的 S&P500 指數每日股價資料，先利用股價資料計算出技術指標，並透過爬蟲獲取路透社(Reuters)的財經新聞，並將財經新聞文本利用 Word2Vec 轉換成詞嵌入，將詞嵌入與技術指標結合後，最後

使用 RCNN 建立預測模型來預測 S&P500 每日收盤價漲跌走勢。結果提出加入新聞文本變數能有效提升預測準確度，而使用結合新聞文本與技術指標的 RCNN 能比單純只使用新聞或只使用技術指標的其他演算法有更佳預測表現。

而我們的研究則指出，結合員工情感以及文本特徵也能一定程度的提升預測準確度，但需要經過適當的特徵選取，萃取出較佳的特徵子集才能達到有效提升預測準確度。

第五章、研究結論與建議

5.1 結論

股價預測一直是投資人以及金融界相當感興趣且困難的研究議題，由於股市資料的非線性、非平穩性，使其也是學術研究上熱門的一大主題。近年來社群平台的發展，許多社群平台、論壇等都有人在討論股票相關話題，提供自己購股、選股的心得作為其他投資人決策之參考。過去研究多採用單純數值型特徵進行建模，並以基本面、技術面、籌碼面為主，消息面也多是只使用新聞資料或是論壇資料，並沒有深入探討其他類型的社群資料利用之可能性。

本研究採用國外知名的員工評論網站 Glassdoor 文本特徵來源，並以大小資本將公司進行分組，進而分析利用 Glassdoor 員工文本、情感特徵於大小資本公司對於預測模型之影響，並透過以下方式進行實驗：(1)分別考量大小資本之預測模型建立、(2)分別考量上漲類別預測模型、下跌類別預測模型之建立、(3)考量 COVID-19 疫情影響期間之資料集切割。以上皆使用四種演算法建立。

基於本研究之實驗結果，於學術面及實務面有以下之貢獻：(1)於學術方面，利用 Glassdoor 員工評論之情感分析變數及評論變數作為新面向的文字探勘社群資料，以提升股價漲跌預測準確性；(2)於實務方面，提供投資者一預測模型，使其能透過參考員工評論增進預測準確率。

5.2 研究限制

本研究使用 Glassdoor 之員工評論作為文本資料集，採用文字探勘技術 TF-IDF 以及 Vader 情感分析，並以隨機森林、極限梯度提升、循環神經網路、長短期記憶神經網路演算法建立預測模型，然而受到以下限制上有不周全之處：

1. 本研究實驗文本資料來自於 Glassdoor，在小資本的公司中，由於員工人數較少，故會有部分公司比起大資本公司其線上評論數量減少很多，可能造成小資本公司之預測模型不夠周全。
2. 本研究涵蓋產業別僅三種(「科技」、「消費者非必須消費品」、「通訊服務」)，未對 S&P500 中其他產業(如:「金融」等)進行不同的資料收集及資料前處理，可能造成分析不夠全面。

5.3 未來研究方向

於資料集方面，本研究採用 Glassdoor 作為文本資料來源，未來可考慮加入其他員工網站(如 Blind、Indeed.com)之文本，藉以補足小資本公司因資料過少而可能分析不夠周全之情況。此外，除了員工資料之外，未來也可考慮更多結合不同的消息面資訊納入分析，如新聞、社群文章等。

於模型方面，本研究著重於員工資料之特徵於股價漲跌之影響，未來可以再更細節探討員工特徵中是否能進一步找出更有效之特徵，以提升預測模型之有效性，使投資人於投資時有更佳的參考依據，亦可朝向研究深度學習網路之方面，使得模型預測能力得以更加優化。

參考文獻

中文文獻

查欣瑜（2011）。*法人籌碼對台股未來走勢影響之研究*。國立交通大學。

范聖培（2014）。*三大法人之買賣超行為對股價短期報酬之研究*。國立中央大學。

張維碩, 張智淵, & 張書豪（2018）。以向量自我迴歸模式探討台灣 50 成分股報酬率與技術面及籌碼面之關聯性。全球商業經營管理學報 (*10*), 177-187. http://doi.org/10.29967/JGBOM

黃宇翔 & 王百祿（2008）。*ARIMA 與適應性 SVM 之混合模型於股價指數預測之研究*。國立成功大學。

黃祺敦（2012）。*運用當日籌碼面變數預測隔日股價方向*。國立中正大學。

黃韻欣（2020）。*籌碼指標對股票報酬率之影響*。東海大學。

蔡尚翰(2017)。*籌碼面選股結合技術分析之投資績效研究*。國立高雄應用科技大學。

謝璁賦（2011）。*應用類神經網路於台股權值股籌碼面的知識發現*。國立交通大學。

英文文獻

Adebiyi, A. A., Ayo, C. K., Adebiyi, M., & Otokiti, S. O. (2012). Stock price prediction using neural network with hybridized market indicators. *Journal of Emerging Trends in Computing and Information Sciences*, *3*(1). https://doi.org/10.1109/UKSim.2014.67

Ballings, M., Van den Poel, D., Hespeels, N., & Gryp, R. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert systems with Applications*, *42*(20), 7046-7056. https://doi.org/10.1016/j.eswa.2015.05.013

Bharathi, S., Geetha, A., & Sathiynarayanan, R. (2017). Sentiment analysis of twitter and RSS news feeds and its impact on stock market prediction. *International Journal of Intelligent Engineering and Systems*, *10*(6), 68-77. https://doi.org/10.22266/ijies2017.1231.08

Billah, M., Waheed, S., & Hanifa, A. (2016). *Stock market prediction using an improved training algorithm of neural network* 2016 2nd International Conference on Electrical, Computer & Telecommunication Engineering (ICECTE), Piscataway, New Jersey, United States.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, *2*(1), 1-8. https://doi.org/10.1016/j.jocs.2010.12.007

Bramer, M. (2007). Avoiding overfitting of decision trees. *Principles of data mining*, 119-134. https://doi.org/10.1007/978-1-84628-766-4_8

Bustos, O., Pomares, A., & Gonzalez, E. (2017). *A comparison between SVM and multilayer perceptron in predicting an emerging financial market: Colombian stock market*

2017 Congreso Internacional de Innovacion y Tendencias en Ingenieria (CONIITI), Bogota, Colombia.

Chavarnakul, T., & Enke, D. (2008). Intelligent technical analysis based equivolume charting for stock trading using neural networks. *Expert systems with Applications*, *34*(2), 1004-1017. https://doi.org/10.1016/j.eswa.2006.10.028

Chen, H.-L., Chow, E. H., & Shiu, C.-Y. (2013). Ex-dividend prices and investor trades: Evidence from Taiwan. *Pacific-Basin Finance Journal*, *24*, 39-65. https://doi.org/10.1016/j.pacfin.2013.02.004

Chen, Q., Zhang, W., & Lou, Y. (2020). Forecasting stock prices using a hybrid deep learning model integrating attention mechanism, multi-layer perceptron, and bidirectional long-short term memory neural network. *IEEE Access*, *8*, 117365-117376. https://doi.org/10.1109/ACCESS.2020.3004284

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,

Chen, Y., & Hao, Y. (2017). A feature weighted support vector machine and K-nearest neighbor algorithm for stock market indices prediction. *Expert systems with Applications*, *80*, 340-355. https://doi.org/10.1016/j.eswa.2017.02.044

Chen, Y., Lin, W., & Wang, J. Z. (2019). A dual-attention-based stock price trend prediction model with dual features. *IEEE Access*, *7*, 148047-148058. https://doi.org/10.1109/ACCESS.2019.2946223

Deng, S., Mitsubuchi, T., Shioda, K., Shimada, T., & Sakurai, A. (2011). *Combining technical analysis with sentiment analysis for stock price prediction* 2011 IEEE ninth international conference on dependable, autonomic and secure computing, Sydney, NSW, Australia.

Ding, C., & Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, *3*(02), 185-205.

Duan, J., & Zeng, J. (2013). *Mining opinion and sentiment for stock return prediction based on web-forum messages* 2013 10th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Shenyang, China.

Earnest, A., Chen, M. I., Ng, D., & Sin, L. Y. (2005). Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore. *BMC Health Services Research*, *5*(1), 1-8. https://doi.org/10.1186/1472-6963-5-36

Edmans, A. (2011). Does the stock market fully value intangibles? Employee satisfaction and equity prices. *Journal of Financial economics*, *101*(3), 621-640. https://doi.org/10.1016/j.jfineco.2011.03.021

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179-211. https://doi.org/10.1207/s15516709cog1402_1

Emir, S., Dinçer, H., & Timor, M. (2012). A stock selection model based on fundamental and technical analysis variables by using artificial neural networks and support vector machines. *Review of Economics & Finance*, *2*(3), 106-122. https://api.semanticscholar.org/CorpusID:10271923

Fama, E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, *25*(2), 383-417. https://doi.org/10.2307/2325486

Feng, S. F. (2020). Job Satisfaction, Management Sentiment, and Financial Performance: Text Analysis with Job Reviews from Indeed. com. *Authorea Preprints*. https://doi.org/10.22541/au.159596366.69672878

Göçken, M., Özçalıcı, M., Boru, A., & Dosdoğru, A. T. (2016). Integrating metaheuristics and artificial neural networks for improved stock price prediction. *Expert systems with Applications*, *44*, 320-331. https://doi.org/10.1016/j.eswa.2015.09.029

Gao, P., Zhang, R., & Yang, X. (2020). The application of stock index price prediction with neural network. *Mathematical and Computational Applications*, *25*(3), 53. https://doi.org/10.3390/mca25030053

Gao, T., & Chai, Y. (2018). Improving stock closing price prediction using recurrent neural network and technical indicators. *Neural computation*, *30*(10), 2833-2854. https://doi.org/10.1162/neco_a_01124

Green, T. C., Huang, R., Wen, Q., & Zhou, D. (2019). Crowdsourced employer reviews and stock returns. *Journal of Financial economics*, *134*(1), 236-251. https://doi.org/10.2139/ssrn.3002707

Guo, K., Sun, Y., & Qian, X. (2017). Can investor sentiment be used to predict the stock price? Dynamic analysis based on China stock market. *Physica A: Statistical Mechanics and its Applications*, *469*, 390-396. https://doi.org/10.1016/j.physa.2016.11.114

Ho, T. K. (1995). *Random decision forests* Proceedings of 3rd international conference on document analysis and recognition, Montreal, QC, Canada.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, *9*(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554-2558. https://doi.org/10.1073/pnas.79.8.2554

Hu, Y., Shao, L., La, L., & Hua, H. (2021). Using Investor and News Sentiment in Tourism Stock Price Prediction based on XGBoost Model. 2021 IEEE/ACIS 6th International Conference on Big Data, Cloud Computing, and Data Science (BCD),

Hu, Z., Zhao, Y., & Khushi, M. (2021). A survey of forex and stock price prediction using deep learning. *Applied System Innovation*, *4*(1), 9. https://doi.org/10.3390/asi4010009

Huang, C.-J., Yang, D.-X., & Chuang, Y.-T. (2008). Application of wrapper approach and composite classifier to the stock trend prediction. *Expert systems with Applications*, *34*(4), 2870-2878. https://doi.org/10.1016/j.eswa.2007.05.035

Huang, M., Li, P., Meschke, F., & Guthrie, J. P. (2015). Family firms, employee satisfaction, and corporate performance. *Journal of Corporate Finance*, *34*, 108-127. https://doi.org/10.1016/j.jcorpfin.2015.08.002

Huang, Q., Yang, J., Feng, X., Liew, A. W.-C., & Li, X. (2019). Automated trading point forecasting based on bicluster mining and fuzzy inference. *IEEE Transactions on Fuzzy Systems*, *28*(2), 259-272. https://doi.org/10.1109/TFUZZ.2019.2904920

Hutto, C., & Gilbert, E. (2014). *Vader: A parsimonious rule-based model for sentiment analysis of social media text* Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, Michigan, United States.

Jianqiang, Z., Xiaolin, G., & Xuejun, Z. (2018). Deep convolution neural networks for twitter sentiment analysis. *IEEE Access*, *6*, 23253-23260. https://doi.org/10.1109/ACCESS.2017.2776930

Jin, Z., Yang, Y., & Liu, Y. (2020). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing and Applications*, *32*(13), 9713-9729. https://doi.org/10.1007/s00521-019-04504-2

Jordan, M. I. (1997). Serial order: A parallel distributed processing approach. In *Advances in psychology* (Vol. 121, pp. 471-495). Elsevier. https://doi.org/10.1016/S0166-4115(97)80111-2

Joshi, S., & Li, Y. (2016). What is corporate sustainability and how do firms practice it? A management accounting research perspective. *Journal of Management Accounting Research*, *28*(2), 1-11. https://doi.org/10.2308/jmar-10496

Kamble, R. A. (2017). *Short and long term stock trend prediction using decision tree* 2017 International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India.

Khare, K., Darekar, O., Gupta, P., & Attar, V. (2017). *Short term stock price prediction using deep learning* 2017 2nd IEEE international conference on recent trends in electronics, information & communication technology (RTEICT), Bangalore, India.

Kim, R. Y. (2020). The Value of Followers on Social Media. *IEEE Engineering Management Review*, *48*(2), 173-183. https://doi.org/10.1109/EMR.2020.2979973

Kirlić, A., & Orhan, Z. (2017). Measuring human and Vader performance on sentiment analysis. *Invention Journal of Research Technology in Engineering & Management (IJRTEM)*, *1*(12), 42-46.

Ko, C.-R., & Chang, H.-T. (2021). LSTM-based sentiment analysis for stock price forecast. *PeerJ Computer Science*, *7*, e408. https://doi.org/10.7717/peerj-cs.408

Kordonis, J., Symeonidis, S., & Arampatzis, A. (2016). *Stock price forecasting via sentiment analysis on Twitter* Proceedings of the 20th Pan-Hellenic Conference on Informatics, Patras, Greece.

Kumar, M., & Thenmozhi, M. (2006). *Forecasting stock index movement: A comparison of support vector machines and random forest* Indian institute of capital markets 9th capital markets conference paper, Navi Mumbai, Maharashtra, India.

Kyoung-Sook, M., & Hongjoong, K. (2019). PERFORMANCE OF DEEP LEARNING IN PREDICTION OF STOCK MARKET VOLATILITY. *Economic Computation & Economic Cybernetics Studies & Research*, *53*(2). https://doi.org/10.24818/18423264/53.2.19.05

Lee, C.-Y., & Soo, V.-W. (2017). *Predict stock price with financial news based on recurrent convolutional neural networks* 2017 conference on technologies and applications of artificial intelligence (TAAI), Taipei, Taiwan.

Lee, Y.-T., Liu, Y.-J., Roll, R., & Subrahmanyam, A. (2004). Order imbalances and market efficiency: Evidence from the Taiwan Stock Exchange. *Journal of Financial and Quantitative Analysis*, *39*(2), 327-341. https://doi.org/10.1017/S0022109000003094

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, *69*, 14-23. https://doi.org/10.1016/j.knosys.2014.04.022

Li, Y., & Pan, Y. (2022). A novel ensemble deep learning model for stock prediction based on stock prices and news. *International Journal of Data Science and Analytics*, *13*(2), 139-149.

Liu, Y., Qin, Z., Li, P., & Wan, T. (2017). *Stock volatility prediction using recurrent neural networks with sentiment analysis* International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems, Arras, France.

Lo, A. W. (2005). Reconciling efficient markets with behavioral finance: the adaptive markets hypothesis. *Journal of investment consulting*, *7*(2), 21-44. https://doi.org/https://ssrn.com/abstract=1702447

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35-65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Lu, W., Li, J., Wang, J., & Qin, L. (2021). A CNN-BiLSTM-AM method for stock price prediction. *Neural Computing and Applications*, *33*(10), 4741-4753. https://doi.org/10.1007/s00521-020-05532-z

Lubitz, M. (2017). *Who drives the market? Sentiment analysis of financial news posted on Reddit and Financial Times*

Luo, N., Zhou, Y., & Shon, J. (2016). *Employee satisfaction and corporate performance: Mining employee reviews on glassdoor. com* 2016 International Conference on Information Systems, Dublin, Ireland.

Mizuno, H., Kosaka, M., Yajima, H., & Komoda, N. (1998). Application of neural network to technical analysis of stock market prediction. *Studies in Informatic and control*, *7*(3), 111-120. https://sceweb.uhcl.edu/boetticher/ML_DataMining/Kimoto.pdf

Mohan, S., Mullapudi, S., Sammeta, S., Vijayvergia, P., & Anastasiu, D. C. (2019). *Stock price prediction using news sentiment analysis* 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService), Newark, CA, USA.

Moniz, A., & de Jong, F. (2014). *Sentiment analysis and the impact of employee satisfaction on firm earnings* European conference on information retrieval, Amsterdam, The Netherlands.

Namdari, A., & Li, Z. S. (2018). *Integrating fundamental and technical analysis of stock market through multi-layer perceptron* 2018 IEEE Technology and Engineering Management Conference (TEMSCON), Evanston, IL, United States.

Nelson, D. M., Pereira, A. C., & de Oliveira, R. A. (2017). *Stock market's price movement prediction with LSTM neural networks* 2017 International joint conference on neural networks (IJCNN), Anchorage, AK, United States.

Nofsinger, J. R. (2001). The impact of public information on investors. *Journal of Banking & Finance*, *25*(7), 1339-1366. https://doi.org/10.1016/S0378-4266(00)00133-3

Paliari, I., Karanikola, A., & Kotsiantis, S. (2021). A comparison of the optimized LSTM, XGBOOST and ARIMA in Time Series forecasting. 2021 12th International Conference on Information, Intelligence, Systems & Applications (IISA),

Peng, D. (2019). *Analysis of investor sentiment and stock market volatility trend based on Big Data strategy* 2019 International Conference on Robots & Intelligent System (ICRIS), Haikou, China.

Picasso, A., Merello, S., Ma, Y., Oneto, L., & Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. *Expert systems with Applications*, *135*, 60-70. https://doi.org/10.1016/j.eswa.2019.06.014

Polanco-Martínez, J. M. (2019). Dynamic relationship analysis between NAFTA stock markets using nonlinear, nonparametric, non-stationary methods. *Nonlinear Dynamics*, *97*(1), 369-389. https://doi.org/10.1007/s11071-019-04974-y

Quah, T.-S. (2008). DJIA stock selection assisted by neural network. *Expert systems with Applications*, *35*(1-2), 50-58. https://doi.org/10.1016/j.eswa.2007.06.039

Ranco, G., Aleksovski, D., Caldarelli, G., Grčar, M., & Mozetič, I. (2015). The effects of Twitter sentiment on stock price returns. *PloS one*, *10*(9), e0138441. https://doi.org/10.1371/journal.pone.0138441

Ren, R., Wu, D. D., & Liu, T. (2018). Forecasting stock market movement direction using sentiment analysis and support vector machine. *IEEE Systems Journal*, *13*(1), 760-770. https://doi.org/10.1109/JSYST.2018.2794462

Roubaud, D., & Arouri, M. (2018). Oil prices, exchange rates and stock markets under uncertainty and regime-switching. *Finance research letters*, *27*, 28-33. https://doi.org/10.1016/j.frl.2018.02.032

Schneider, B., Hanges, P. J., Smith, D. B., & Salvaggio, A. N. (2003). Which comes first: employee attitudes or organizational financial and market performance? *Journal of applied psychology*, *88*(5), 836. https://doi.org/10.1037/0021-9010.88.5.836

Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. *ACM Transactions on Information Systems (TOIS)*, *27*(2), 1-19. https://doi.org/10.1145/1462198.1462204

Shaikh, A., Panuganti, A., Husain, M., & Singh, P. (2021). *Stock Market Prediction Using Machine Learning* Proceedings of International Conference on Intelligent Computing, Information and Control Systems, Secunderabad, India.

Shim, S., & Pourhomayoun, M. (2017). *Predicting movie market revenue using social media data* 2017 IEEE International Conference on Information Reuse and Integration (IRI), San Diego, CA, United States.

Shiva Nandhini, J., Bari, C., & Pradip, G. (2020). Stock Market Prediction Using Machine Learning. *Journal of Computational and Theoretical Nanoscience*, *17*(4), 1584-1589. https://doi.org/10.1166/jctn.2020.8405

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., & Deng, X. (2013). *Exploiting topic based twitter sentiment for stock prediction* Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Sofia, Bulgaria.

Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018). *A comparison of ARIMA and LSTM in forecasting time series* 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), Orlando, FL, United States.

Singh, D., & Singh, B. (2020). Investigating the impact of data normalization on classification performance. *Applied Soft Computing*, *97*, 105524. https://doi.org/10.1016/j.asoc.2019.105524

Skuza, M., & Romanowski, A. (2015). *Sentiment analysis of Twitter data within big data distributed environment for stock prediction* 2015 Federated Conference on Computer Science and Information Systems (FedCSIS), Lodz, Poland.

Smith, A. (1937). *The wealth of nations: An inquiry into the nature and causes*. Modern Library. http://www.math.chalmers.se/~ulfp/Review/smith1.pdf

Tan, Z., Yan, Z., & Zhu, G. (2019). Stock selection with random forest: An exploitation of excess return in the Chinese stock market. *Heliyon*, *5*(8), e02310. https://doi.org/10.1016/j.heliyon.2019.e02310

Teoh, T.-T., Lim, W., Koh, K., Soh, J., Tan, T., Liu, S., & Nguwi, Y.-Y. (2019). *From Technical Analysis to Text Analytics: Stock and Index Prediction with GRU* 2019 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), Bangkok, Thailand.

Vargas, M. R., De Lima, B. S., & Evsukoff, A. G. (2017). Deep learning for stock market prediction from financial news articles. 2017 IEEE international conference on computational intelligence and virtual environments for measurement systems and applications (CIVEMSA),

Vu, T.-T., Chang, S., Ha, Q. T., & Collier, N. (2012). *An experiment in integrating sentiment features for tech stock prediction in twitter* Proceedings of the Workshop on Information Extraction and Entity Analytics on Social Media Data, Mumbai, India.

Wang, B., Huang, H., & Wang, X. (2012). A novel text mining approach to financial time series forecasting. *Neurocomputing*, *83*, 136-145. https://doi.org/10.1016/j.neucom.2011.12.013

Wooley, S., Edmonds, A., Bagavathi, A., & Krishnan, S. (2019). *Extracting Cryptocurrency Price Movements from the Reddit Network Sentiment* 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), Boca Raton, FL, United States.

Yetis, Y., Kaplan, H., & Jamshidi, M. (2014). *Stock market prediction by using artificial neural network* 2014 World Automation Congress (WAC), Waikoloa, Hawaii, United States.

Yu, P., & Yan, X. (2020). Stock price prediction based on deep neural networks. *Neural Computing and Applications*, *32*(6), 1609-1628. https://doi.org/10.1007/s00521-019-04212-x

Zhao, Y., & Chen, Z. (2021). Forecasting stock price movement: New evidence from a novel hybrid deep learning model. *Journal of Asian Business and Economic Studies*. https://doi.org/10.1108/JABES-05-2021-0061