逢甲大學
資訊工程學系
碩士論文

基於財經字典與分析指標的神經網路預測股價趨勢

Predicting Stock Price Trend Using Neural Network Based on Financial Lexicon and Technical Indicators

指導教授：張哲誠博士

研究生：潘亮晴

中華民國一百一十一年七月

誌　　謝

首先我要感謝我的指導老師，張哲誠教授，一開始進入實驗室時我有許多需要努力改進的部分，感謝教授耐心的給予指導。與教授一對一研究討論時幫助我更快的釐清問題，還有怎麼把複雜的事情簡單化；生活上也會時不時的關心實驗室的研究生近況，在與教授談天時也能向老師學習不只是學業上的事情，能當老師的研究生真是太好了！

在碩士的這段期間經歷了許多學業上的挫折，感謝讓我能夠渡過這些挫折並且陪我一起經歷的實驗室朋友、畢業的學長姊。在實驗室時常有問題就可以就近得到解答，也能夠隨時一起嘻笑，上課時分工合作，下課時一起吃飯，每一件事看似平凡，但卻是日常生活中抒發學業壓力的重要管道，多虧了有他們的陪伴與幫忙，我的研究生生活很美好。

也要感謝我的家人，時有電話關心，平常會教我怎麼做人、教我怎麼選擇，雖然我不是每一件事都會依照他們的意思，但還是會繼續關愛著我、支持著我，讓我更有動力去完成這份學業，非常感謝她們。

摘　　要

股票市場中影響投資人做出決策的面向有非常多元，無論是公司的主要營運項目、公司財務狀況還有市場的景氣等等都會受影響，因此在分析股價時，必須將這些因素都納入考量。以往有關預測股價的論文中，大多都會使用新聞、公司個體財報資訊或是歷史股價來預測股價趨勢，很少研究同時將這三種決策資訊納入考量。

本研究旨在使用多元線性迴歸模型 (multiple linear regression) 和人工神經網路模型 (Artificial neural network) 預測股價，以有大量新聞的公司台積電作為研究對象，擷取其在新聞、分析指標、歷史股價上的多方資訊。本文蒐集來自富果網站上的財金新聞，並將財金新聞做「台積電」、「大盤相關新聞」兩大分類，使用自製情感字典計算出兩大分類的新聞情緒分數，自製爬蟲程式蒐集台積電的分析指標與歷史股價，最後將兩大新聞情緒分數、分析指標、歷史股價作為預測股價的特徵。

關鍵詞：情感分析、人工神經網路、線性迴歸、股價預測

Abstract

There are many aspects of the stock market that affect investors' decisions, including main operating projects, financial condition of a company and the market sentiment, etc. These factors must be taken into account while analyzing stock prices. In previous papers on predicting stock prices, most of them have used news, the company's financial situation or historical stock prices to predict stock price trends. Few studies have taken all of them into consideration at once.

This study uses multiple linear regression (MLR) model and Artificial neural network (ANN) model to predict the stock price, taking TSMC, a company with a large amount of news, as the study object. Collecting TSMC historical stock price, news, and financial statements. This study crawls financial news published by various newspapers from January 1, 2017 to December 31, 2021. We tag financial news as two major categories: "TSMC Related News" and "Market Related News" and calculate the news sentiment scores of the two major categories using a customized sentiment analysis dictionary. We also use a customized crawler program to collect financial statement and technical indicators. news sentiment scores, technical indicators and historical stock prices are used as features to predict stock price trends.

Keywords: Sentiment analysis, Artificial neural network, Linear regression, Stock price prediction

目　　錄

圖目錄

表目錄

第一章　緒論

1.1 研究背景

1.1.1 台灣的投資市場概要

台灣的金融市場有多元投資標的：黃金、基金、ETF、股票、期貨、選擇權、加密貨幣，而其中基金、ETF、股票都是風險相對較小，要求自有資本相對較少的投資標的，因此吸引眾多投資人。

影響股票的因素有很多種，環境因素像是股票以外的投資標的如果在當時的投資風險相對較小、獲得報酬相對較高，則將會有可能使投資人變賣股票轉而投資其他投資標的；在台灣股票市場有許多不同的投資人角色，包含散戶或法人，而最能影響台灣大盤的角色正是法人，法人每日成交量動輒上億，因此會影響股價，而在台灣的法人又可以分為「外資」、「投信」、「自營商」，他們被稱為三大法人；景氣和世界前兩大經濟體美國、中國將左右投資人的投資意願；政府政策或政治因素可能帶動特定產業類型的發展，讓投資人投資特定類別的股票。這些都是投資人在投資股票時必須考量的要素，因此每天都有眾多的新聞釋出這些資訊以利投資人做出投資決策。

股票市場的投資人有本國自然人、本國法人、僑外自然人、僑外法人，其中自然人就是投資市場中俗稱的”散戶”，一般而言散戶與法人之間存在資訊不對等的問題，法人在經濟規模、利益關係上的優勢，較容易比散戶更快取得更準確的消息。

根據台灣證券交易所2020年資料顯示，台灣總成交股數為906,809,114 張股票[1]，而台灣投資人類別成交值比重統計[2]如下：

表1. 1 台灣投資人類別成交值比重統計

本國自然人		本國法人		僑外自然人		僑外法人	
買進	賣出	買進	賣出	買進	賣出	買進	賣出
31.25	30.82	5.68	5.56	0.02	0.03	13.05	13.59
62.07		11.24		0.05		26.64	

單位：百分比

根據台灣證券交易所2020年資料顯示，台灣年度上市公司資本來源明細表年報[3]如下：

表1. 2 台灣年度上市公司資本來源明細表年報

上市公司資本來源	比例
政府機構	5.03
本國金融	6.27
本國證券	1.18
本國公司	23.09
本國其他	2.71
僑外金融	0.72
僑外法人	9.87
僑外證券	14.45
本國自然人	36.1
僑外自然人	0.44
庫藏股票	0.14

單位：百分比

由兩表顯示散戶的交易量比重與上市公司投資金額比重大約佔股市整體62%與36.5%，可見散戶積極投資股市，但就金額而言其對上市公司的影響力只有不到四成。

1.1.2 景氣下的股價變動因素

台灣股市大盤與景氣息息相關，影響景氣的因素可以是政府政策或是國際影響。當經濟走向衰退，政府將會使用貨幣寬鬆政策刺激經濟發展與消費動能；[4]指出，台灣政治在股票市場上有『政治景氣循環』現象，執政黨為了得到選票而在選舉前一年提出貨幣寬鬆政策。

國際議題包括商業、科技、政治等等也都會影響到景氣，研究指出COVID-19爆發下，對於整體股市而言，[5]收集了 Market Watch、紐約時報和路透社於 2020 年 1 月 23 日和 2020 年 6 月 22 日期間有關 COVID-19 的新聞，使用 Google BERT 進行情緒分析，分析新聞對於股票市場標準普爾 500 指數讀影響，證明新聞具有積極影響力；對於特定類型的股票而言，全球首例滅活 COVID-2019疫苗臨床試驗的宣布發現對醫藥股產生了積極影響，證實人們在疫情流行期間對有關疫苗研發的消息很敏感[6]。

1.1.3 應用機器學習的股價趨勢預測

由於投資股票能為投資人帶來可觀的收益，因此股價趨勢預測已經是學術上的熱門研究議題。現今已出版大量的股價預測相關的學術論文，無論在國內亦或是國外都有大量的學術論文可考。機器學習的發展與盛行，讓金融領域得以和人工智慧合作，絕大部分的股價趨勢預測實驗都是以機器學習進行，並且使用的機器學習非常多樣化，針對的金融議題也有所不同。[7]做市場層級的研究：利用新聞情緒作為發行股票公司之間的關聯性，使用股價預測股價趨勢。[8]做產業層級的研究：研究對於所有三種產業—IT、銀行和醫療保健，MARS已被證明是研究中股票預測表現最好的模型。[9]做公司層級的研究：將台積電作為研究對象，針對短期數據建構了技術分析的神經網絡。

除了研究對象規模不同，這些與股價趨勢預測的機器學習輸入的特徵也有所差異，輸入的特徵可能是新聞分析[10]、基本面分析、技術分析[9]，又會從這三者中做不同的資料處理做成資料集。

互聯網技術的發展使投資者透過電子媒體更容易獲取股票市場資訊，新聞對股市的影響，主要有三個方面：（1）公司特定新聞文章的基本面資訊影響投資者的交易活動;（2）新聞喚起公眾情緒，投資者決策受公眾情緒影響因而被干預投資決策;（3）網路媒體對股票的影響因新聞內容和公司特徵而異[11]。

根據[12]的說法，基本面分析主要基於三個基本方面（1）宏觀經濟分析，如國內生產總值（GDP）和居民消費價格指數（CPI），分析宏觀經濟環境對公司未來利潤的影響，（2）行業分析，根據行業現狀和前景估計公司的價值，（3）公司分析，分析公司的當前運營和財務狀況，以評估其內部價值。

預測股價趨勢所輸入機器學習的特徵如基本面分析由於數量眾多，故[13]利用決策樹和多元迴歸方法預測銀行業，並研究結果顯示輸入變量的減少對模型的預測性能有積極的影響。

1.2 研究動機

影響股價變動的因素眾多，就規模而言，景氣、產業、單一公司都有不同的影響因子，投資者需要經常透過新聞資訊取得景氣、產業、單一公司的相關資訊以茲投資決策。

投資者除了新聞，還可以從單一公司的會計報表判斷目前公司經營狀況，並且為了投資者更容易判讀會計數據，投資者將會計資訊使用 EPS、ROE、毛利率等等這些分析指標藉以分析數據，這些分析指標資訊可以從網路上易於取得而不用投資者自行計算或繪圖，因此是眾投資者時常使用的股價分析工具。

在短期投資決策內，歷史股價也是一個可以參考的資訊，投資人可能藉由近期股價的動盪幅度判斷購買時機。

本研究希望使用新聞資訊、分析指標、歷史股價這些股價參考資訊，並考量景氣與公司個體影響股價之因素，使用機器學習做出一套輔助投資者投資股票的預測單一公司股價趨勢模型。

1.3 研究目的

本研究蒐集歷史股價、股市新聞、財務報表做為資料集，研究台灣2017年1月1日至2021年12月31日的股市資料，探討使用多元特徵輸入模型的預測股價方法是否合適；以及不同切割資料集的方法是否能提升模型預測準確率。並計算均方根誤差(Root Mean Squared Error, RMSE)、準確率(Accuracy)、精確率(Precision)、召回率(Recall)、F1-Score 評估本研究的模型效益，以訓練出最適合預測股價的模型。

1.4 章節介紹

本文第二章節介紹論文背景技術與文獻回顧；第三章介紹本研究的研究設計與實施方法；第四章是研究設計方法的實證結果與分析；第五章為結論與未來工作。

第二章　文獻探討

2.1 機器學習

2.1.1 多元線性迴歸模型(Multiple Linear Regression, MLR)

普通最小均方誤差線性迴歸(Ordinary least squares Linear Regression.) 擬合係數為 $\mathrm{w} = w_1, w_2, \ldots, w_n$ 的線性模型，藉由線性近似去最小化觀察目標與預測目標之間的誤差平方和。起先被應用於統計領域，後因機器學習的興起，因其適合被應用在模型的輸入與輸出資料之間的關係屬於線性的研究，因此被廣泛應用在序列模型或是分類模型的機器學習案例中。

在輸入的變量為單變量的案例中，線性迴歸稱為簡單線性迴歸；在輸入的變量為多變量的案例中，線性迴歸稱為多元線性迴歸(Multiple Linear Regression)。多元線性迴歸須通過迭代以求出最小均方誤差和。

多元線性迴歸

$$\mathrm{y} = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

其中x_i為輸入變量，w_i為係數，b為殘差，y為輸出變量。

2.1.2 人工神經網路(Artificial neural network, ANN)

人工神經網路是現今熱門的議題，在歷史的演進下，電腦的運算能力大幅提升，發展到電腦可以承受神經網路的龐大矩陣運算量，致使應用神經網路的機器學習開始盛行。神經網路可以應用於時間序列預測以及分類器，可以輸入單一特徵及多元特徵，資料的種類可以是文字、語音、影像等，可以透過改變神經元的數量、隱藏層的層數、激發函數讓模型達到更好的預測效果。

人工神經網路本身是一種反饋神經網路(Backward Propagation Neural Network)，因其透過多微分方程式或是迭代的方式，將運算後的結果利用損失函數進行參數更新，讓模型係數朝著最小化損失函數的方向更新。它的特點是迭代速度快、學習精度高、能夠處理非線性關係數據[9]。

2.1.3 損失函數

從每個係數的隨機值開始，模型使用訓練數據和真實數據的誤差來優化係數值至最小化誤差的操作，在機器學習中被稱為梯度下降。梯度下降的目的就是將損失函數最小化，多元線性迴歸和 ANN 皆可使用均方誤差和(Mean-square Error, MSE)作為損失函數(Loss Function)，梯度下降直至無法將損失函數收斂至更小的數值時，此時即找到了誤差最少的最佳解，代表模型已經訓練完成。

損失函數

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(observed_i - predicted_i)^2$$

其中$observed_i$為觀察目標，也就是真實資料；$predicted_i$為預測目標，也就是預測資料。

2.2 新聞情緒分析

新聞中所透漏的情緒可以影響投資者造成市場波動[11]，對新聞文本之情緒分析也因而開始盛行。新聞文本會顯示出公司個體、市場整體情況，同時帶來短期或是長期的影響。透過分析新聞，我們可以得到一些有用的資訊，將之作為影響股價的因素。

2.2.1 基於字典的情感分析

詞頻(term frequency)是一個詞出現在一篇文本中的頻率，除以文本總共使用的詞語總數，來表示一詞在文本中的佔比，佔比高的詞語表示該詞的意義越為重要。無意義的詞語可以使用反文檔頻率(Inverse Document Frequency)過濾，因整個資料集中越高頻率出現的詞語越不具重要含意。

文獻經常使用字典作為特徵提取的方法，一部字典裡所含的詞語具有特定特徵，定義的字典越精確，就越能精確的提取出文本的特定特徵。字典也時常搭配詞頻，一篇文本中頻繁出現字典中的詞語，則表示該文本具有該項特定特徵。特定特徵時常被劃分為正向或是負向，例如[10]中字典的特定特徵為股價波動，針對股市新聞的正向含意表示股價上漲；負向含意表示股價下跌，並透過詞頻計算出情感分數，若一篇文章的正面情緒詞個數-負面情緒個數>0，則預測為漲；若一篇文章的正面情緒詞個數-負面情緒詞個數<0，則預測為跌。[10]、[14]都使用同樣的方式提取文本中的情感特徵，同時使用其他研究所建立的字典與自行建立的字典進行研究。

第三章　研究設計與實施

3.1 實驗流程

本研究所提出預測短期股價趨勢的機器學習流程圖如圖所示：

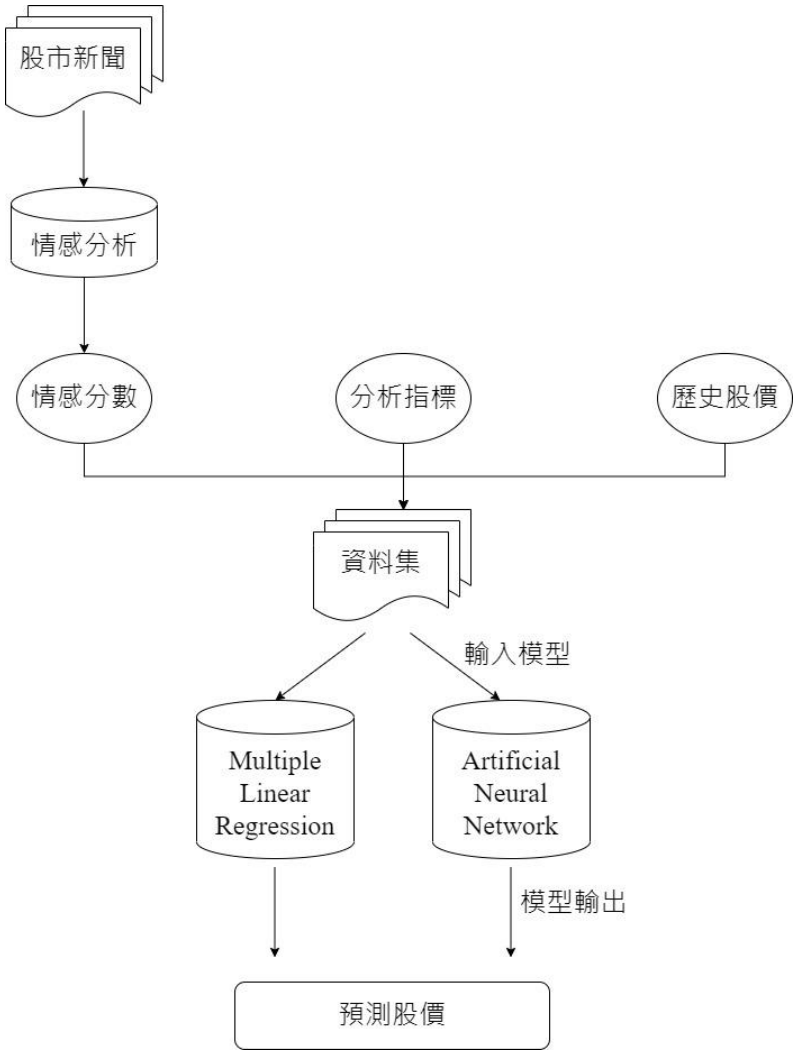

圖3. 1 實驗架構圖

A. 股市新聞：利用 Python 程式語言撰寫程式，呼叫富果網(Fugle)上的新聞 API，蒐集2017年1月1日至2021年12月31之台積電股市相關新聞，去除掉重複資料後，總計筆數979筆。

B. 情感字典：使用台灣大學情感分析字典 NTUSD 與自製財經字典 VFinDict 作為情感分析模型使用的字典。

C. 情感分析模型： 具有將新聞文章斷句、分類斷句、利用情感字典計算情感分數的功能。

D. 情感分數：情感分數可代表新聞屬於意味著股價上漲的正向情緒或是股價下跌的負面情緒，有兩大類情感分數，分別為景氣情感分數與公司個體情感分數。

E. 分析指標：在富果網(Fugle)上有多個針對公司個體的分析指標，分析指標共有營收、EPS、利潤比率、ROE 及 ROA、成長能力、經營能力、償債能力這七大類分析指標，蒐集2017年1月1日至2021年12月31的分析指標。

F. 歷史股價：使用套件 FinMind 抓取歷史股價資訊，蒐集2017年1月1日至2021年12月31日的成交股數、成交金額、開盤價、最高價、最低價、收盤價、漲跌價差、成交筆數。

G. 資料集：將兩大類情感分數、十一大類分析指標、歷史股價資訊整合成2017年1月1日至2021年12月31日每日資料集，並將2017年1月1日到2020年12月31日資料做為訓練資料集，約莫為整份資料集八成的資料；2021年1月1日到2021年12月31日做為測試資料集，約莫為整份資料集兩成的資料。

H. 神經網路模型：使用 Python 程式撰寫多元線性迴歸模型(Multiple Linear Regression)、人工神經網路模型(Artificial neural network)。

I. 預測股價：模型輸出為預測明天或後天的股價。

J. 評估成果：使用準確率(Accuracy)評估情感分數與股市漲幅關係；均方根誤差(Root Mean Squared Error, RMSE)機器學習模型評估指標，評估預測成果。

3.2 資料蒐集

3.2.1 爬蟲程式

本研究使用 Python 撰寫程式腳本爬取富果網上提供的歷年新聞，觀察網站前端向後端呼叫之 API 何者屬於回應歷史新聞資料的 API，並透過對 API 結構觀察出以 GET 方法傳遞股票代號與起始日期參數即可得到該起始日以後與該股票代號相關的若干筆日新聞資料。由於本研究的研究對象為上市公司台積電，因此以 Python 套件 requests 以 GET 方法傳遞台積電的股票代號2330與2017年到2021年每日的時間戳呼叫 API，取得每日的新聞。

API 回傳的內容是 json 格式，因此為了要結構化讀取內容，使用了 pandas 套件將內容讀取成 DataFrame 的資料類型，並進一步觀察資料內容。

資料內容包含若干筆新聞資料，每一筆新聞都有唯一值的 ID，爬取時會有新聞重複出現的情況，因此如遇到已存在於新聞資料集的 ID 則不會加入新聞資料集，避免重複新聞存在於新聞資料集；時間戳則使用套件 datetime 將愈爬取的日期轉換為電腦時間，例如：「 2021-01-01 00:00:0」轉換成電腦時間「1609430400000」。

3.2.2 富果網的股市新聞

富果網[15]針對每個股票顯示個別的股票新聞，也就是要先查詢股票才能夠得到該股票相關的新聞。本研究針對台積電於富果網上爬取的資料每日約有2~15筆新聞，其新聞來源來自鉅亨網、Moneydj 理財網、中央社、時報資訊、東森財金、財訊快報等多家報社發布的財金新聞，每筆新聞資料都包含新聞ID、新聞標題、新聞內容、出版新聞報社、新聞發布時間戳、新聞 URL，如表3. 1 爬取新聞內容所示。

表3. 1 爬取新聞內容

欄位	欄位敘述
_id	新聞 ID
title	新聞標題
content	新聞內容
source	出版新聞報社
timestamp	新聞發布時間戳
url	新聞 URL

3.2.3 分析指標

富果網[15]中有多方面為投資人考量的設計，例如提供投資人分析股票時可能會對公司個體做基本面、消息面、技術面、籌碼面、財務報表這五大面向分析，這些面向的分析富果網皆有提供。基本面實際上就是從財務報表衍伸出來的分析方式，將公司擁有的四大報表：資產負債表、損益表、現金流量表、股東權益變動表中的資訊加以使用數學公式計算，使投資人更易於了解財務報表顯示出的公司狀況。

本研究的分析指標使用基本面分析法，從富果網呼叫與基本面分析對應的API 取得2017至2021年的基本面分析資料，將資料結構化後統整之資料如表3. 2

分析指標所示：

表3. 2 分析指標

基本面	計算期間	分析指標
本益比	日	本益比(PER)
股價淨值比	日	股價淨值比(PBR)
營收	月	營收
EPS	季	EPS
利潤比率	季	毛利率
		營業利益率
		稅後淨利率
ROE 及 ROA	季	ROE
		ROA
成長能力	季	營收季成長率
		每股盈餘季成長率
		毛利季成長率
		營業利益淨成長率
		稅後淨利淨成長率
經營能力	季	應收帳款週轉率
		存貨週轉率
		不動產及設備週轉率
		總資產週轉率
償債能力	季	流動比率
		速動比率
		利息保障倍數

3.2.4 歷史股價

FinMind 套件提供以台股為主，超過 50 種金融開源數據(open data)，希望讓大數據、資料分析，減少資料收集的門檻[16]。本研究使用其提供的股價日成交資訊來下載2017年1月1日至2021年12月31日的歷史股價資料，包含「date、stock_id、Trading_Volume、Trading_money、open、max、min、close、spread、Trading_turnover」，如表3. 3 歷史股價所示。

表3. 3 歷史股價

欄位	欄位敘述
date	當日日期
stock_id	股票代號
Trading_Volume	成交股數
Trading_money	成交金額
open	開盤價
max	最高價
min	最低價
close	收盤價
spread	漲跌價差
Trading_turnover	成交筆數

3.3 資料預處理

3.3.1 斷詞

閱讀文章時，閱讀者可以在閱讀時正確地將每一字句正確的斷詞，因此可得知整個句子、整篇文章想表達的正確語意。在中文裡，要是斷詞錯誤，就會被判讀成錯誤的資訊，所以在進行語意分析任務中，必須要將中文文本的詞正確的拆解，需要使用斷詞工具。斷詞工具普遍具有內建字典與自身的演算法，包含巨量詞語的字典可以更正確的斷詞，而如遇新詞語有一些斷詞工具也可以使用自身演算法為新詞斷詞，以下為 Python 中文斷詞工具的比較：

1. 結巴(jieba)[17]：

   是一款 Python 中文分詞套件。簡體字和繁體字雖然字體互通，但是兩者慣用詞、慣用句是不一樣的，而結巴內建字典是簡體字文本，所以對簡體字斷詞表現較好，內建字典共有349,046個詞語。

   結巴使用演算法 HMM(Hidden Markov Model)中的 Viterbi 演算法為未曾出現過的語句斷詞。

   其斷詞效率優良快速，但斷詞較不準確，需要自行建立使用者字典提升精準度。

2. CKIP tagger[18]：

   一款由中央研究院資訊科學研究所 CKIP Lab 所研發基於深度學習模型之斷詞工具， 訓練文本的資料來源於中央社、維基百科、中央研究院現代漢語標記語料庫，字典以詞向量的形式儲存。其斷詞效果非常準確，但因其使用深度學習模型進行分詞，因此分詞效率慢。

3.3.2 CKIPTagger

本研究選用 CKIPTagger 作為斷詞工具，表3. 4 斷詞範例為經過 CKIPTagger 斷詞後的文章範例。

表3. 4 斷詞範例

文本內容	台積電即將在19日召開法人說明會，由於大立光法說釋出利多訊息，市場關注台積電對下半年科技業景氣看法及蘋果(Apple)新機拉貨力道，加上台積電現金股息即將發放，外資領到股息後會回頭買股還是匯出，也關乎指數未來發展方向。
斷詞後	台積電 即將 在 19日 召開 法人 說明會 ， 由於 大立光 法說 釋出 利多 訊息 ， 市場 關注 台積電 對 下 半 年 科技業 景氣 看法 及 蘋果 (Apple) 新機 拉貨 力道 ， 加上 台積電 現金 股息 即將 發放 ， 外資 領到 股息 後 會 回頭 買股 還是 匯出 ， 也 關乎 指數 未來 發展 方向 。

3.3.3 情感字典

在 NLP 任務中，使用情感字典分析文本十分常見，可以透過帶有情感的詞語計算出整個文本帶有正向情緒或是負向情緒，也因此字典中的詞語量與精確度越高，越能準確判別一個文本的情感。

投資人視所獲得的消息判斷股市漲跌，因此可將投資人預期股票上漲視為正向情緒，投資人預期股票下跌視為負向情緒，故本研究以此為準則手動建立一個屬於財經領域的情感字典 VFinDict（ 表3. 5 VFinDict 情感字典範例)；本研究同時也加入 NTUSD[19]台灣大學自然語言處理研究室建立之中文情緒字典，藉由通用的情緒詞語判別新聞做一般性敘述時的文本情緒。字典中所有詞語統計如表3. 6 字典詞語統計所示。

表3. 5 VFinDict 情感字典範例

正面詞	負面詞
熱錢湧入	訂單流失
擴大市佔	商譽受損
法人看好	收賄弊案
財報亮眼	資金出逃
銷售一空	擦鞋童

表3. 6 字典詞語統計

字典	正面詞數量	負面詞數量	總數
NTUSD	2,812	8,276	11,088
VFinDict	412	237	649

清除重複詞語後，本研究總計使用之正面詞數量3,159個、負面詞數量8,472個、總數11,631個。

3.3.4 台股總覽

在使用 CKIPTagger 進行斷詞時，由於 CKIPTagger 無法去識別股票公司名稱，有可能將「大立光」分詞成「大立」、「光」，因此需要先取得台股所有股票列表，本研究使用 FinMind 下載台股總覽，台股總覽列出台灣所有上市上櫃的股票名稱，代碼和產業類別，並將台股總覽（表3. 7 台股總覽）加入 CLIPTagger 使用者自建字典中。

表3. 7 台股總覽

產業類別	股票代碼	股票名稱	交易所	發行日期
ETF	0050	元大台灣50	twse	2021-10-05
ETF	0051	元大中型100	twse	2021-10-05
ETF	0052	富邦科技	twse	2021-10-05

3.3.5 文句特徵提取

從富果網取得的新聞有以下情況：(1) 新聞只有提到台積電、 (2) 新聞討論台灣股市整體現況、(3) 新聞提及眾多股票公司。在這樣的情況下，必須將新聞做分類，才能將研究對象聚焦在台積電，而不會將其他不相關的新聞也加入資料集中。

本研究參考[10]的方法，依照這些狀況進行處理，首先將新聞斷句，也就是將一個句子的開頭到一個句子的句末標點符號「。、？、！」視為一個句子，並用台股總覽找出句子中提到的公司，依照該公司的產業類別、股票代號、股票名稱儲存成「年月_股票代號_公司名稱.csv」的 csv 檔，將句子放入其中（表3. 8 201712_2330_台積電檔案中的資料範例、表3. 9 201712_9919_康那香檔案中的資料範例）。

表3. 8 201712_2330_台積電檔案中的資料範例

出版日期	含有台積電的句子
2017-12-01	台積電 ADR 30日隨著美國科技股回穩，台積電今跳空開高2.5元為228.5元，盤中穩步走揚至230元之上，終場收復231元季線關卡，最高觸及233.5元。

表3. 9 201712_9919_康那香檔案中的資料範例

出版日期	含有康那香的句子
2017-12-01	近日台灣飽受空汙之苦，國人防疫意識興起，康那香 (9919) 、美德醫療-DR (9103) 、恆大 (1325) 挾著不織布、口罩題材持續發燒，股價逆勢揚升，康那香漲幅達2%，美德醫大漲5.7%，恆大收漲1.5%。

斷句後就能避免將非台積電的新聞被放到資料集中，進而讓模型誤判的情形。

本研究也以一些特定的關鍵字將句子做分類，例如只要有提到**景氣**的就會被歸類到**景氣**的檔案下儲存，因為景氣會影響到投資市場。本研究也因此將新聞做兩大分類：(1)台積電相關新聞 (2)大盤相關新聞（表3. 10 關鍵字分類(一)）。

表3. 10 關鍵字分類(一)

新聞分類	關鍵字
台積電相關新聞	半導體、電子、晶圓、台積電、奈米、三星
大盤相關新聞	台股、大盤、外資、投信、自營商、法人、加權指數、美股、台灣、美國、 景氣

這兩大分類又可以依關鍵字細分：如果文本提到該關鍵字且文本具有正向情感，則表示台積電的股票價格會上漲；如果文本提到該關鍵字且文本具有正向情感，則表示台積電的股票價格會下跌這兩種情況（表3. 11 關鍵字分類(二)）。

表3. 11 關鍵字分類(二)

新聞分類	句子具有正向情感則台積電股價上漲之關鍵字	句子具有正向情感則台積電股價下跌之關鍵字
台積電相關新聞	半導體、電子、晶圓、台積電、奈米	三星、英特爾
大盤相關新聞	台股、大盤、外資、投信、自營商、法人、加權指數、台灣、 景氣、美股、美國	

3.3.6 情感分數計算

經過文句特徵提取後，便可以開始計算分數。首先將每一則新聞都使用斷詞工具 CKIPTagger 斷詞後，將斷詞與情感字典中的詞語比對，如果斷詞與情感字典中的詞語一致，則表示該詞語具有情感特徵。

讓股價上漲的情感特徵視為正向情感特徵，讓股價下跌的情感特徵視為負向情感特徵，每比對到一個斷詞與正向情感特徵詞語一致，則計分1分；美比對到一個斷詞語負向情感特徵詞語一致，則計分-1分。經過比對當日的所有文本後將分數加總，如果總和分數大於0，則意味著當日新聞有著正向情感，股價可能因而上漲；如果總和分數小於0，則意味著當日新聞有著負向情感，股價可能因而下跌。情緒分數計分範例如表3. 12 情感分數計分範例所示。

表3. 12 情感分數計分範例

日期	2017-02-13	2017-02-13	總計
文句斷詞	在 台幣升值 態勢 ， 熱錢湧 入 明顯 下 ， 台股 今日 再 收上 9700 關卡 壓力 ， 目 前 台積電 正在 區間 上緣 位 置 ， 一旦 上 攻 突破 前 高 ， 台股 將 持續 加速 上 漲 幅度 。	台北 晶圓 代工 大廠 台積電 (2330) 將 在 本 周 二 (14 日) 舉行 董事會 ， 可望 公布 股利 政策 ， 市場 預 期 ， 台積電 今年 現金 股利 將 從 去年 6 元 起跳 ， 上' 看 8 元 ， 利多 帶動 下 台 積電 今 (13) 日 股價 走 強 ， 漲幅 逾 2% ， 站穩 多頭 均線 之上 。	
情感比對	0 -1 0 0 1 0 0 0 0 0 0 1 0 0 -1 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 1 0 1 0 0 0 1 1 0 0 0	
情感分數	4	8	12

3.3.7 預測交易日投入之資料時間

本研究旨在預測股價，並將欲預測股價之當日稱為「預測交易日」，而凡是股市有開市日子都稱為「交易日」，統一用詞防止後續敘述時間造成混淆。

由於本研究輸入多種資料，每種資料都有不同的時間區間，因此需要個別定義。假定預測交易日為2021年12月31日，則將預測交易日前一天的日歷史股價、預測交易日前一天的日新聞情感分數、預測交易日前一天的日分析指標、預測交易日前一月的月分析指標、預測交易日前一季的季分析指標做為特徵輸入模型中，而如果前一日休市而無資料的話，則找最近的交易日遞補資料。資料與資料標籤日期調整範例表3. 13 資料與資料標籤日期調整範例。

表3. 13 資料與資料標籤日期調整範例

預測交易日之標籤	預測交易日之特徵資料
2021-12-24(五) 收盤價	23日歷史股價 23日新聞情感分數 11月分析指標 第三季分析指標
2021-12-27(一) 收盤價	24日歷史股價 26日新聞情感分數 11月分析指標 第三季分析指標

3.3.8 機器學習資料集

將日歷史股價、日新聞情感分數、日分析指標、月分析指標、季分析指標提取之後，依照預測交易日投入之資料時間製作成資料集，接著將資料集依照最大值-最小值正規化方法正規化。需要將資料集最大值-最小值正規化是因為，每一種特徵的數值範圍都不一樣，例如流動比率的數值只會落在0到1之間，然而營收的數字卻是上億的數字，將會導致機器學習模型受巨大的數字差距影響導致影響學習成效。經過最大值-最小值正規化方法正規化後每一個特徵的數值都會介於0到1之間。

最大值-最小值正規化方法(Min-Max Normalization)

$$\frac{X - X_{min}}{X_{max} - X_{min}} \in [0,1]$$

其中X為單一屬性資料，X_{max}為單一屬性資料的最大值，X_{min}為單一屬性資料的最小值。

最後，將目前為止含有2017年1月1日到2021年12月31日數據集，使用兩種不同的方法切割數據集投入機器學習訓練：

(1) 傳統切割方法：訓練出一個可以預測未來一整年的模型。

表3. 14 傳統切割法資料集

訓練資料集	訓練資料筆數	測試資料集	測試資料筆數
2017/01/03 ~ 2020/12/30	979	2021/01/03 ~ 2021/01/17	244

依照時間，將2017年1月1日到2020年12月31日切分為訓練資料集(約佔總資料集的80%)，2021年1月1日到2021年12月31切分為測試資料集(約佔總資料集的20%)。

(2) 滑動窗格法(sliding window)：訓練資料集和測試資料集隨著時間改變。該法更貼近真實狀況，每一期的訓練均包含了近期的數據。

圖3. 2 滑動視窗法切割資料集概念圖

表3. 15 滑動窗格法各期數資訊

期數編號	訓練資料集	訓練資料筆數	測試資料集	測試資料筆數
1	2017/01/03 ~ 2020/12/30	979	2021/01/03 ~ 2021/01/17	10
2	2017/01/03 ~ 2021/01/14	989	2021/01/17 ~ 2021/01/31	10
3	2017/01/03 ~ 2021/01/28	999	2021/01/31 ~ 2021/02/23	10
4	2017/01/03 ~ 2021/02/22	1009	2021/02/23 ~ 2021/03/10	10
5	2017/01/03 ~ 2021/03/09	1019	2021/03/10 ~ 2021/03/24	10
6	2017/01/03 ~ 2021/03/23	1029	2021/03/24 ~ 2021/04/11	10
7	2017/01/03 ~ 2021/04/08	1039	2021/04/11 ~ 2021/04/25	10
8	2017/01/03 ~ 2021/04/22	1049	2021/04/25 ~ 2021/05/10	10
9	2017/01/03 ~ 2021/05/09	1059	2021/05/10 ~ 2021/05/24	10
10	2017/01/03 ~ 2021/05/23	1069	2021/05/24 ~ 2021/06/07	10

11	2017/01/03 ~ 2021/06/06	1079	2021/06/07 ~ 2021/06/22	10
12	2017/01/03 ~ 2021/06/21	1089	2021/06/22 ~ 2021/07/06	10
13	2017/01/03 ~ 2021/07/05	1099	2021/07/06 ~ 2021/07/20	10
14	2017/01/03 ~ 2021/07/19	1109	2021/07/20 ~ 2021/08/03	10
15	2017/01/03 ~ 2021/08/02	1119	2021/08/03 ~ 2021/08/17	10
16	2017/01/03 ~ 2021/08/16	1129	2021/08/17 ~ 2021/08/31	10
17	2017/01/03 ~ 2021/08/30	1139	2021/08/31 ~ 2021/09/14	10
18	2017/01/03 ~ 2021/09/13	1149	2021/09/14 ~ 2021/09/30	10
19	2017/01/03 ~ 2021/09/29	1159	2021/09/30 ~ 2021/10/17	10
20	2017/01/03 ~ 2021/10/14	1169	2021/10/17 ~ 2021/10/31	10
21	2017/01/03 ~ 2021/10/28	1179	2021/10/31 ~ 2021/11/14	10
22	2017/01/03 ~ 2021/11/11	1189	2021/11/14 ~ 2021/11/28	10
23	2017/01/03 ~ 2021/11/25	1199	2021/11/28 ~ 2021/12/12	10
24	2017/01/03 ~ 2021/12/09	1209	2021/12/12 ~ 2021/12/26	10
25	2017/01/03 ~ 2021/12/23	1219	2021/12/26 ~ 2021/12/29	4

每一期的訓練集都會新增10筆資料，並預測最近的10天。

3.4 機器學習模型

3.4.1 多元線性迴歸模型

使用 sklearn 套件線性模型中的多元線性迴歸模型，將訓練資料集輸入模型中訓練，接著將測試資料集輸入訓練好的模型中做預測，將預測結果使用圖形呈現，並計算出模型的 RMSE 以評估模型表現。

3.4.2 人工神經網路模型

使用 Tensorflow 的線性堆疊模型構成多層的神經元模型（圖3. 3 人工神經網路模型架構），將訓練資料集的輸入模型，接著將測試資料集輸入訓練好的模型中做預測，將預測結果使用圖形呈現，並計算出模型的 RMSE 以評估模型表現，再持續依照 RMSE 的成效調整模型參數。

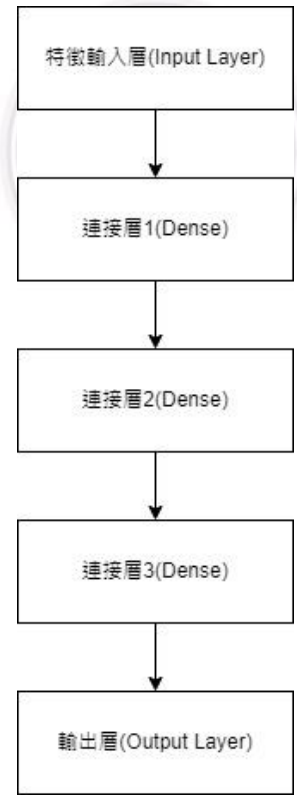

圖3. 3 人工神經網路模型架構

表3. 16 人工神經網路模型參數說明

參數名稱	資料型態	說明
seed	int	模型權重的隨機亂數種子，產生初始神經元的權重
input_dim	int	輸入層神經元個數，同時為訓練集特徵數
units	int	神經元個數
activation	string	激發函數
optimizer	string	決定更新權重方式的優化器
loss function	string	均方誤差作為損失函數以用來收斂模型
learning_rate	float	決定每一次更新的步長
decay	int	每次權重更新後學習率衰減值
momentum	float	用於加快在收斂方向上的速度
nesterov	boolean	決定是否使用 nesterov 動量
batch_size	int	每一次迭代時訓練集樣本個數
epochs	int	迭代所有批次一次為一期，epochs 決定迭代的期數

3.5 實驗設計

3.5.1 實驗方法

使用兩種模型(ANN、MLR)預測股價趨勢，將情感分數、分析指標、歷史股價資訊做為輸入特徵，在實驗中調整模型參數組合 random seed, Layer Units, learning rate, decay, momentum, nesterov, optimizer, loss, epochs, batch size，找出最適合預測短期股價的模型。

表3. 17 實驗模型類別

模型類別	特徵	資料集	模型架構
傳統法-MLR	情感分數 分析指標 歷史股價資訊	傳統法	MLR
傳統法-ANN		傳統法	ANN
視窗法-ANN		滑動視窗法	ANN

3.6 成果評估方法

3.6.1 字典法評估指標

使用準確率(Accuracy)、精確率(Precision)、召回率(Recall)、F1-score 做為使用字典法計算出的情感分數的評估指標，評估分數與股市漲跌幅之間的關係，觀察其與股價漲跌狀況的準確率。

若情感分數為正數且隔一天股價上漲，則該類別為 TP(True Positive)；

若情感分數為負數且隔一天股價下跌，則該類別為 TN(True Negative)；

若情感分數為正數且隔一天股價下跌，則該類別為 FP(False Positive)；

若情感分數為負數且隔一天股價上漲，則該類別為 FN(False Negative)。

表3. 18 情感分數評估指標矩陣

	隔天股價上漲	隔天股價下跌
情感分數為正數	TP	FP
情感分數為負數	FN	TN

準確率(Accuracy)：在所有類別中，正確判斷的比率。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

精確率(Precision)：所有情感分數為正的類別中，股價上漲的比率。

$$Precision = \frac{TP}{TP + FP}$$

召回率(Recall)：股價上漲的情況下，情感分數為正的比率。

$$Recall = \frac{TP}{TP + FN}$$

F1(F1-score)：綜合上述兩種指標評估情感模型。

$$F1 - score = \frac{2 * Precision * Recall}{Precision + Recall}$$

3.6.2 模型評估指標

使用均方根誤差(Root Mean Squared Error, RMSE)作為模型評估指標，評估預測成果。

RMSE 可以透過預測資料與真實資料間的差距評估預測效能，算法為：

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2}$$

其中 m 為預測資料總數，y_i為真實資料，即真實收盤價；$\hat{y_i}$為預測資料，即預測收盤價。

與情感分數相同，使用準確率(Accuracy)、精確率(Precision)、召回率(Recall)、F1-score 做為模型預測結果的評估指標，評估模型預測漲跌幅與股市漲跌幅之間的關係。

若預測股價上漲且隔天股價上漲，則該類別為 TP(True Positive)；
若預測股價下跌且隔天股價下跌，則該類別為 TN(True Negative)；
若預測股價上漲且隔天股價下跌，則該類別為 FP(False Positive)；
若預測股價下跌且隔天股價上漲，則該類別為 FN(False Negative)。

表3. 19 模型評估指標矩陣

	隔天股價上漲	隔天股價下跌
預測股價上漲	TP	FP
預測股價下跌	FN	TN

準確率(Accuracy)：在所有類別中，正確判斷的比率。
精確率(Precision)：所有預測隔天股價上漲的類別中，隔天股價上漲的比率。
召回率(Recall)：隔天股價上漲的情況下，預測隔天股價上漲的比率。
F1(F1-score)：綜合上述兩種指標評估情感模型。

3.6.3 模擬投資

預測股價趨勢的目的是要在實際投資上獲利，股票的獲利可以分為股利及售出股本的獲利，本研究以買低賣高來賺取價差的目標作短期交易，企圖僅以售出股本的方式獲利而不考慮股利，進行投資模擬以評估模型預測表現。在台灣證券交易中，買進股票時須花費買入價的0.1425%手續費，賣出股票時須花費賣出價的0.3%交易稅與0.1425%手續費，因此需考量手續費與交易稅的成本，例如使用10,000元買一張股價10元的股票，則賣出時股價至少需要超過約10.06才能夠獲利。

利用各模型類別進行2021/1/1~2021/12/31為期一年的模擬投資，每一天取得今天的所有特徵數值後，預測明天的股價，並於明天採取行動，過程中僅會持有一支股票(1000股)，或是沒有持有股票這兩種狀況。股票委託買賣則以最高價≥委買價格≥最低價為交易成功，超出範圍則交易失敗。委託買賣價則會依照台灣證券交易所公告之台股各商品升降單位對照表[21]出價。

表3. 20 台灣股票升降單位對照表

最低股價	最高股價	股票升降單位
0.01元	10元	0.01
10元	50元	0.05
50元	100元	0.10
100元	500元	0.5
500元	1000元	1.00
1000元	以上	5.00

舉例說明，台積電在2017/01/03的股價為183.0元，其適用之股價升降單位為每次0.5元，隔天能以每股183.5元出價購買台積電；台積電在2021/12/29的股價為628.0元，其適用之股價升降單位為每次1.00元，隔天能以每股629.0元出價購買台積電。

模擬投資參考[10]，當持有股票時，若模型預測為上漲，則在隔日的開盤買進一張(1000股)的台積電股票。當持有股票時，若模型預測為下跌，則於隔日開盤時賣出台積電股票。若在交易截止日的最後一天（2021/12/29）仍持有股票，則會於當日的開盤價賣出，清空所有股票。

模擬投資的投資成果將會使用績效報表[22]，績效報表如下表：

表3. 21 投資成果績效報表

項目	說明
年份	交易年份
總損益	總獲利、總虧損之總和
交易總次數	買入賣出為一次，顯示該年份交易總次數
交易產生總費用	買入手續費、賣出手續費、賣出交易稅之總和
勝率	買入賣出為一次， $\text{勝率} = \frac{\text{獲利次數}}{\text{總次數}}$
最大損失	年度中虧損最大的一次
最大獲利	年度中獲利最大的一次
投資報酬率	$\text{投資報酬率}\% = \frac{\text{總損益}}{\text{總投入資本}} \text{ x } 100\%$ 總投入資本為交易時最高的買進價格。
平均交易報酬率	$\text{平均交易報酬率}\% = \frac{\text{投資報酬率}\%}{\text{總次數}}$

3.7 模型訓練系統

執行機器學習任務往往會花費許多時間，因此提高效率是必須的。本研究使用 Tensorflow 執行機器學習訓練，每一次執行會跑約略5000多個模型，接著根據數據結果再進行模型調整，例如模型跑完一批量 learning rate=0.00001但其餘參數皆不同的模型後，觀察這樣的學習率下是否有利於模型收斂再進行微調。因此使用平行式擴充運算效能，縮短此次微調至下一次微調的時間以便更快的找出最佳模型。

本研究採用多進程與分散式的方式訓練模組，總共使用3台電腦，同時間可以執行至少8個進程，每個進程各訓練1個模型，亦即同時間可以訓練8個模型以增進找出最佳模型的速度。

3.7.2 多進程訓練模組

本研究同時用同一個腳本執行三個程式，由於每執行一個程式，也就是每生成一個進程，Tensorflow 就會有記憶體流失(memory leak)的狀況，如果使用單個進程長時間執行或是同時多個進程執行機器學習訓練任務記憶體會不堪使用造成程式崩潰，記憶體使用情形如下圖：

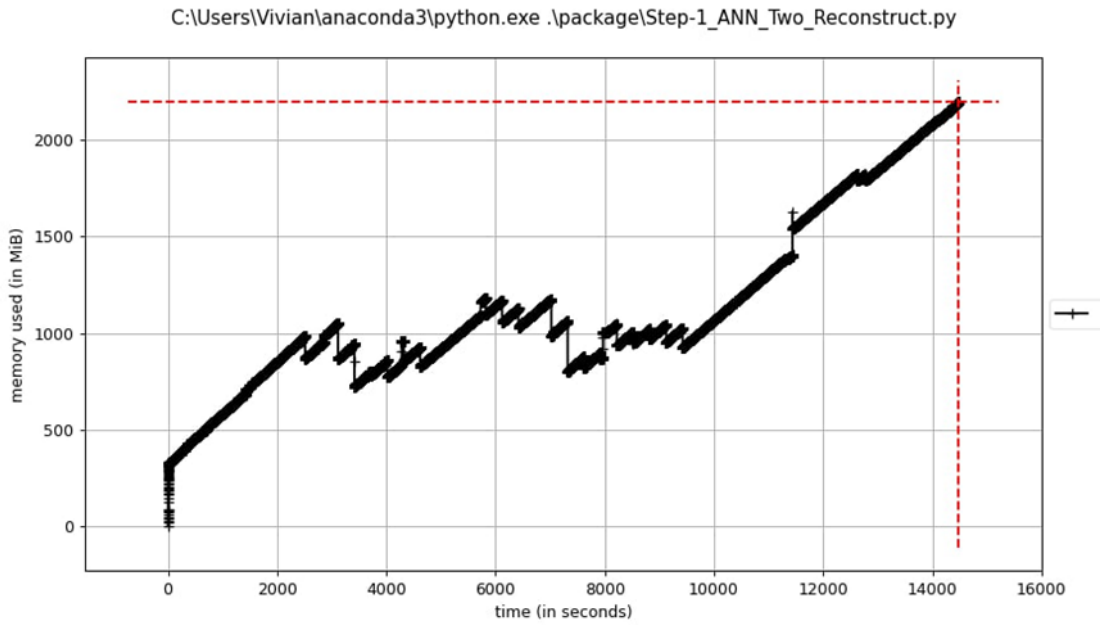

圖3. 4 記憶體流失情形

橫軸代表程式執行的次數，縱軸代表記憶體使用量，可以明顯看出有記憶

體流失的狀況。

為了解決記憶體流失的問題，需要使用 Python 中的 multiprocessing 套件，將訓練模型的程式部分額外再生成一個子進程，進程結束後就會釋放記憶體，如圖所示：

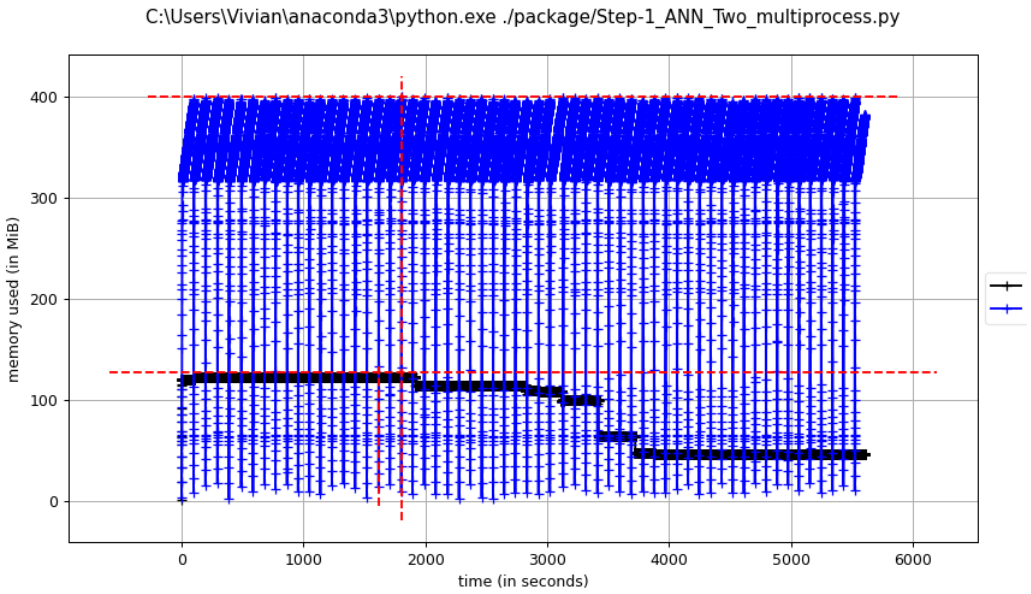

圖3. 5 改善記憶體流失

橫軸代表程式執行的次數，縱軸代表記憶體使用量，黑線代表主程式，藍線代表新產生的訓練模型進程。從圖中可以看出每一次訓練完模型後子進程所使用的記憶體就會被釋放。

每一個主進程所輸出的模型參數與訓練結果皆會被寫入屬於各自進程的 csv 檔儲存，避免多個主進程同時讀寫一個檔案造成髒讀、資料遺失的情況。透過觀察工作管理員中三台電腦執行機器學習任務的情形，電腦1最多可以同時執行3個主進程、電腦2最多可以同時執行3個主進程、電腦3最多可以執行2個主進程。

程式每一次的模型參數 random_seed、隱藏層參數都是隨機的，為了避免多個進程重複訓練相同參數的模型，使用 windows 系統內建的排程工具 schtasks，設定每30分鐘會將儲存各進程訓練結果的檔案資料統一匯入到一個同樣是 csv 檔的總儲存檔案，使程式每一次訓練模型前都會先檢查在總儲存檔案中該組參數

是否有被訓練過，如果訓練過則更換參數；在多進程訓練模組中皆使用 csv 檔儲存資料的目的是為了要讓其更容易部署在不同電腦上。

3.7.3 分散式訓練模組

將多進程訓練模組部署在多台電腦，以達到平行擴充運算效能的目的。本研究在電腦1架設關聯式資料庫 SQL Server，建立一張儲存所有結果的最終結果表，與專門儲存來自各電腦的資料表。多進程訓練模組的排程任務在匯合資料至總檔案後，會再將總檔案透過 sqlalchemy 套件上傳至 SQL Server 上專屬於這台電腦的資料表。SQL Server Agent 再以每30分鐘一次的頻率將多台電腦的資料從各自的資料表上匯入最終結果表。

與多進程訓練模組同理，為了避免多台電腦重複訓練相同的參數模型，讓程式每次訓練模組前都會先檢查該組參數是否存在最終結果表，如果訓練過則更換參數。

使用 SQL Server 的好處是用 SQL 快速的查到想要的值，並且在進程中檢查是否已經是訓練過的參數組時，只要用 sqlalchemy 對 SQL 做資料查詢即可，其查詢速度是非常快的，且查詢後傳遞資料較少，可以更快的在各電腦間傳輸。

表3. 22 排程任務

訓練模組	排程工具	任務頻率	任務內容
多進程訓練模組	Windows schtasks	30 分/次	匯合多進程的訓練結果並去除重複值，上傳至 SQL Server 中專屬表。
分散式訓練模組	SQL Server Agent	30 分/次	匯合多張表的訓練結果並去除重複值，匯合到最終結果表。

結合多進程與分散式訓練方法後，完整的訓練模型系統架構圖如圖3.6所示。一台電腦運行多個進程，每一個進程產生一個結果檔案；透過 windows 內建排程工具 schtasks 排程任務，每30分鐘集合一次來自各 process 的結果，並寫入位於電腦1上的 SQL Server 屬於各電腦的 SQL table 中；將各 SQL table 透過 SQL Server Agent 排程任務每30分鐘集合一次來自各 SQL table 的結果，存入紀錄所有結果的 final SQL table 中。

在程式執行的過程當中，各 process 會在訓練模型之前檢視圖中3個紅色標示的紀錄，查看是否該參數有被訓練過，如果訓練過則更換下一個參數組合。

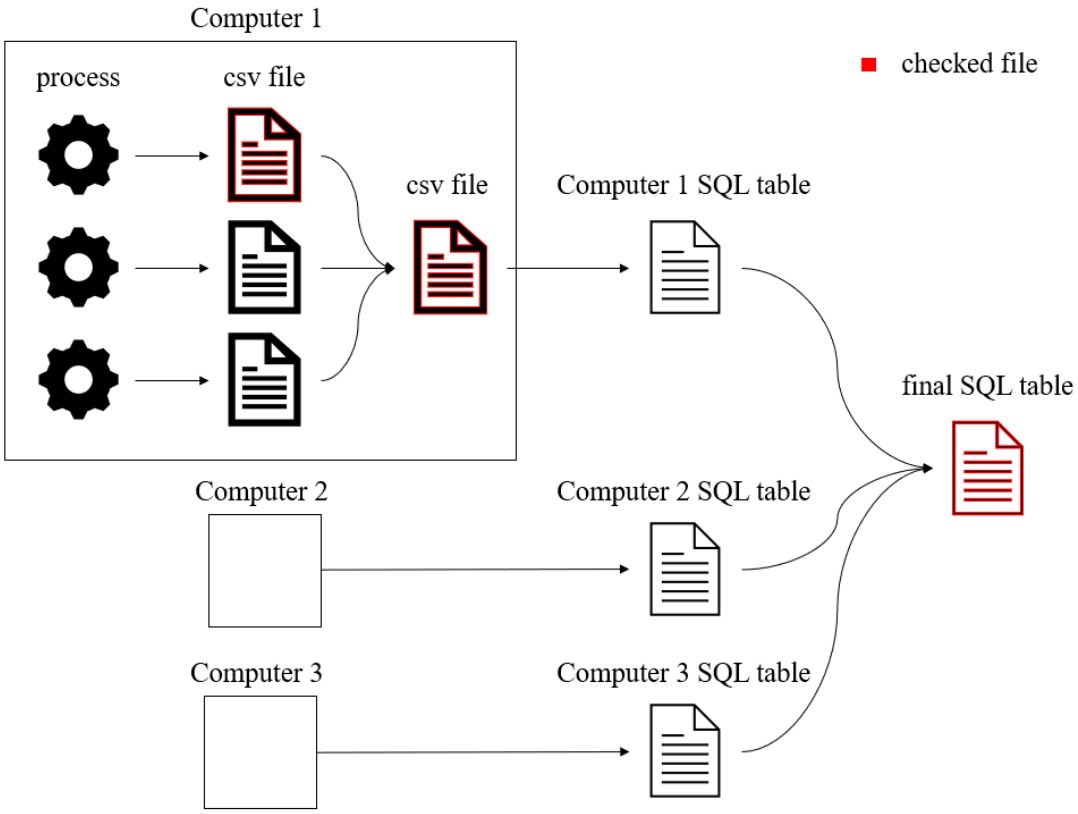

圖3. 6 多進程訓練模組與分散式訓練模組架構圖

第四章　實證結果與分析

4.1 實驗設計

本研究實驗有三個部分：

實驗一：情感分數與股價漲跌關係研究。

實驗二：使用傳統法資料集進行模型的訓練、股價預測與效益評估。

實驗三：使用滑動視窗法資料即進行模型的訓練、股價預測與效益評估。

4.2 實驗環境

由於本研究的機器學習模型不需要複雜的模型架構即可有良好的訓練成果，經過電腦2實測發現，CPU 訓練模型約150秒，GPU 訓練模行約500秒，使用 CPU 訓練模型相較於使用 GPU 訓練模型的速度還要快至少3倍，因此所有電腦皆使用 CPU 訓練模型。

表4. 1 軟體與硬體設備—電腦1

電腦1	
設備與環境	說明
中央處理器	Intel® Core™ i7-9700
記憶體	16GB RAM
顯示卡	-
作業系統	Windows 10
開發環境	Python 3.8
Tensorflow	2.8.0
可運行進程數量	3

表4. 2 軟體與硬體設備—電腦1

電腦2	
設備與環境	說明
中央處理器	Intel® Core™ i7-6700
記憶體	16GB RAM
顯示卡	GeForce GTX 1060 3GB
作業系統	Windows 10
開發環境	Python 3.8
Tensorflow	2.8.0
可運行進程數量	3

表4. 3 軟體與硬體設備—電腦1

電腦3	
設備與環境	說明
中央處理器	Intel® Core™ i5-8250U
記憶體	8GB RAM
顯示卡	GeForce GTX 1060 3GB
作業系統	Windows 10
開發環境	Python 3.8
Tensorflow	2.9.1
可運行進程數量	2

4.3 實驗一：情感分數與股價漲跌關係研究

4.3.1 新聞篇數統計

使用自製爬蟲程式取得的新聞自2017/01/01~2021/12/31總共12796筆，當中來自時報資訊共6554筆、中央社共3524筆、Moneydj 理財網共1721、財訊快報共511筆、東森財經共176筆、鉅亨網共219筆、其他來源91筆。

4.3.2 新聞情感分數預測效益

使用字典法計算出新聞的情感分數後，情感分數為正數即代表使用字典法預測明天的股價應是上漲；情感分數為負數即代表使用字典法預測明天的股價應是下跌。因此使用準確率、精確率、召回率、F1值評估使用字典法預測股價的表現。經過計算，字典法預測之成效如表4.4所示。

表4. 4 字典法預測成效

	準確率	精確率	召回率	F1
字典法	56.66%	95.57%	57.24%	71.60%

將情感分數使用散佈圖呈現，如圖4. 1 情感分數分佈圖，可以觀察到2017年到2021年情感分數的分布情形，總計1223天，平均在80.41分，最小值為-47，最大值為660。

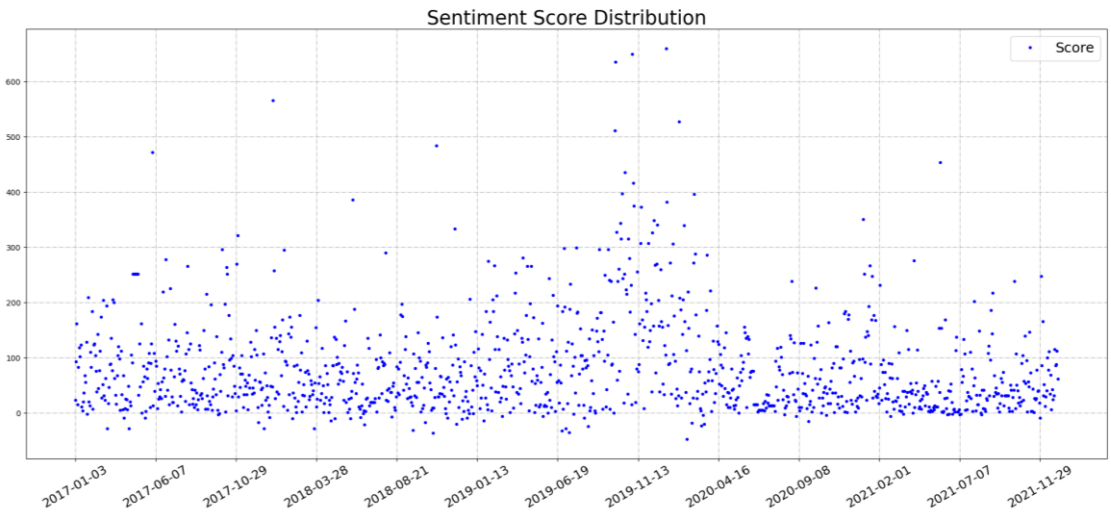

圖4. 1 情感分數分佈圖

圖4. 2 正確預測的情感分數分佈圖將正確預測的情感分數標示，透過散佈圖觀察預測正確之情況分布。

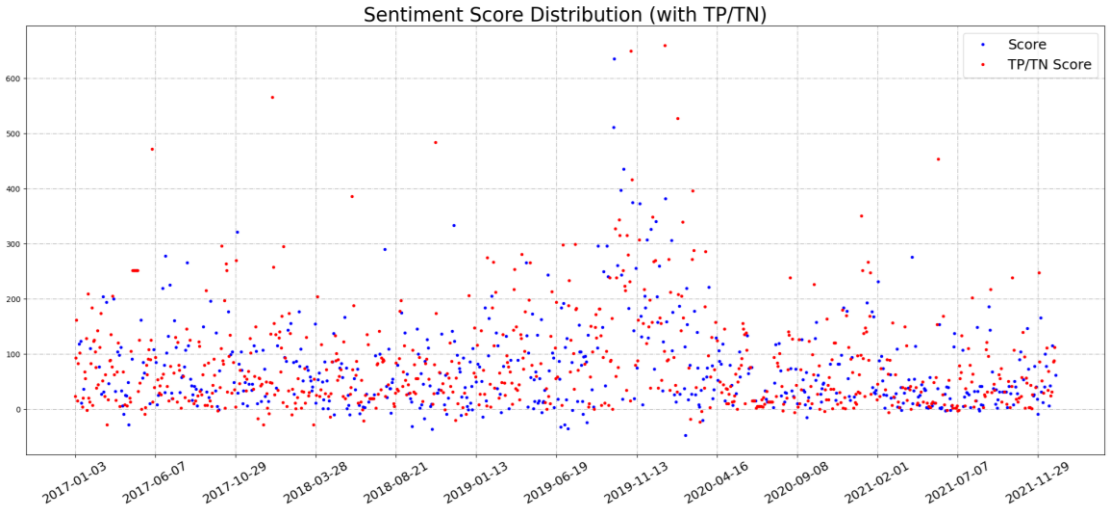

圖4. 2 正確預測的情感分數分佈圖

依照表4. 5 情感分數區間準確率分佈圖分數區間檢視情感分數的預測效能，可以看出準確率在分數區間200~299、400~499、500~599、600~699表現較佳。

表4. 5 情感分數區間準確率分佈圖

情感分數區間	準確率	出現次數
600~699	66.67%	3
500~599	66.67%	3
400~499	80.00%	5
300~399	50.00%	20
200~299	67.95%	79
100~199	54.96%	242
0~99	56.83%	815
-99~-1	45.10%	56
總計		1223

4.4 實驗二：傳統法資料集進行模型訓練

4.4.1 傳統法-MLR 模型

使用 sklearn 套件中的 LinearRegression 做訓練，輸入的參數特徵數23個、copy_X 為 True、fit_intercept 為 True、n_jobs 為 None、normalize 為 False。

訓練成果如圖4. 3 傳統法-MLR 股價趨勢預測圖所示：

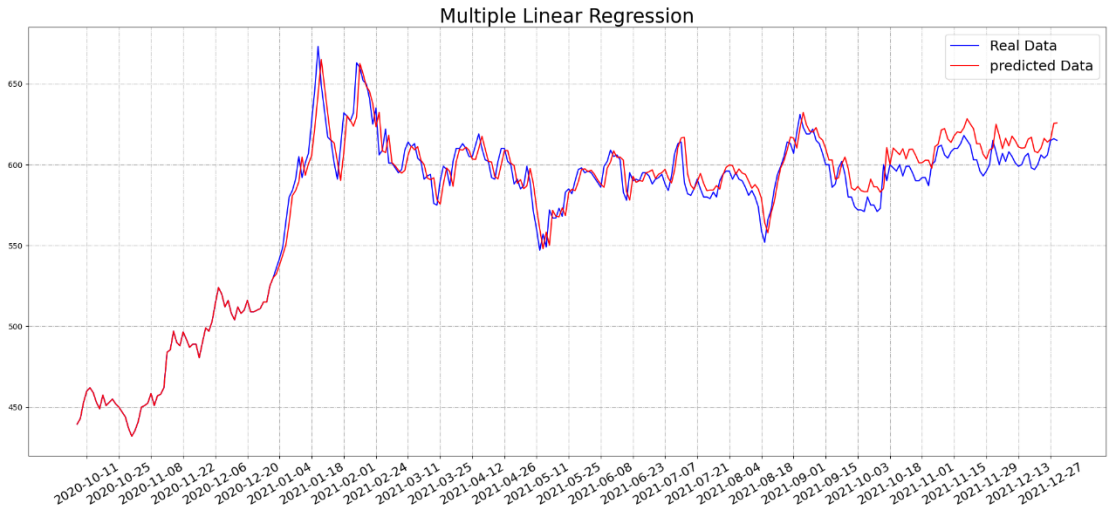

圖4. 3 傳統法-MLR 股價趨勢預測圖

傳統法-MLR 的預測成效如表4. 6 傳統法-MLR 預測成效所示。

表4. 6 傳統法-MLR 預測成效

	RMSE	準確率	精確率	召回率	F1
傳統法-MLR	10.2498	70.08%	57.03%	80.22%	66.67%

4.4.2 傳統法-ANN 模型

本研究使用多組參數輸入機器學習訓練，當中最好的模型參數如下所示：輸入23個特徵資料，並經過兩層隱藏層運算後，輸出層輸出一個目標值，該目標值即預測的收盤價；隱藏層第一層的神經元個數為40，隱藏層第二層的神經元個數為58，每一層皆使用激發函數 relu。

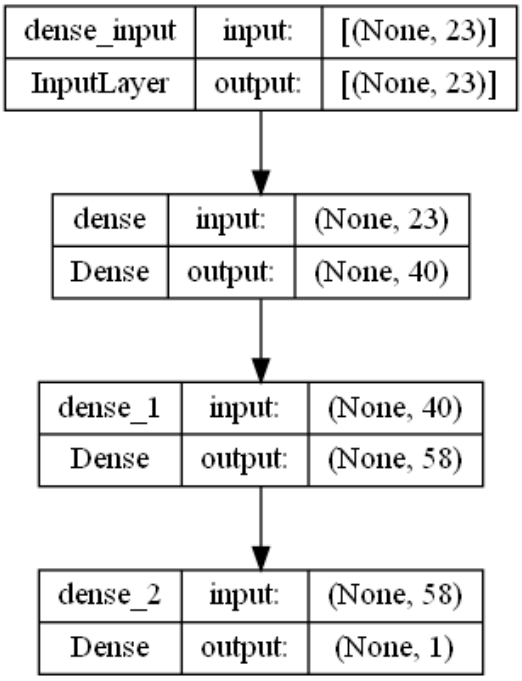

圖4. 4 傳統法-ANN 模型結構圖

傳統法-ANN 模型訓練參數如表4. 7 傳統法-ANN 模型參數所示，電腦1訓練該模型花費160秒。

表4. 7 傳統法-ANN 模型參數

參數名稱	參數值
seed	200
input_dim	23
layer1-units	40
activation	relu
Layer2-units	58
activation	relu
optimizer	SGD
loss function	mean_square_error
learning_rate	0.00001
decay	0
momentum	0.9
nesterov	True
batch_size	10
epochs	2000

傳統法-ANN 的預測結果如圖4. 5 傳統法-ANN 股價趨勢預測圖所示：

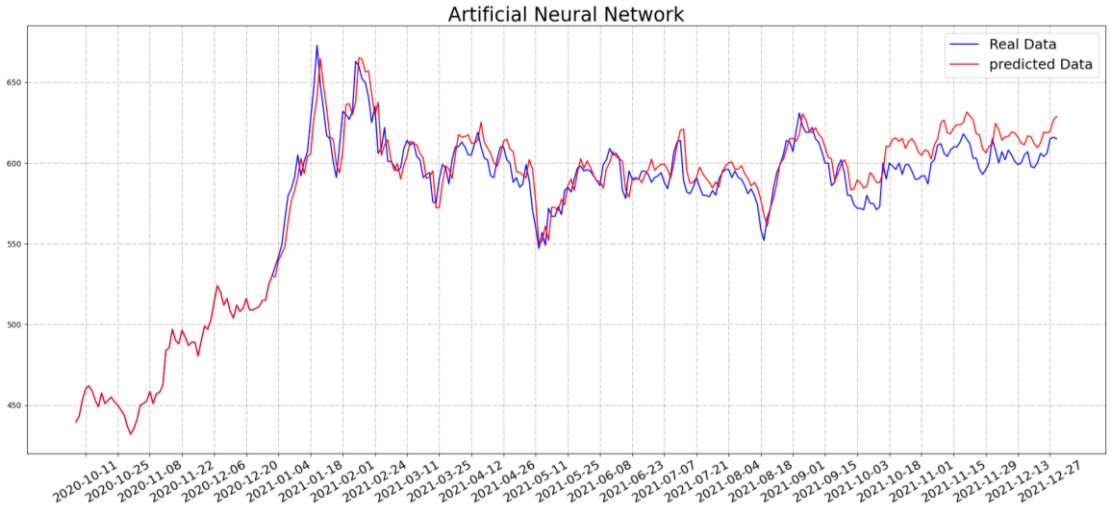

圖4. 5 傳統法-ANN 股價趨勢預測圖

傳統法-ANN 預測成效如表4. 8 傳統法-ANN 預測成效所示：

表4. 8 傳統法-ANN 預測成效

	RMSE	準確率	精確率	召回率	F1
傳統法-ANN	11.6379	68.03%	49.21%	82.89%	61.76%

4.4.3 模擬投資

將投資策略應用在2021年的資料集，形成模擬投資，試算使用實驗二之模型在2021年購入台積電股票的損益、報酬率等。

表4. 9 傳統法-MLR 模擬投資績效報表

	傳統法-MLR	傳統法-ANN
年份	2021年	2021年
總損益	114,837	(-103,851)
交易總次數	15	19
交易產生總費用	(-24,985)	(-32,329)
勝率	66.67%	26.32%
最大損失	(-11,982)	(-28,958)
最大獲利	33,952	50,927
投資報酬率	18.79%	(-16.38%)
平均交易報酬率	1.25%	(-0.86%)

表4.6為投資台積電的績效結果，使用傳統法-MLR 預測出來的模型可以獲利並且獲利114,837元；而使用傳統法-ANN 不能獲利且虧損103,851元。傳統法-MLR 交易次數較傳統法-ANN 少，交易產生的總費用也較少，勝率多了兩倍以上，其最大損失是由於在2021/05/11預測股票會上漲，用579元買入股票，但當天股票實則下跌到560元，隨後旋即在2021/05/12預測股票會下跌以567元賣出股票，當天股票確實下跌至547元，意味著因為買入時的預測失誤而造成虧損；其最大獲利是在2021/01/14以587元買進股票，當天股票確實上漲至601元，於下週一2021/01/17以621元賣出，但當天股票時則上漲至607元，意味著股票尚呈現漲勢但卻提早賣出，並且剛好在2021/01/17用高額開盤價賣出，因而有高收益的成果。而傳統法-ANN 則各項表現都遜於傳統法-MLR，其最大損失是由於在2021/03/31以596買進股票後，一直到2021/05/11都預測明天會漲，預測的價格有起伏可是都比實際價格還高，顯示這段期間內預測值全部高估最後在2021/05/12賣出；最大獲利則是於2021/05/13買進547元，跟最大損失一樣，直到賣出前的預測幾乎都屬於高估的，直到2021/06/01以598元賣出。

4.5 實驗三：滑動視窗法資料集進行模型訓練

4.5.1 使用 ANN 模型

輸入23個特徵資料，並經過兩層隱藏層運算後輸出一個目標值，該目標值即預測的收盤價；隱藏層第一層的神經元個數為21，隱藏層第二層的神經元個數為151，每一層皆使用激發函數 relu。

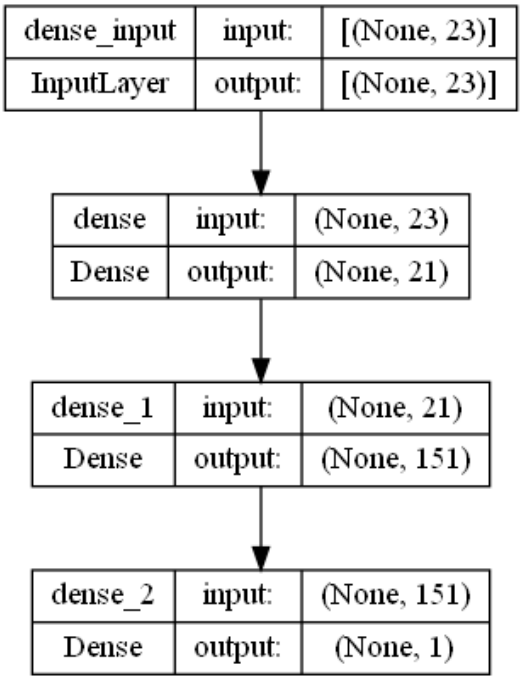

圖4. 6 視窗法-ANN 模型結構圖

視窗法-ANN 模型訓練參數如 表4. 10 視窗法-ANN 參數 所示，電腦3訓練該模型平均每期花費402.9秒。。

表4. 10 視窗法-ANN 參數

參數名稱	參數值
seed	39
input_dim	23
layer1-units	21
activation	relu
Layer2-units	151
activation	relu
optimizer	SGD
loss function	mean_square_error
learning_rate	0.000001
decay	0
momentum	0.9
nesterov	True
batch_size	10
epochs	4000

視窗法-ANN 第1期的預測結果如圖4. 7 視窗法-ANN 股價趨勢預測圖(第1期)所示：

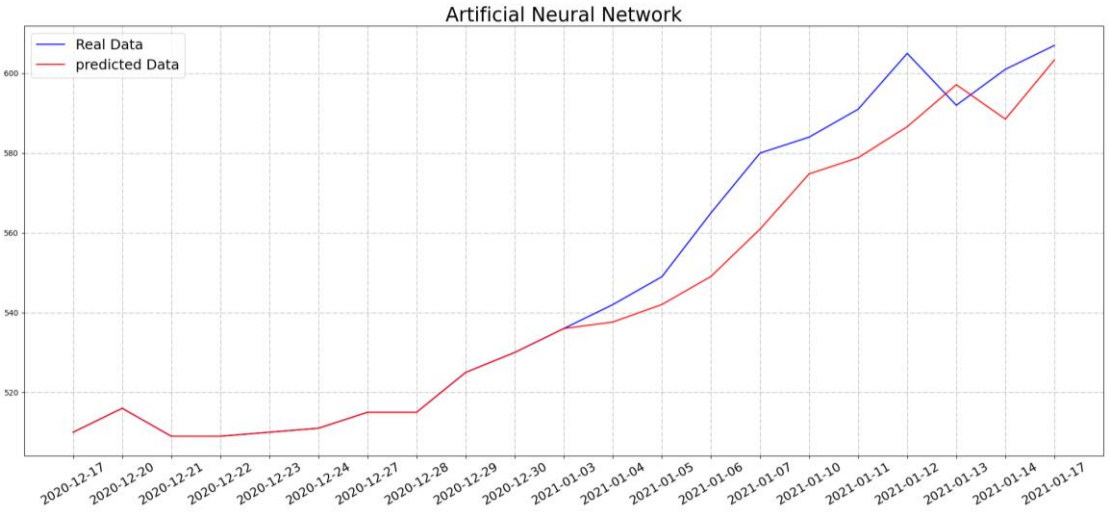

圖4. 7 視窗法-ANN 股價趨勢預測圖(第1期)

視窗法-ANN 第2期的預測結果如圖4. 8 視窗法-ANN 股價趨勢預測圖(第2期)所示：

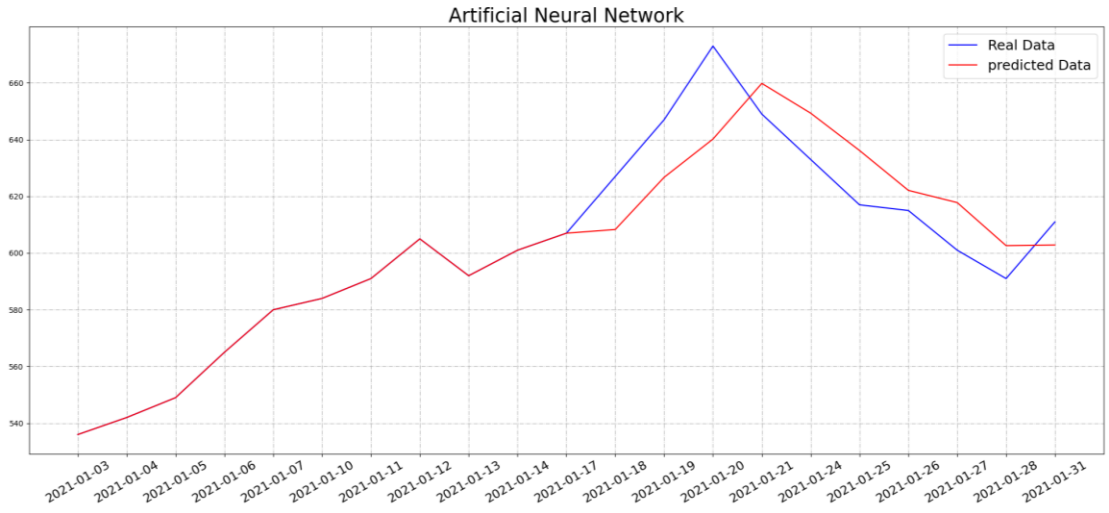

圖4. 8 視窗法-ANN 股價趨勢預測圖(第2期)

視窗法-ANN 全期數的預測結果如圖4. 9 視窗法-ANN 股價趨勢預測圖(全期數)所示：

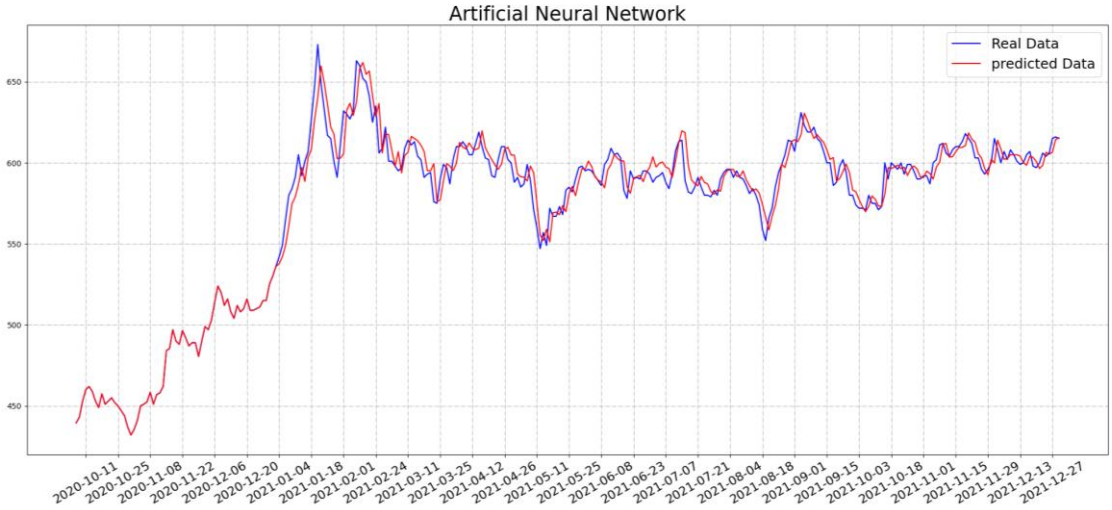

圖4. 9 視窗法-ANN 股價趨勢預測圖(全期數)

視窗法-ANN 全期數的預測成效如表4. 11 視窗法-ANN 預測成效所示。

表4. 11 視窗法-ANN 預測成效(全期數)

	RMSE	準確率	精確率	召回率	F1
視窗法-ANN	9.1557	75%	71.88%	78.63%	75.1%

視窗法-ANN 各期數的 RMSE 表現如表4. 12 視窗法-ANN 之 RMSE 成果：

表4. 12 視窗法-ANN 之 RMSE 成果(各期數)

期數	RMSE	期數	RMSE
1	12.05	14	5.24
2	17.68	15	4.06
3	14.47	16	9.54
4	12.2	17	6.93
5	10.39	18	8.81
6	7.95	19	7.3
7	7.85	20	3.93
8	10.66	21	5.81
9	10.54	22	5.07
10	4.89	23	6.33
11	9.71	24	4.42
12	6.76	25	4.98
13	12.43		

4.5.3 模擬投資

將投資策略應用在2021年的資料集，形成模擬投資，試算使用實驗三之模型在2021年購入台積電股票的損益、報酬率等，並為了將結果與實驗二中資料集切割方法不同但同樣使用 ANN 模型的傳統法-ANN 做比較，同樣將傳統法-ANN 呈現在表4. 13 視窗法-ANN 模擬投資績效報表中。

表4. 13 視窗法-ANN 模擬投資績效報表

	視窗法-ANN	傳統法-ANN
年份	2021年	2021年
總損益	(-339,651)	(-103,851)
交易總次數	53	19
交易產生總費用	(-89,355)	(-32,329)
勝率	20.75%	26.32%
最大損失	(-46,933)	(-28,958)
最大獲利	33,952	50,927
投資報酬率	(-51.15%)	(-16.38%)
平均交易報酬率	(-0.97%)	(-0.86%)

兩種模型表現皆不佳，視窗法-ANN 的投資報酬率甚至更遜於傳統法-ANN，虧損比傳統法高出三倍之多。以視窗法的最大損失來看，其在2021/01/25以642元買入，但當天股價實際是跌到617元，後2021/02/01以595元賣出，但當天股價實際是漲到632元，也就是當天應預測為上漲卻預測為下跌，又剛好開盤時買在股市最低點；最大獲利為2021/02/02以629元買入，預測為上漲實則下跌，後2021/02/17以663元賣出，預測為下跌而實際上也下跌。

第五章　結論

5.1 研究結論

傳統法資料集與視窗法資料集實驗結果比較：

本研究將多個特徵輸入資料集，再以不同切割資料集的方法輸入 ANN 模型做訓練，一般來說使用傳統法會因為只有訓練過訓練集的資料而到後期預測資料會越不準確，如果過度準確又有過度擬合之疑慮，而股價資訊屬於時間序列之資料，會因為時間的不同數據又產生變化，例如傳統法的訓練資料集涵蓋的時間範圍為2017~2020年，但在2021年初開始爆發新冠肺炎疫情，傳統法卻沒有辦法及時更新疫情的影響資訊。因此想要使用滑動視窗法解決傳統法訓練資訊無法更新的問題。但就研究結果發現，雖滑動視窗法的 RMSE 較傳統法的 RMSE 少2.4822、準確率多6.97%、精確率多22.67%、F1表現多13.34%，看似成效皆較好，甚至是3種模型中預測成效最好的，但在模擬投資評估時，其投資報酬率卻是(-51.15%)，跟傳統法比虧損三倍之多，產生虧損的原因應是(1) RMSE 仍舊過高，需要 RMSE 表現更低的模型；(2) 字典法的預測成效表現不佳，情感分數準確率僅有56.66%，因為情感分數是模型特徵之一，可能因此受到該特徵影響；(3) 投資策略不適用該模型。

5.2 研究貢獻

本研究將股價資訊、分析指標、以及使用字典法計算出的情感分數投入模型，實測這組特徵組合是否能在多元線性迴歸模型、人工神經網路模型上有良好的預測結果。

5.3 研究限制

本研究進行文句特徵處理時，因為將文句以關鍵字、股票名稱作為提取的準則，有時會發生一句話中又提及競爭對手做比較的內容，在此情況下會無法辨別主要是在敘述目標公司表現良好、亦或競爭對手表現良好，可能造成誤判的情況。

計算情感分數需要足夠的新聞數量，因此本研究亦受到目標公司的新聞報導量限制，如果樣本數不足造成情感分數大多數空值，將會影響情感分數對於股價預測的準確率。

5.4 未來研究建議

本研究認為，視窗法-ANN 的預測成效是三個模型中表現最好的，可是其在模擬投資的表現卻是三個模型中最差的，原因也許是投資策略不適用該模型，可以嘗試添加今日與預測價差容忍度提高獲利可能性，例如價差在10元以內不做任何投資動作。

參考文獻

[1] 歷年股票市場概況表 URL: https://www.twse.com.tw/zh/statistics/statisticsList?type=07&subType=232

[2] 投資人類別交易比重統計表 URL: https://www.twse.com.tw/zh/statistics/statisticsList?type=07&subType=262

[3] 年度上市公司資本來源明細表年報 URL: https://www.twse.com.tw/zh/statistics/statisticsList?type=07&subType=257

[4] 黃巧雯，“台灣政治循環下股票市場投資組合績效之探討”，國立高雄第一科技大學，高雄市。

[5] M. Costola, M. Nofer, O. Hinz, and L. Pelizzon, “Machine Learning Sentiment Analysis, Covid-19 News and Stock Market Reactions,” *SSRN Electron. J.*, 2020.

[6] X. Kewei and L. Yuanyuan, “A-share Stock Reactions to the Approval of COVID-19 Vaccine Clinical Trial: An Event Study Model of Listed Pharmaceutical Firms’ Returns,” *2020 2nd International Conference on Economic Management and Model Engineering (ICEMME)*, pp. 404–407, Nov. 2020, Chongqing, China.

[7] J. Liu, Z. Lu, and W. Du, “Combining Enterprise Knowledge Graph and News Sentiment Analysis for Stock Price Prediction,” Hawaii International Conference on System Sciences, 2019.

[8] A. Chatterjee, H. Bhowmick, and J. Sen, “Stock Price Prediction Using Time Series, Econometric, Machine Learning, and Deep Learning Models,” *2021 IEEE Mysore Sub Section International Conference (MysuruCon)*, pp. 289–296, Oct. 2021, Hassan, India.

[9] P.-C. Lan, W.-L. Kung, Y.-L. Ou, C.-Y. Lin, W.-C. Hu, and Y.-H. Wang, “Machine learning model with technical analysis for stock price prediction: Empirical study of Semiconductor Company in Taiwan,” *2019 International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS)*, pp. 1–2. Dec. 2019, Taipei, Taiwan.

[10] 邱彥誠，“應用人工智慧於股市新聞與情感分析預測股價走勢”，國立臺北大學資訊管理研究所碩士論文，新北市。

[11] Q. Li, T. Wang, P. Li, L. Liu, Q. Gong, and Y. Chen, “The effect of news and public mood on stock movements,” *Inf. Sci.*, vol. 278, pp. 826–840, Sep. 2014.

[12] Y. Hu, K. Liu, X. Zhang, L. Su, E. W. T. Ngai, and M. Liu, “Application of evolutionary computation for rule discovery in stock algorithmic trading: A literature review,” *Appl. Soft Comput.*, vol. 36, pp. 534–551, Nov. 2015.

[13] Z. D. Aksehir and E. Kilic, “Prediction of Bank Stocks Price with Reduced Technical Indicators,” *2019 4th International Conference on Computer Science and Engineering (UBMK)*, pp. 206–210, Sep. 2019, Samsun, Turkey.

[14] V. Kalyanaraman, S. Kazi, R. Tondulkar, and S. Oswal, “Sentiment Analysis on News Articles for Stocks,” *2014 8th Asia Modelling Symposium*, pp. 10–15, Sep. 2014, Taipei, Taiwan.

[15] Fugle. URL: https://www.fugle.tw/

[16] FinMind. URL: https://finmind.github.io/

[17] jieba. URL: https://github.com/fxsjy/jieba

[18] CkipTagger URL: https://github.com/ckiplab/ckiptagger

[19] NTUSD. URL: http://nlg.csie.ntu.edu.tw/download.php

[20] R. I. Rasel, N. Sultana, and N. Hasan, “Financial instability analysis using ANN and feature selection technique: Application to stock market price prediction,” *2016 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pp. 1–4, Oct. 2016, Dhaka, Bangladesh.

[21] 台股各商品升降單位對照表 URL: https://www.twse.com.tw/zh/page/products/trading/introduce.html

[22] 李家瑋，“基於 C-RNN-GAN 神經網路的股票價格趨勢預測模型之研究-以美國股票市場為例”，私立輔仁大學資訊管理研究所碩士論文，新北市。