國立雲林科技大學資訊管理系

碩士論文

Department of Information Management

National Yunlin University of Science & Technology

Master Thesis

利用 BERT 於中大型股票新聞資料探勘以探討其股價漲跌情形-以台積電為例

Use BERT to Mine Medium and Large-Cap Stock News Data to Explore The Rise and Fall of Its Stock Price - Taking TSMC as An Example

阮宗墉

Zong-Yong Juan

指導教授：陳重臣 博士

Advisor:Jong-Chen Chen,Ph.D.

中華民國 112 年 1 月

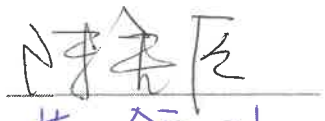
January 2023

國立雲林科技大學
碩士班學位論文考試委員會審定書
National Yunlin University of Science and Technology
Thesis Oral Defense Approval Form

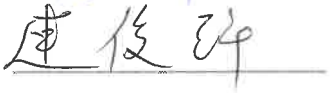
本論文係 阮宗墉 君在本校 資訊管理系碩士班 所提論文 利用BERT於中大型股票新聞資料探勘以探討其股價漲跌情形-以台積電為例 碩士資格水準，業經本委員會評審認可，特此證明。

The student JUAN,ZONG-YONG enrolled in the Master's program in Department of Information Management has satisfactorily passed the oral defense of the thesis Use BERT to Mine Medium and Large-Cap Stock News Data to Explore The Rise and Fall of Its Stock Price - Taking TSMC as An Example.

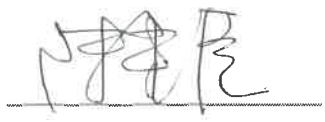
口試委員：
Oral defense committee members

陳重臣 

黃錦法 

連俊瑋 

指導教授：
Advisor(s)

陳重臣 

所長：
Dean of Graduate Institute 

中　華　民　國　111　年　12　月　24　日

摘要

在股票市場中所擁有的資訊量越多，則勝率越高。在目前的股票市場資訊，新聞扮演著相當重要的角色。相對的，如何從這些資訊中擷取重要的資訊更是一大難題。本研究使用新聞詞頻率機率化的方式來解決同一天存在多則新聞但只有一個漲跌訊號的問題，模型使用 Transformers 的 Encoder 架構所訓練出的 Bert 預訓練模型對股票新聞與股價進行漲跌預測的分類任務。研究分為兩部分，分別是台積電新聞與中型股票新聞的漲跌預測。台積電新聞在研究中為測試資料前處理的方法是否有效，而後把有效的方法應用於中型股票新聞，且中型股票大多會遇到新聞資料量過少的問題，研究中使用皮爾森相關係數找出正相關的股票群對資料集進行增強。

研究後發現，若是使用未經同一天單一漲跌標籤修正處理的台積電新聞資料集對 Bert 進行微調，會得到損失分數無法下降的失敗模型，但在修正後可以產出損失分數明顯下降且在測試集預測成功率達 0.69 的微調模型；使用相同方法於中型股票資料集中也有相同效果，部分中型股的資料增強不僅對於預測成功率的提升有幫助，同時也讓新聞量偏少的中小型股票有新的研究方式。

關鍵字：Bert、股票新聞、台積電、中型股票、漲跌預測、資料增強

ABSTRACT

This study uses the probability method of news word frequency to solve the problem that there are multiple news on the same day but only one rising and falling signal. The model uses the Bert pre-training model trained by Transformers' Encoder architecture to predict the rise and fall of stock news and stock prices in classification task. The study is divided into two parts, the up and down forecasts for TSMC news and mid-cap stock news. In the research, TSMC News tested whether the data preprocessing method was effective, and then applied the effective method to the news of mid-cap stocks, and most of the mid-cap stocks would encounter the problem of too little news data. In the study, the Pearson correlation coefficient was used to find the positive related stock groups enhance the dataset.

After research, it is found that if Bert is fine-tuned using the TSMC news data set that has not been corrected by a single rise and fall label on the same day, a failure model with a loss score that cannot be reduced will be obtained, but after the correction, the loss score can be significantly reduced and tested. A fine-tuning model with a prediction success rate of 0.69; using the same method in the data set of mid-cap stocks also has the same effect. The data enhancement of some mid-cap stocks is not only helpful for the improvement of the prediction success rate, but also allows small and medium-sized stocks with less news volume to have the same effect. New research methods.

Keywords: Bert, stock news, TSMC, mid-cap stocks, up and down forecasts, data augmentation

目錄

表目錄

圖目錄

第一章　緒論

1.1 研究背景與動機

當我們研究中小型股票時，常常會因為資訊量的不足而產生誤判或是不知道下手買賣的時機。資訊通常包含了技術面、籌碼面和消息面，若在股票市場中所擁有的資訊量越多，則勝率越高。舉兩個極端的例子，如果半點資訊都沒有，那勝率就是瞎猜來的；如果能預知未來，那肯定能有一個非常完美的勝率。而中小型股票雖然也有完善的技術面指標，但是新聞量遠不及大型股票，對此窘境勢必要有彌補或增強的作法。

常在股市網站看到股票連動性(stock price co-movement)這個詞，股票連動意指，當一檔股票漲或跌時，會連帶影響與其具有相關性的股票，其中連動的關係包含了類股連動、集團連動、概念股連動和國際股市連動。另外一個相近的名詞就是比價效應，兩者的意思差不多，在財經網站出現的頻率也不低。買股票前若有蒐集資訊的習慣，而蒐集的資料往往要追求廣度，例如買個股也可能會去參考相關族群的新聞或發布的消息，為的就是想避開同族群的利空而被拖累，或是看到同族群有利多消息而去買還沒起漲的股票想分一杯羹。例子比比皆是，如 2018 年左右的被動元件族群，由指標股國巨(2327)所帶動整個被動元件族群的瘋漲潮，當股民看到國巨的相關利多新聞包括缺料跟漲價…，他們也會思考如果指標股大漲那其他體質也不錯的相關個股是不是也會漲，進而引發這波被動元件潮流。

利用指標股的漲跌去跟進投資相關股票或是相關中小型股票，換句話說，手上缺乏資訊的中小型股也能利用指標股或是相關類股的資訊去類推漲跌。資訊除了數據面之外，還有就是新聞或是各種公告，近年來這種非結構化資料的研究日益被人們所重視，也就是文字資訊處理，文字資訊處理包含在自然語言處理(Natural Language Processing，NLP)的領域中。從早期的機率統計方式，像是簡單貝氏(Naïve

Bayes)、隱藏式馬可夫模型(Hidden Markov models)、CBOW、Skip-gram…，到現在很熱門的神經網路，像是循環神經網路(Recurrent Neural Network,，RNN)、長短期記憶(Long Short-Term Memory，LSTM)和 Transformers，也發展出了預訓練模型(Pre-Trained Model)如 ELMO、BERT、GPT3、T5…，這一切的技術迭代都讓自然語言處理的能力更加強大，取得上也更方便。

對於中小型股票常常會發生當天沒有新聞可以參考的狀況，在思考如何資訊增強(Data Augmentation)時，想到是否能利用其他具有相關性股票的新聞來加入預測的資料中是此篇論文的動機所在，而為了達到資料量的多多益善，同時要避免垃圾進垃圾出的狀況，則是這篇論文想深究的方向。而在面對同一天有很多新聞存在時，卻只能用當天的漲或跌統一上標，這也存在是不是所有新聞都是屬於正向或反向的迷思(夏鶴芸，2020)。

1.2 研究目的

i.使用網路爬蟲爬取的股票相關新聞與真實收盤價數據對其進行內文與標籤的處理，並解決新聞與股票多對一的問題。

ii.試驗何種資料前處理方式對台積電資料集所產出的 Bert 模型具有較好的預測力。

iii.於 Bert 模型中帶入台積電資料前處理的方式對中型股及相關股票資料集進行股市漲跌的預測並且分析資料探勘方式是否有效。

第二章 文獻探討

本章節介紹使用新聞預測股價的正當性，股價之間連動性的關係與 Bert 模型的架構介紹。

2.1 新聞對股票的影響與相關研究

財經新聞對於股票是否有影響性，又或是說資訊的揭露對於股市的影響，從很早以前就有相關研究，Eugene F. Fama 於 1970 年所提出的市場效率假說 (Efficient-market hypothesis， EMH)，該理論存在三個假設:

i. 市場價格會隨著最新資訊即時反應，所以股價呈現隨機走勢。

ii. 市場資訊為隨機性出現，即是好壞資訊是隨機出現的。

iii. 市場上的投資者都是理性而且追求最大利潤，每個投資者皆為獨立，不相互影響。

這三個假設衍伸出市場的三個性質，弱效率市場、半強效率市場跟強效率市場，簡單來說，在股票市場的相關資訊已經充分被利用，投資人無法使用資訊不對等的策略來進行套利行為，且投資人都相當理性，不會在風險異常的情況下追高殺低。該假說大概只活躍了十幾年，在 1980 年代後發現，其實投資人並不總是理性，也並不是所有的資訊都不存在套利空間。以台灣來說，投資者從事投資分析時，會將過去報酬的走勢當成重要考量(古金尚，2003)。

投資人對於股市的敏感度也會因為市場的狀態而有差異，在牛市時，投資人對於正面新聞的關注度會提高，反應也變快，而對於負面新聞的反應則較慢或是產生延遲；熊市時投資人則對於負面新聞敏感度提高。另外無論市場狀態，負面新聞的傳遞速度都會比較慢(王彥均，2017)。股票市場的狀態也會影響投資人對於新聞的注意力分佈，投資人會參考新聞，但存在因為市場狀態而產生的延遲。

台灣各家媒體的財經新聞對於股票市場表現存在顯著影響，不同媒體對股票價格的影響力不盡相同(王釗東，2017)，該研究對一年的財經新聞進行文字前處理後，使用支援向量機(Support Vector Machine，SVM)進行分析，這也說明利用新聞資訊來預測股票是有其正當性和研究性的。

大部分對於新聞與股票之間的相關性都是採用量化研究，蒐集大量新聞數據以文本分析來佐證股票漲跌，而差別只是在於使用的方法。本研究想要對新聞文本使用預訓練模型(pre-training model)進行情緒分析，並且用其結果來探討跟股價之間的關係。相關研究如夏鶴芸(2019)，使用 BERT 建立新聞分類模型，並且發現使用股價漲跌真實數據的準確度會比使用情緒字典來得好。

2.2 連動性(相關性)，相關研究

連動性就是相關性的研究，在金融市場中研究的範圍相當廣泛，有國與國之間的股市是否有相關，或是兩檔股票的比較，也有不同金融商品之間的相關性研究，如債券與股市、匯率與股市…，也有跨越市場的比較，像是房市與股市、油價或金價對於股市的相關性…，簡單來說就是兩種以上的組合探討其中的關係。在台灣股市或是新聞中，跟連動性(Co-movement)相關的詞彙還有外溢效應(Spillover effect)跟比價效應(Price Comparison effect)，這三個詞很常出現在台股新聞中。外溢效應指的是當某一事件或行為的出現，常會影響到其他事情的發展所產生的外部效果(吳青山，2018)，比價效應是同類型的公司，其生產的產品、公司的規模、流通股本、業績或是所面對的上下游廠商差不多的狀況下，例如同個產業中其中一間公司的股價開始因為獲利原因而大漲，市場會開始猜測是否其他同樣體質的公司也會受益。就名詞解釋來看，連動性與外溢效應所包含的範圍比較大，允許跨類股、跨市場或跨領域的相關研究，而比價效應所形容的通常都是具有同質的事物。

金融市場相關性的研究大部分探討關於投資組合的風險管控、市場或金融商品

互相影響的因果關係。在外資與投信法人持股比率變化對股價報酬率影響之研究(陳彥豪，2002)中使用相關係數(Correlation coefficient)分析外資與投信對於電子股的持股比率的相關性很高(係數為 0.8455)，可以看出這兩大法人對於電子股的投資策略是相近的。之後加入電子股的指數作相關係數可以發現，指數漲的時候外資跟投信的持股比率也是在增加，雖然這對於一般投資分析來說算是一個明顯易懂的現象，在股價上漲之前通常成交量能會慢慢湧現，而量能變化配上此分析也可以說明，在研究期間電子股的交易量通常是與投信跟外資有關的。

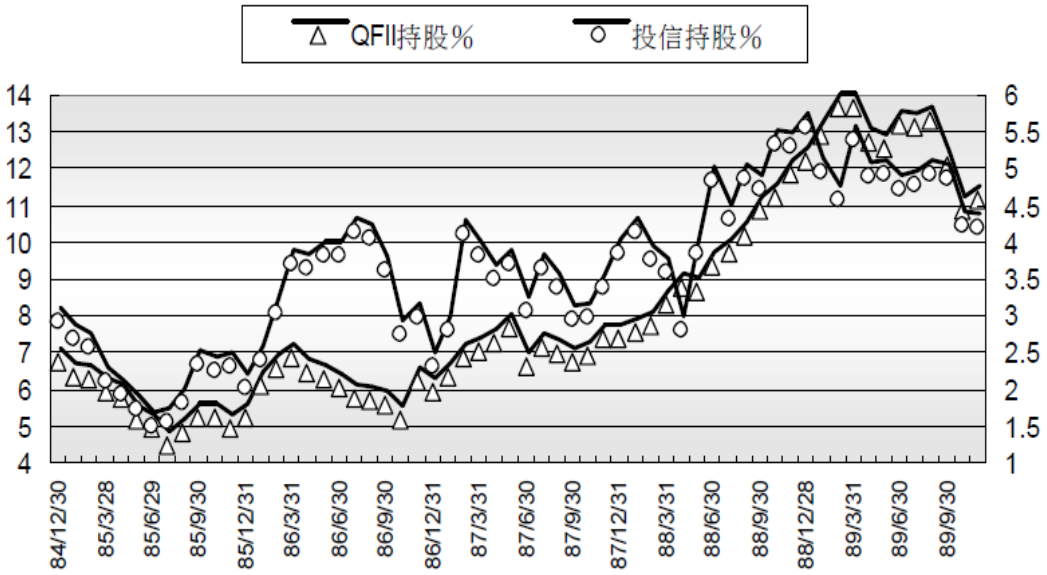


圖 1　外資與投信對整體上市電子股的持股比率

(資料來源: 陳彥豪(2002)。外資與投信法人持股比率變化對股價報酬率影響之研究-以上市電子股為例)

有關於相關係數，最常使用的是皮爾森積差相關係數(Pearson product-moment correlation coefficient)或稱皮爾森相關係數，主要是要找相關性，至於因果關係則要做其它更嚴謹的分析。Francois-Serge Lhabitant 於 2011 年發表的文章中指出，若是使用股價當日漲跌幅(日均漲幅)的數據進行皮爾森相關的驗證，那得到的係數或許會是負相關，以圖 2-2 為例，以每日平均來計算股票 1 與股票 2 的相關性(如圖 2

左圖)，那麼會得到一個負相關的係數，這時用的數據相當於是日均漲幅，兩者都有漲，只不過股票 1 漲贏平均值而股票 2 漲輸平均值，然而一正一負的後果就是負相關。又如果數據換成是股票的歷史平均數據的話(如圖 2 右圖)，兩檔股票大部份時間會呈現正正或負負，所得到的相關係數也就會是正相關了。這表示數據的如何選用也會影響到結果的呈現，而驗證的方法之一就是能否找到反例或是資料視覺化也可以看出一點端倪。

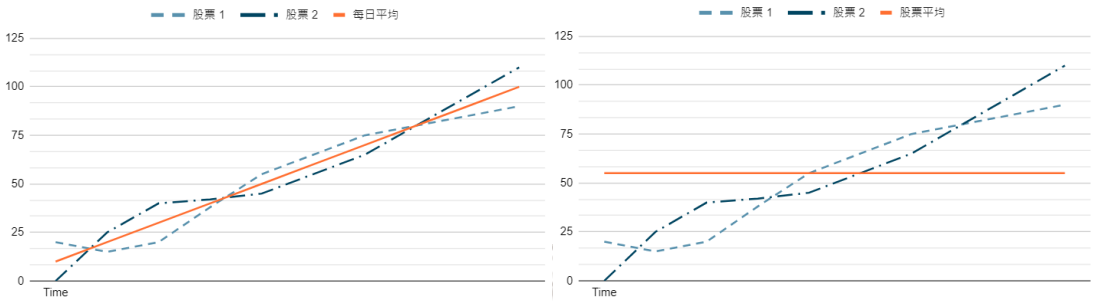


圖 2　對兩檔股票使用不同定義的平均

若是除了相關性之外還想知道更多時間序列資料的資訊，那就必須做其它的檢定，例如單根檢定(Unit root test)判斷是否為定態(Stationary)，Granger 因果關係檢定(Granger causality test)則可以判斷出資料中領先跟落後的關係。若是單根檢定的結果為非定態(Non-stationary)也沒有進行差分(Difference)和共整合檢驗(Cointegration Test)處裡的話，則後續的迴歸分析可能會產生假性迴歸(Spurious regression)。金價、銅價對道瓊工業平均指數與美國工業生產指數關係之研究(楊智欽，2021)中使用單根分析得到銅價為定態序列，金價、道瓊指數與美國工業生產指數為非定態序列，對其進行一階差分處理後皆呈定態序列以利後續研究，並且在 VAR 模型中得到最佳的落後期為 1 期，之後再用 Granger 因果關係檢定得知金價領先銅價、金價領先道瓊工業指數…，可得知金價可以做為銅價的參考指標。

2.3 Bert(transformers)模型+情感分析

Bert 全名為 Bidirectional Encoder Representations from Transformers，是在 2018 年由 Google AI Language(Devlin, Chang et al., 2018)所提出的預訓練模型(Pre-training model)。模型方面是使用 Transformer 的 Encoder 架構，Transformer 模型是一個基於注意力機制(Attention & Self attention)的 seq2seq 結構，是由 Google 於 2017 年所發表的論文「Attention is all you need」提出，其結構包含了 Encoder 和 Decoder。

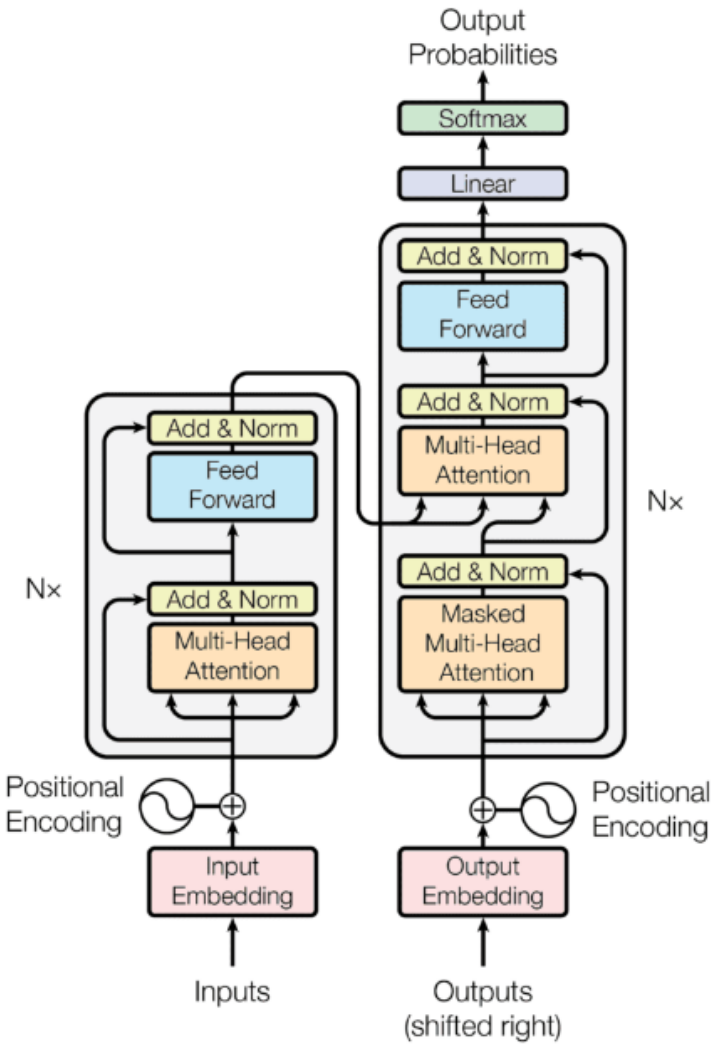


圖 3　Transformer 模型架構圖

資料來源:Vaswani, Shazeer et al. (2017).Attention is all you need.

Bert 就是 Transformer 裡的 Encoder(如圖 3 左半部)，細部的差別在於 Self-Attention 層(Multi-Head Attention)有幾層，訓練部分以 Google 釋出的初版 Bert-Base 中有 24 層、16 個自注意力頭(Self-Attention Head)、340M 的神經網路參數，訓練的語料使用 BooksCorpus 和英文的維基百科，並且是用無監督式學習(Unsupervised learning)，無監督式學習是指訓練的資料不需要標記，也就是說直接給文本但是不給答案讓 Bert 自己去找出裡面的關聯性，好處就是可以省非常多的成本且人工註記也曠日廢時。一般來說 Bert 分成兩階段的訓練，第一階段為預訓練(Pre-training)，通常使用極為大量的數據且訓練時數也長，通常會是由資源多的公司或是實驗室訓練且發佈，像是 Google 的 Bert-Base、台灣中研院 CKIP Lab 的 bert-base-chinese-ws 則是繁體中文的預訓練模型。第二階段是微調模型(Fine tuning)，微調則會根據下游任務的不同給 Bert 一些該任務或領域的標註資料，通常資料量不需要太多就會有不錯的效果，常見的自然語言處理(Natural Language Processing, NLP)任務有文本分類、問答、翻譯、摘要…。

第三章 研究方法

本章節介紹研究中的流程，包含資料蒐集、相關性分析、資料前處理(Data preprocessing)與 Bert 模型的訓練跟評估。本研究的資料集分別為股價與新聞，之後的處理包含股票之間的匹配、新聞與股價的匹配，最後使用自然語言技術 Bert 模型對資料集進行正負面新聞分類的訓練、評估至使用。

3.1 實驗流程

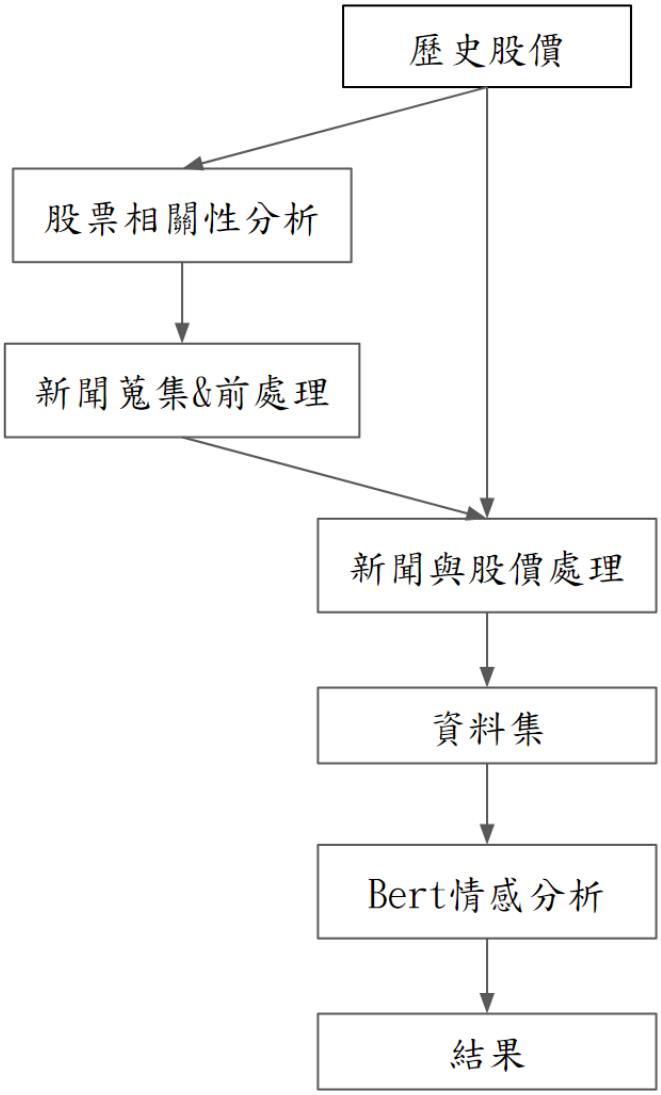


圖 4　研究架構

3.2 資料來源

3.2.1 歷史股價

使用 XQ 全球贏家程式看盤交易軟體匯出股票的還原股價。還原股價是用來計算股票投資報酬率的調整後股價，調整是為了要排除除權息的影響(Mr. Market 市場先生，2018)，讓股價不會因為除權息後而有程度不一的下跌，且若是未對除權息做處理，則對於當日的相關新聞情緒分析的結果會有很大的影響(模型會認為除權息是負面訊號)。

圖 5　還原日 K 圖

參考資料: XQ 全球贏家程式看盤交易軟體

圖 6　日 K 圖

參考資料: XQ 全球贏家程式看盤交易軟體

由上面兩張圖可見，股價於(08/12)有無進行還原的差距可能導致模型誤判，還原的方法就是把八月十三日之前的股價扣掉除息值。

3.2.2 股票新聞蒐集

使用 python 的爬蟲程式套件於鉅亨網(Anue)爬取 2020 年 3 月至 2022 年 3 月，約兩年的新聞。爬取的股票包含與本研究之目標股票的相關係數大於 0.9 的股票群。新聞資料的欄位資訊有新聞標題(Title)、內文(Content)、時間戳(Date)，時間包含日期、時、分，因收盤時間為下午一點半，故以此時間點為每日新聞的間隔。

表 1　新聞格式表

日期	標題	內文
2020/04/01(09:00)	個股分析/超眾 日本電產富爸爸加持	新冠肺炎疫情全球蔓延，全球各大企業紛紛採取異地分區工作，遠端及視訊會議帶動雲端伺服器及商務 NB 需求強勁，大陸新冠肺炎疫情受到控制後加速進行 5G 基礎建設，連原本市場看衰的大陸智慧型手機亦因 4G 手機銷售優於預期，庫存去化後近期重啟拉貨，…
2020/04/01(22:14)	研華攻遠距醫療 攜台中榮總、工研院打造智慧照護平台	工業電腦大廠研華(2395-TW) 今 (1) 日宣布，攜手臺中榮民總醫院、工研院產科國際所三方合作，打造智慧醫療照護平台，串聯患者從出院到返家後照護需求，強化遠距醫療布局。…
…	…	…
2022/03/31(21:06)	金管會畫四紅線嚴禁違法徵求委託書 違者表決權不予計算	為遏止委託書徵求亂象...

參考資料:鉅亨網新聞(https://www.cnyes.com)

3.3 資料處理

3.3.1 股票相關性分析

相關性分析將採用皮爾森相關係數，目的是得到兩檔股票的相關程度，1 為完全正相關、0 為不相關而-1 為完全負相關。本研究將使用台股市值前 300 大的公司股票進行相關性分析，研究主要挑選的股票採用元大中型 100(0051.TW)基金裡的 100 檔中型股，不選小型甚至是微型股，這類的小型股票通常波動性非常大或是人為操控性很大，例如某某投顧老師帶領會員進場某檔小型股票，通常會造成該股票劇烈震盪，但以消息面的新聞卻看不出任何端倪，這會更偏向技術面分析。

使用的資料為兩檔股票的還原股價做為樣本求得樣本平均數(公式 1)，再使用樣本每日收盤價與樣本平均數得到樣本標準差(公式 2)樣本共變異數(公式 3)，最後使用樣本共變異數除上樣本標準差得到相關係數(公式 4)，之後從相關係數矩陣中挑出與市值前 300 大股票中有多個相關的中型股進行後續分析。

表 2　股票相關矩陣

	股票 1	股票 2	…	股票 299	股票 300
股票 1	相關係數矩陣				
股票 2					
…					
股票 299					
股票 300					

樣本平均數:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{n} x_i \quad , \quad \overline{y} = \frac{1}{N}\sum_{i=1}^{n} y_i \tag{1}$$

樣本標準差 S:

$$S_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad , S_y = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{2}$$

樣本共變異數S_{xy}:

$$S_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{3}$$

相關係數 r 為:

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{4}$$

$$-1 \le r \le 1$$

x_i 、y_i，i={1，2…n}：x 與 y 的樣本序列

3.3.2 新聞資料前處理

新聞文字為非結構化數據，並且常常一篇新聞中存在一個段落討論多檔股票或是多個段落分別討論多種股票的情況。如下表 3.3，常常一篇新聞中含有多檔股票，這種情況下使用摘要系統(Text summarization)想要提取出重點的話都不會有太好的結果，例如 TextRank 或是 Transformer 架構的摘要生成預訓練模型，因為這些演算邏輯或預訓練模型不知道我們想提取的重點。

表 3 新聞範例

農金概念 Q1 營收大增 東鹼年增 125%創同期新高
俄烏衝突導致全球化肥供給大幅下滑，年初迄今肥料需求旺，氯化鉀、硫酸鉀、尿素等化肥因供不應求報價應聲飆漲，激勵農金概念股第一季營收大增，東鹼(1708-TW)第一季更創下單季歷史新高。東鹼 3 月營收 6.48 億元，較上月下滑 4.88%，年增 111.7%，累計第一季營收 19.11 億元，季增 20.9%、年增 125.7%。東鹼表示，第一季因遞延出貨 2 月出貨量拉高，預估第二季出貨狀況將維持正常，產品將依國際報價即時轉嫁給客戶。台肥 (1722-TW) 3 月營收 14.8 億元，月增 33.8%、年增 53.7%，第一季營收 40.8 億元，季減 18.8%、年增 38.5%，創 8 年高點。近期尿素飆漲，台肥尿素銷售以市價為主，成為第一季營收成長動能之一。惠光(6508-TW) 3 月營收 2.73 億元，月增 90.2%、年增 18.2%，累計第一季營收 6.32 億元，季增 23.6%、年增 20.4%，創 6 年新高。
鉅亨網記者張博翔 台北 2022/04/08 19:21

參考資料:鉅亨網新聞(https://www.cnyes.com)

為了從新聞中雜亂的資訊提取出特定股票的語句或段落，將對於所有爬取的新聞內容做斷句(Sentence segmentation)處理(如下)，可以看到一段新聞被分成六個獨立的句子，每個句子中的主題也明顯易讀，接下來就是要對個別句子抽取其股票特徵。

表 4　斷句處理

Sentence 1: “俄烏衝突導致全球化肥供給大幅下滑，年初迄今肥料需求旺，氯化鉀、硫酸鉀、尿素等化肥因供不應求報價應聲飆漲，激勵農金概念股第一季營收大增，東鹼(1708-TW) 第一季更創下單季歷史新高。”
Sentence 2: “東鹼 3 月營收 6.48 億元，較上月下滑 4.88%，年增 111.7%，累計第一季營收 19.11 億元，季增 20.9%、年增 125.7%。”
Sentence 3: “東鹼表示，第一季因遞延出貨 2 月出貨量拉高，預估第二季出貨狀況將維持正常，產品將依國際報價即時轉嫁給客戶。”
Sentence 4: “台肥 (1722-TW) 3 月營收 14.8 億元，月增 33.8%、年增 53.7%，第一季營收 40.8 億元，季減 18.8%、年增 38.5%，創 8 年高點。”
Sentence 5: “近期尿素飆漲，台肥尿素銷售以市價為主，成為第一季營收成長動能之一。”
Sentence 6: “惠光 (6508-TW) 3 月營收 2.73 億元，月增 90.2%、年增 18.2%，累計第一季營收 6.32 億元，季增 23.6%、年增 20.4%，創 6 年新高。”

抽取句子中的股票特徵將使用 CKIP Tagger 的斷詞技術(Tokenization)，CKIP Tagger 是中研院的詞庫小組所釋出的開源工具，且在繁體中文上的斷詞表現非常優異。

把一個句子分成多個詞(Token)之後，對分詞做股票比對，所以除了分詞之外還需要股票列表進行比對，本研究準備的股票列表為前 300 大市值的股票名，選擇新聞的好處之一是比較少出現股票的暱稱，如宏達電(2498-TW)不會使用「紅茶店」等暱稱，也較少使用「宏達國際電子」之類公司全名。

表 5　斷詞處理

,俄烏,衝突,導致,全球,化肥,供給,大幅,下滑,，,年初,迄,今,肥料,需求,旺,，,氯化鉀,、,硫酸鉀,、,尿素,等,化肥,因,供不應求,報價,應聲,飆漲,，,激勵,農金,概念股,第一,季,營收,大增,，,東鹼,(,1708,-TW,),第一,季,更,創下,單季,歷史,新高,。,

可以看到上表(表 5)，分詞後再比對股票列表得到該段落的股票特徵為「東鹼」，並且把該句分類成東鹼 2022/04/08 19:21 的股票新聞以利後續分析。

新聞的時間處理為當日下午一點半(收盤)之後的新聞都歸類成隔日新聞，周末或未開盤新聞則往後推到開盤第一日的新聞，如 2022/04/08 19:21，將被標記成 2022/04/09；而六日的新聞則被歸類成周一的新聞。

3.2.3 新聞資料結合股價漲跌的前處理

選定要分析的中型股與其相關股票群後，將新聞資料集提取相關股票的資料整合並且依照時間排列，漲跌的標籤則使用中型股的隔日股價漲跌(公式 5)對比實驗自定義的漲跌百分比比較後進行標註，若漲跌大於自定義的漲跌百分比則定義為漲，小於則為跌，相等則是平盤。

漲跌計算為:

(當日收盤還原股價 − 前一交易日收盤還原股價)/ 前一交易日收盤還原股價　(5)

表 6　新聞&股價處理範例

時間	2020.4.1	2020.4.2	2020.4.3	…	2022.3.28	2022.3.29
股價漲跌	漲	漲	跌	…	平	漲
時間	2020.3.31	2020.4.01	2020.4.02	…	2022.3.27	2022.3.28
中型股	新聞 a		新聞 b		新聞 c	新聞 d
相關股			新聞 e			新聞 f
相關股		新聞 g			新聞 h	

一天可能有多篇新聞的存在，但一天只會有一個漲跌的標註，這是使用股票漲跌對新聞進行標註一定會遇到的問題，其中的難點在於，當天的新聞一定全部都是正面或全部都是反面嗎(表 6)?為了探究這個問題，本研究採用詞頻率統計的方式，先將有標註漲跌的新聞資料集進行停用詞(Stop words)處理，再統計新聞中前 500 個常用辭彙於漲或跌情境中的出現頻率。

此處理把出現在兩個不同情境(漲、跌)的常用詞次數記錄下來，接著進行停用詞處理，之後用該詞於漲或跌出現的次數(公式 6)除以該詞的新聞總數(公式 7)得到機率化(公式 9、10)的數字，後續再使用機率化的數字進行比較。

停用詞(Stop words)表參考中文停用詞表(cn_stopwords)、哈工大停用詞表(hit_stopwords)，以及本研究自行從分詞中挑出的股票新聞相關停用詞，例如，月線、季線、均線、類股、盤勢…。

表 7　新聞詞彙數量統計表

常用辭彙	新聞數量(漲)	新聞數量(跌)	總和(sum)
詞彙一(x_1)	x_{p1}	x_{n1}	T_{x1}
詞彙二(x_2)	x_{p2}	x_{n2}	T_{x2}
詞彙三(x_3)	x_{p3}	x_{n3}	T_{x3}
…	…	…	…
x_{500}	x_{p500}	x_{n500}	T_{x500}
總和(sum)	T_p	T_n	T

個別情境的新聞數總和:

$$T_p = \sum_{i=1}^{500} x_{pi} \text{、} T_n = \sum_{i=1}^{500} x_{ni} \tag{6}$$

個別詞彙出現於所有新聞情境的總和:

$$T_{xi} = x_{pi} + x_{ni} \text{，} i = \{1\text{，}2\text{，} \dots \text{，}500\} \tag{7}$$

新聞數量總和:

$$T = \sum_{i=1}^{500} T_{xi} = T_p + T_n \tag{8}$$

機率化後標註漲的個別詞彙:

$$\frac{x_{pi}}{T_p} \text{，i} = \{1\text{，}2 \dots \text{，}500\} \tag{9}$$

機率化後標註跌的個別詞彙:

$$\frac{x_{ni}}{T_n} \text{，i} = \{1\text{，}2 \dots \text{，}500\} \tag{10}$$

表 8　機率化詞彙表

常用辭彙	新聞數量(漲)	新聞數量(跌)
詞彙一(x_1)	$\frac{x_{p1}}{T_p}$	$\frac{x_{n1}}{T_n}$
詞彙二(x_2)	$\frac{x_{p2}}{T_p}$	$\frac{x_{n2}}{T_n}$
詞彙三(x_3)	$\frac{x_{p3}}{T_p}$	$\frac{x_{n3}}{T_n}$
…	…	…
x_{500}	$\frac{x_{p500}}{T_p}$	$\frac{x_{n500}}{T_n}$
總和(sum)	1	1

對於個別情境的新聞數量的機率化，為的是讓兩邊的新聞占比一致。例如，下表的詞彙一(x_1)以直覺來判斷會認為該詞彙代表漲的意思，但進行機率化後得到漲為 1/100(1000/10 萬)，跌為 1/10(500/5 千)，代表該詞彙在漲的情境每百則才會出現一次，而跌的情境為十則新聞就出現一次。在股價下跌時更常出現詞彙一(x_1)，所以應該被歸類於使股價下跌的詞。

表 9　詞彙數量統計示意

常用辭彙	新聞數量(漲)	新聞數量(跌)
詞彙一(x_1)	1000 則	500 則
…	…	…
總和(sum)	10 萬則	5 千則

之後對詞彙進行加權分數處理，處理方式為:

$$\frac{x_{pi}}{T_p} - \frac{x_{ni}}{T_n} \text{，i} = \{1\text{，}2 \dots \text{，}500\} \quad (11)$$

分數為正代表該詞彙對整個句子有正向(漲)的作用，反之，本實驗會對一段分詞化(Tokenize)的文本比對常用辭彙進行加權處理(公式 11)，並且經過加權分數處理後設定一個閥值(詳見第四章)，超過閥值則將新聞的標籤從上漲改成下跌，或是從下跌改成上漲。

表 10　加權分數詞彙表

常用辭彙	新聞數量(漲)	新聞數量(跌)	加權分數
詞彙一(x_1)	$\frac{x_{p1}}{T_p}$	$\frac{x_{n1}}{T_n}$	$\frac{x_{p1}}{T_p} - \frac{x_{n1}}{T_n}$
詞彙二(x_2)	$\frac{x_{p2}}{T_p}$	$\frac{x_{n2}}{T_n}$	$\frac{x_{p2}}{T_p} - \frac{x_{n2}}{T_n}$
詞彙三(x_3)	$\frac{x_{p3}}{T_p}$	$\frac{x_{n3}}{T_n}$	$\frac{x_{p3}}{T_p} - \frac{x_{n3}}{T_n}$
…	…	…	
x_{500}	$\frac{x_{p500}}{T_p}$	$\frac{x_{n500}}{T_n}$	$\frac{x_{p500}}{T_p} - \frac{x_{n500}}{T_n}$
總和(sum)	1	1	0

3.4 Bert 新聞情感分析

將處理過的新聞股價資料集，資料集包含被研究之股票於 2020 年 3 月至 2022 年 3 月的新聞文本(Content)與股價(Label)，股價標註為上漲(Positive)、持平(Flat)或下跌(Negative)。

使用的 BERT 預模型為「bert-base-chinese」，fine-tuning 後的模為 Multi-Class Classification 形式(單一句子的分類任務，Sequence Classification)。原理為擷取 BERT 模型輸出的第一個 token「[CLS]」，將其再輸入一個線性分類層(Linear classifier)裡做線性轉換得到預測的標註，最後再以測試集(Test set)的新聞數據丟入模型中驗證股票漲跌的預測準確率。

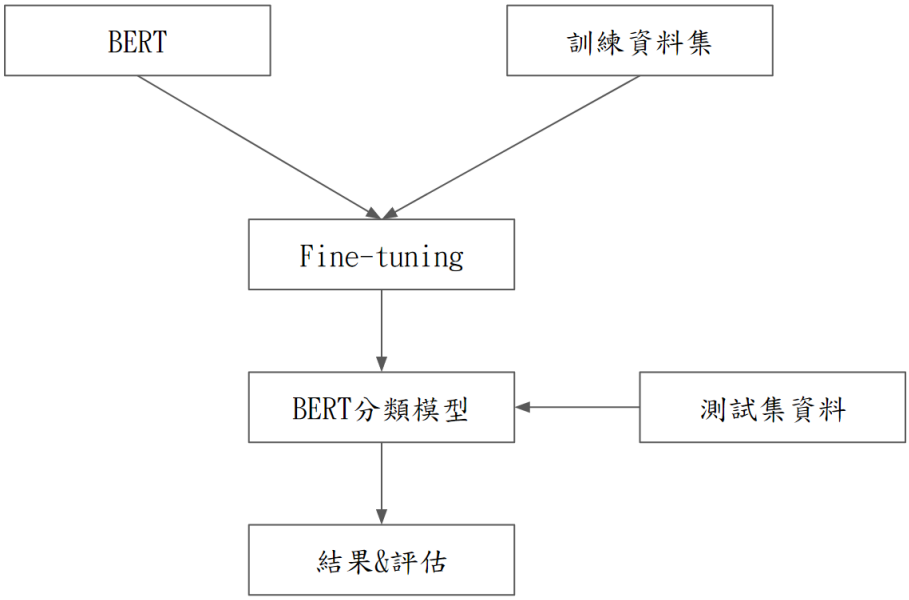


圖 7　模型訓練流程

第四章 研究結果

本次研究實驗環境均在 Google Colaboratory 上運行,所做的實驗為兩個項目，分別是使用 Bert 分類在相關係數股票群與台積電相關新聞做股票漲跌預測。

相關係數股票群為本次實驗研究之對象，而台積電則是參考多篇文獻後，認為可以拿來當對照組的資料，其原因在於，新聞資料多且於其他研究有被成功做出模型過。

以下實驗順序為台積電、相關股票群，實驗結果也包含失敗的文本與標籤前處理設定。

股票新聞經爬蟲後得到 68,828 筆未整理資料，該資料的前處理為對文本分詞後比對股票代號初步歸類。

4.1 台積電股票新聞漲跌預測

4.1.1 台積電的成功案例

在 6 萬多筆新聞資料中，經文本切割與股票比對後一共撈出 16,150 筆未標記漲跌(Label)的新聞內容，切割方式為句號(。)、分號(；)、And 符號(&)。

漲跌幅定義使用 0%，大於 0%標記漲，小於是跌，等於則標記平盤，共 3 個 Label。資料整體分佈如下:

表 11　資料集的 Label 分佈

漲	8249 筆
跌	7176 筆
平	725 筆

後續對原始台積電資料集做同日新聞單一漲跌的處理(本研究 3.2.3)，經分析漲與跌兩類的新聞共同詞彙並去除停用詞(Stop words)後，得到有句子加權分數並把台積電資料集分成漲、跌、平三個子資料集後對閥值進行處理(如下圖)。

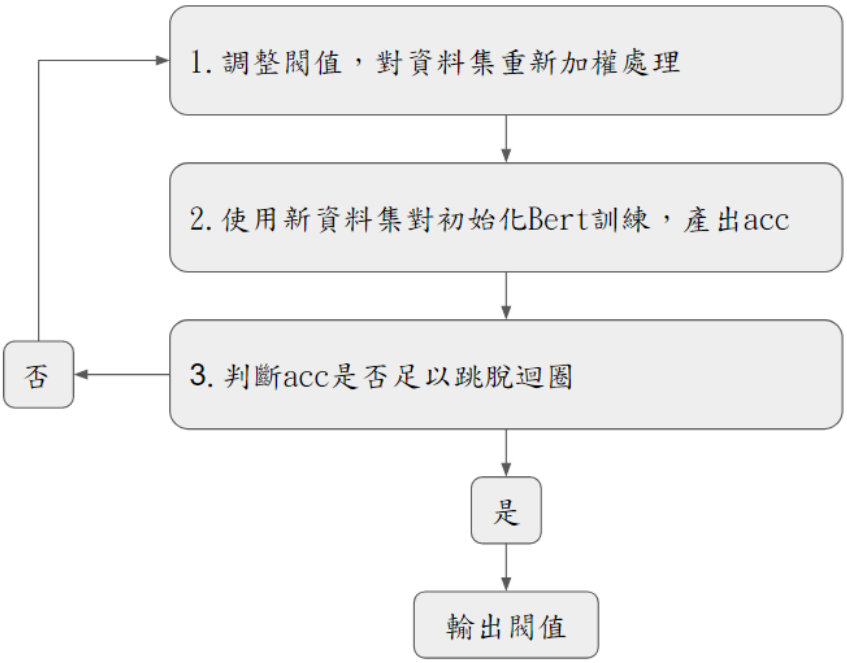


圖 8　閥值設定流程圖

對於漲資料及所進行的處理如下表，句子加權分數小於閥值−0.004則歸類成下跌，閥值是基於資料偏移量、天數偏移量與後續 Bert 模型訓練成效所設計成有自動修正機制的參數(偏移量越少而且模型預測力越好為佳)。

資料偏移量為:0.1092(901/8249，單位:筆)

天數偏移量為:0.0318(8/251，修正後變成跌的天數/原始漲的總天數)

表 12　漲資料集的資料偏移表

資料偏移	漲	跌
修正前的漲資料集	8249 筆	0 筆
修正後的漲資料集	7348 筆	901 筆
天數偏移	漲	跌
修正前的漲資料集	251 天	0 天
修正後的漲資料集	243 天	8 天

對於跌資料集及所進行的處理如下表，句子加權分數大於閥值0.005則修改標籤歸類為漲。

資料偏移量為:0.0814(584/7176，單位:筆)

天數偏移量為:0.0043(1/232，修正後變成跌的天數/原始漲的總天數)

表 13　跌資料集的資料偏移表

資料偏移	漲	跌
修正前的跌資料集	0 筆	7176 筆
修正後的跌資料集	584 筆	6604 筆
天數偏移	漲	跌
修正前的跌資料集	0 天	232 天
修正後的跌資料集	1 天	231 天

對於平盤資料集及所進行的處理為，句子加權分數小於閥值−0.004則歸類成跌，大於0.005則改成漲。

資料偏移量為:0.1462((31+75)/725，單位:筆)

天數偏移量為:0.4815(13/27，修正後變成跌的天數/原始漲的總天數)

表 14　平盤資料集的資料偏移表

資料偏移	漲	平	跌
修正前的平盤資料集	0 筆	725 筆	0 筆
修正後的平盤資料集	31 筆	607 筆	75 筆
天數偏移	漲	平	跌
修正前的平盤資料集	0 天	27 天	0 天
修正後的平盤資料集	4 天	14 天	9 天

資料處理完後，以 8:1:1 切割資料集為訓練集(Training dataset)、驗證集(Validation dataset)、測試集(Testing dataset)，並對訓練集使用亂數排序(Shuffle)，之後在 Pytorch 框架上對 Bert_base_chinese(語言層)與 BertClassifier(分類層)做 10 輪(Epochs)微調訓練(Fine-tuning)，並且對 Bert 各層的參數不進行凍結。

其中模型手動更改過的參數或方法有:

Loss function = Cross Entropy
Optimizer = Adam
Learning rate = 1e-5
Batch size = 16
Activation function = GELU

最後再以測試集資料驗證其準確度(Accuracy)，測試集是否亂數排列並不影響模型預測，而在本研究中不進行亂數的另一原因為比較好蒐集同一天不同新聞的預測結果再加以總和分析。

在此實驗中，經過多次訓練且每次得到最好的 Validation accuracy 都差不多出現在 epoch 第五輪的時候，而此時停止訓練且對測試集做預測的 Test accuracy 約為 0.68。多次訓練意指模型設定、模型初始權重一致但每次都為獨立訓練的 10 輪。

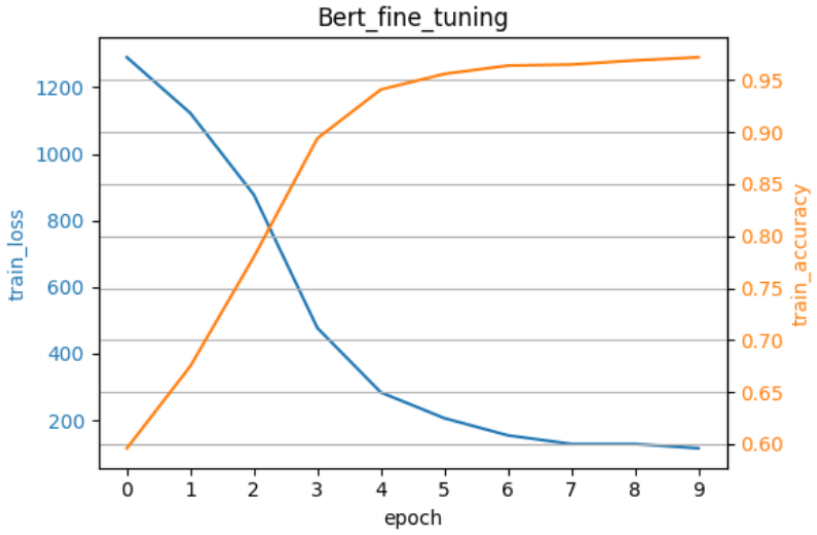


圖 9　Bert 訓練集趨勢圖

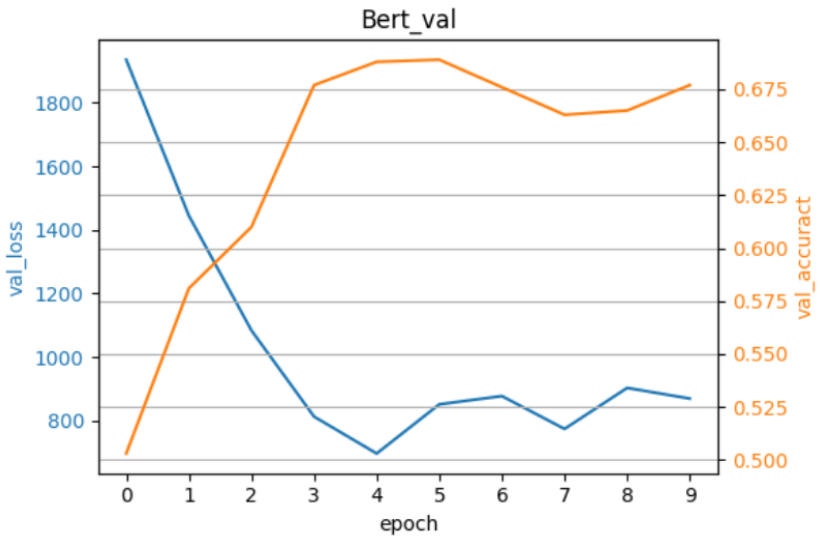


圖 10　Bert 驗證集趨勢圖

取其中一次在測試集(Testing dataset)驗證而預測成功率(Test Accuracy)為 0.6615 的結果分析，Test Accuracy 在本研究其他次的訓練中最高達 0.681。

表 15　台積電預測成功率

	真實資料數	預測成功數	成功率
漲	837	622	0.80787
跌	727	437	0.49299
平	49	8	0.16326
總和	1613	1067	0.66150

由上表可發現，微調之後的模型對於漲相關的新聞預測成功率達 8 成，對於跌的文本分類約 5 成左右，而平盤的話效果最差，平均不超過 2 成。

又以測試集(Testing dataset)丟入 Bert 分類模型後的成果帶入驗證測試集的真實每日股價漲跌，方法為當日所有新聞的預測漲跌標籤相加(大於 0 則預測漲，小於 0 則預測跌)時間為 2022/1/12 至 2022/3/31，若當日所有新聞預測相加結果與當日真實漲跌一致則代表預測成功，得到的最好結果為 0.6923。

表 16　台積電預測與真實漲跌對照

2022/1/12 的新聞	預測漲跌	真實漲跌
新聞 1	漲(1)	漲
新聞 2	漲(1)	
新聞 3	跌(-1)	
新聞 4	跌(-1)	
新聞 5	漲(1)	
Label sum	1+1+(-1)+(-1)+1=1　預測漲	成功預測

4.1.2 台積電的其他訓練實例紀錄

此部分為簡單記錄使用一樣的原始(未前處理)資料集，做不同的前處理之後訓練模型卻失敗、預測效果極差或作法複雜但跟上述成功實例差距微小的實例。以下方法的來源來自:本研究、他人論文、Medium 網站、CSDN 網站…。

模型失敗的定義:Train loss 無法有效下降、Train loss 下降但 Val loss 不變或上升、Val accuracy 跟全部都猜上漲或下跌的機率數值差不多。

本小節研究結論是:

i. 同樣的文本資料處理方式，可能因為資料爬取的時間區段、文本來源的不同或語言不同導致無法有效訓練模型。

ii. 不一樣的漲跌標籤(Label)設定，包含%數定義和 Label 數量，對於模型訓練也有非常重大的影響。

iii. 不一樣的 Bert 語言模型選擇，如使用中文預訓練模型跟多語言(Multilingual)預訓練模型，最終成效還是取決於資料品質。

實作案例:

對於文本(Content)資料的資料萃取方式:

1. 取鉅亨網內新聞標籤(tag)有標記台積電的「台股新聞」內文(Content)。
2. 取鉅亨網內新聞標籤(tag)有標記台積電的「台股新聞」標題(Title)。

對於文本(content)資料的文字前處理方式:

1. 取頭取尾法:Bert 的輸入最大長度(Max length)為 512 個單位，扣除 2 個特殊字元([CLS]、[SEP])，取前面 255 個字，後面 255 個字。
2. 最大長度截斷:取開頭至最大長度，不足則對其補值(Zero padding)。
3. 去除一切符號與空格:將所有標點符號、特殊符號與空格去除，留下純中文字，給模型最大的中文資訊量。
4. 關鍵詞或關鍵句提取法:使用 KeyBert 預訓練模型、PageRank 相關模型提出關鍵詞或關鍵句作為 Bert 分類模型的輸入。

對於停用詞(Stop words)的方法:

1. 僅使用現成的停用詞字典，如中文停用詞表、哈工大停用詞表、結巴停用詞。
2. 使用 TF-IDF，刪除在所有文本中分數低的詞。
3. 不使用停用詞。

對於標籤(Label)的方式:

1. 使用不等於 0%的漲跌定義:例如:使用 1.5%作為漲跌標準，亦即大於 1.5%為漲、小於 1.5%為跌，在此區間為平盤。該方法為 3 個 Label。
2. 使用二元分類:採用 2 個 Label 的分類作法，大於 0%為漲、小於 0%為跌，而等於 0%則剔除出後續訓練與驗證資料集。

對於加權處理的方法:

1. 使用 Log Likelihood 對詞頻機率表進行處理，公式如下

$$\lambda_i = \log\left(\frac{x_{pi}}{T_p} / \frac{x_{ni}}{T_n}\right)，\mathrm{i} = \{1，2 \ldots，500\} \quad (11)$$

對比本研究使用的漲跌機率相減的方法其分數值更能拉開差距，但套入 5.1 的實驗流程後，模型產出的成功率卻沒有明顯變化，約略都在 0.68 左右。

4.2 對中型股票與相關股票漲跌預測

有上述台積電成功訓練出可以收斂的模型之後，使用相同的模式套用在資料量比較少的中型股票與相關股票上，此部分與台積電成功案例的差別在於文本(Content)資料的不同，其餘方法包含資料前處理、漲跌定義、資料偏移修正皆沿用。

對元大中型 100 成分股與台股前 300 大股票的股票收盤價使用相關係數為 0.95 找出與元大中型 100 有相關股票，並從中得到 31 檔股票，其中每檔股票的相關股票群至少有 5 檔以上(詳細參考附件一)，下表 17 為範例。

表 17　中型 100 與其相關股票

元大中型 100 成分股	與其相關的台股前 300 大股票
華新(1605)	中鋼(2002)、元大金(2885)、台泥(1101)、大聯大(3702)、新纖(1409)、統一實(9907)、大統益(1232)、中再保(2851)
裕民(2606)	中鋼(2002)、長榮(2603)、陽明(2609)、萬海(2615)、大成鋼(2027)、慧洋-KY(2637)、中鴻(2014)、新纖(1409)、華紙(1905)
…	…

把目標股票與相關股票群的新聞按照時序先後排列，每則新聞的漲跌標籤使用目標股票的股價漲跌，一共產出 31 個資料集。將資料集以 8:1:1 切割成訓練、驗證與測試用，訓練、驗證資料集標籤經過加權處理，測試資料集中的標籤(Label)則使用未經加權修正的原漲跌標籤產出 Test Accuracy，最終再使用預測漲跌對照真實漲跌並結合日期得到「天數預測成功率」。

從下表中可以發現準確率突破 6 成的只有 4 檔股票，分別是景碩(3189)、潤泰新(9945)、聯強(2347)、中保科(9917)。從實驗結果可以得知可以使用於台積電的方法並不一定適用於本研究其他股票。

表 18　中型股實驗結果

編號	股票名稱	Test Accuracy	天數預測成功率
1	華新(1605)	0.543	0.5676
2	裕民(2606)	0.513	0.4545
3*	景碩(3189)	0.684	0.675
4*	中保科(9917)	0.582	0.631
5	永豐金(2890)	0.576	0.5936
6	裕融(9941)	0.515	0.5135
7	新光金(2888)	0.512	0.5294
8	南紡(1440)	0.544	0.449
9	潤泰全(2915)	0.502	0.4615
10	光寶科(2301)	0.496	0.4419
11	群創(3481)	0.514	0.48214
12*	潤泰新(9945)	0.608	0.6486
13	聯強(2347)	0.627	0.65517
14	仁寶(2324)	0.58	0.52632
15	長榮航(2618)	0.508	0.4643
16	大聯大(3702)	0.552	0.4884
17	華航(2610)	0.492	0.46154
18	大成鋼(2027)	0.481	0.4412
19	台光電(2383)	0.486	0.3333
20	聯華(1229)	0.501	0.5
21	台中銀(2812)	0.561	0.54286
22	國票金(2889)	0.57	0.56863
23	慧洋-KY(2637)	0.567	0.52941
24	永豐餘(1907)	0.544	0.4286
25	大成(1210)	0.504	0.41509
26	中鴻(2014)	0.503	0.55882
27	彩晶(6116)	0.542	0.45614
28	文曄(3036)	0.454	0.41667
29	燁輝(2023)	0.486	0.44444
30	新唐(4919)	0.561	0.5625
31	超豐(2441)	0.544	0.36111

以上表編號 13，聯強(2347)為例，模型訓練後產出資訊如下表:

表 19　聯強(2347)實驗結果

	成功預測資料分布	真實資料分布	成功率
漲	428 筆	541 筆	0.7911
跌	73 筆	258 筆	0.2829
平	0 筆	0 筆	0
SUM	501 筆	799 筆	0.627
預測正確天數		測試集總天數	天數預測成功率
19 天		29 天	0.655

又以景碩(3189)為例，驗證本研究的加權處理是否有效，如下表，可以觀察到使用不經任何處理的資料所得到的 Test Accuracy 與天數預測成功率均低於使用加權處理過的資料集。

表 20　景碩(3189)使用自身與相關股票並對資料進行閥值處理的實驗結果

	成功預測資料分布	真實資料分布	成功率
漲	234 筆	315 筆	0.7429
跌	145 筆	239 筆	0.6067
平	0 筆	0 筆	0
SUM	379 筆	554 筆	0.684
預測正確天數		測試集總天數	天數預測成功率
27 天		40 天	0.675

表 21　景碩(3189)使用自身與相關股票但不做新聞閥值處理的實驗結果

	成功預測資料分布	真實資料分布	成功率
漲	198 筆	315 筆	0.6286
跌	109 筆	239 筆	0.4561
平	0 筆	0 筆	0
SUM	307 筆	554 筆	0.554
預測正確天數		測試集總天數	天數預測成功率
22 天		40 天	0.55

若是只有使用中型股本身的資料集則會遇到資料數過少，測試成功率也不高的狀況，如下表，景碩(3189)有無使用資料增強的資料量差了快 8 倍，而越多有效的參考資料對於後續無論是訓練或是使用新聞做買賣決策都能提供越多資訊參考。

表 22　景碩(3189)使用自身新聞且不做新聞閥值處理的實驗結果

	成功預測資料分布	真實資料分布	成功率
漲	21 筆	45 筆	0.4667
跌	17 筆	31 筆	0.5484
平	0 筆	0 筆	0
SUM	38 筆	76 筆	0.5
預測正確天數		測試集總天數	天數預測成功率
11 天		29 天	0.3793

甚至對於預測率偏低的股票台中銀(2812)也存在提升效果從原本 Test Accuracy 為 0.492 提升至 0.5429，如下兩表:

表 23　台中銀(2812)使用自身與相關股票但不做新聞閥值歸類的實驗結果

	成功預測資料分布	真實資料分布	成功率
漲	244 筆	312 筆	0.7821
跌	24 筆	233 筆	0.103
平	0 筆	0 筆	0
SUM	268 筆	545 筆	0.492
預測正確天數		測試集總天數	天數預測成功率
18 天		35 天	0.5143

表 24　台中銀(2812) 使用自身與相關股票且用新聞閥值歸類的實驗結果

	成功預測資料分布	真實資料分布	成功率
漲	249 筆	312 筆	0.7981
跌	57 筆	233 筆	0.2446
平	0 筆	0 筆	0
SUM	306 筆	545 筆	0.561
預測正確天數		測試集總天數	天數預測成功率
19 天		35 天	0.5429

第五章 研究結論

1、解決新聞與股票多對一的問題與股票預測提升的貢獻

本研究中使用詞彙的加權分數修正同一天多則新聞但對應一個漲跌的問題，並在 31 檔被研究的中型股票測試成功率中平均取得了約 7%的進步，且能用更少量的訓練 Epochs 次數得到更好的模型成效；於台積電資料集的訓練中，加權分數處理的方法讓一開始藉由批量上標而無法訓練的困境也有所突破，並且預測準確率達到 0.67 左右。由此可見，原始資料集的漲跌標籤是存在一定程度的錯誤，而該錯誤輕則讓模型的產出預測低落，重則導致整個模型無法有效的學習到資料集的特徵。

值得注意的是如果資料集修正過多，如閥值的分數門檻設置過低，會讓訓練資料產生過度擬合(Over fitting)的效果；分數過高的話則又會和原始資料集無異。本研究所導入的機制是，模型多次獨立(初始權重一致)訓練下，測試成功率(Test Accuracy)達到最高情況下所使用的閥值參數，算是一種耗時但有效的暴力法。

在參考多篇文獻與網路他人對於中文或英文的股票新聞實作後，發現有些新聞股票漲跌預測的訓練不需經過過多處理就達到約 0.65 以上的結果，以一樣的手法在本研究上卻行不通，推測原因可能為選取時段的差異，本研究選擇的新聞區間為新冠肺炎(COVID-19)發生後兩年，影響原因可能包含，寫作風格改變、股市不確定性更高、…。

2、既然已經有詞彙分數表，為何還要 Bert

詞彙分數表在本研究中是使用來修正漲跌標籤(Label)的，之後還是以新聞原句加上修正後漲跌標籤丟入模型中進行分類訓練，並沒有對訓練文本(Content)進行任何刪減，原因在於 Bert 裡的 self-attention 機制能夠自己去找詞彙之間的關聯性，並且 Bert_base_chinese 對於中文的分詞機制是逐字分詞，並不像 CKIP 或結巴是使

用詞彙來分詞。這也讓本研究多了一個分支，即為刪除所有標點符號與模型無法識別的分詞(Unknown token)，讓輸入的文本有最大文字量，但產出結果並沒有多大差別，其原因推測在於超過 Bert 輸入最大字數為 512 個字，而大部分的資料都符合字數的輸入規定並沒有被篩選掉。

實驗結束後有嘗試使用詞彙分數表直接對台積電測試資料集進行分類，得到測試成功率(Test Accuracy)為 0.52，明顯低於使用 Bert 分類器。這也間接說明還存在特徵是詞彙分數表無法表示或該機制仍有進步空間。

3、關於停用詞與常用詞的選擇

本研究除了導入現成的停用詞字典並加以過篩外，還額外人工自定義的部份，關於停用詞的選擇也是金融研究中難點。以現今機器學習的主流下，是否使用停用詞已經變成一個選項，但是在機率統計的時代則是必要流程，原因在於機器學習可以透過機制修正模型權重或發掘更多資料特徵，如反向傳播與注意力機制。研究中所導入的停用詞則是在加權分數處理漲跌標籤階段，並沒有直接使用停用詞後的文本訓練模型。

其中常用詞中不能隨意剃除罕見詞彙，例如「巴菲特」，波克夏公司(Berkshire Hathaway)於 2022 年 11 月 14 日首度購入台積電 ADR 並影響台股台積電於隔日大漲，該詞彙近幾年來首次出現就造成台積電大漲；使用詞性標註(POS)以詞性批量處理的話也不完全恰當；停用詞使用 TF-IDF 則有些常用且有象徵性的詞彙分數會過低，例如「漲」跟「跌」，這兩個詞在文本大量出現但卻有其參考意義。若是不使用停用詞和常用詞則整體的計算量會變得非常龐大或特徵模糊導致模型訓練成效低落。

4、對中型股票與其有關的股票進行資料增強的做法

本研究使用股票的還原收盤價計算相關係數，並抽出高相關係數的股票群集新聞當成資料增強的來源，若使用單一中型股的新聞資料對本身做資料增強，可以增加同一天的資訊數量但無法增加新聞的總天數，簡單來說，可以增加深度但廣度依然不足，並且這個資料深度還存在批量上標的潛在錯誤風險。

5、加權分數存在被稀釋問題

以上述第 3 點中所提到的「巴菲特」為例，該詞在詞頻率中出現率極低，但影響卻極為重大，導致計算加權分數後該詞的貢獻非常的低，相當於分數被詞頻高的詞彙稀釋了。若是使用 TF-IDF 則可以提升其貢獻程度，但對於其他詞頻高且意義重大的詞一樣有稀釋問題，這也是實驗中可以改良的部分，如何以演算法的方式批量找出有意義但低詞頻與剃除無意義卻高詞頻的詞彙。

參考文獻

Fama, E. F. (1970). Efficient Capital Markets : A Review of Theory and Empirical Work. The journal of Finance, 25, 383-417

Lhabitant, Francois-Serge(2011). Correlation vs. Trends in Portfolio Management: A Common Misinterpretation (April 12, 2011).Available at SSRN: https://ssrn.com/abstract=1808267 or http://dx.doi.org/10.2139/ssrn.1808267

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin(2017). Attention Is All You Need. arXiv:1706.03762

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova(2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805

陳彥豪(2002)。外資與投信法人持股比率變化對股價報酬率影響之研究-以上市電子股為例〔未出版之碩士論文〕。國立中山大學財務管理系研究所

王釗東(2018)。以大數據探究財經新聞對台灣股票市場表現之影響 An Analysis of Financial News Influence on Stock Market in Taiwan〔未出版之碩士論文〕。國立台灣大學社會科學院新聞研究所

古金尚(2003)。台灣股票市場投資者心理情緒影響因素之實證研究〔未出版之碩士論文〕。朝陽科技大學財務金融系

王彥鈞(2017)。不同市場狀態下新聞情緒的預測能力：以台灣五十指數為例〔未出版之碩士論文〕。國立中央大學財務金融學系

吳青山(2018)。外溢效應。教育研究月刊，294 期，135-136

楊智欽(2021)。金價、銅價對道瓊工業平均數與美國工業生產指數關係之研究〔未出版之碩士論文〕。國立成功大學財務金融系

夏鶴芸(2020)。應用深度學習與自然語言處理新技術預測股票走勢 – 以台積電為例〔未出版之碩士論文〕。國立臺北大學資訊管理系

Mr. Market 市場先生(2018/09/03)。什麼是「還原股價」？如何計算和查詢？。https://rich01.com/price-adjusted/

附錄

附件一 股票相關係數表

華新(1605)	中鋼(2002) 相關係數: 0.9544859125376993 元大金(2885) 相關係數: 0.9686854405863773 台泥(1101) 相關係數: 0.9561791847378844 大聯大(3702) 相關係數: 0.9527875007064405 新纖(1409) 相關係數: 0.9601424173376453 統一實(9907) 相關係數: 0.958679408578455 大統益(1232) 相關係數: 0.9652140848101596 中再保(2851) 相關係數: 0.9548532462454639
裕民(2606	中鋼(2002) 相關係數: 0.9591962960629252 長榮(2603) 相關係數: 0.9584098583050986 陽明(2609) 相關係數: 0.9654090434499903 萬海(2615) 相關係數: 0.9536110265990627 大成鋼(2027) 相關係數: 0.9584787723451437 慧洋-KY(2637) 相關係數: 0.9714729865035581 中鴻(2014) 相關係數: 0.9729332875463719 新纖(1409) 相關係數: 0.956309103619794 華紙(1905) 相關係數: 0.9715227761291696
景碩(3189)	富邦金(2881) 相關係數: 0.9504481996518097 中租-KY(5871) 相關係數: 0.9562743342406614 台新金(2887) 相關係數: 0.9589475965157671 南電(8046) 相關係數: 0.9535612630400269 中保科(9917) 相關係數: 0.9630957607412973 潤泰新(9945) 相關係數: 0.9535722304888109 新唐(4919) 相關係數: 0.9564105158356744 強茂(2481) 相關係數: 0.9694129816896732 順德(6412) 相關係數: 0.9671248016406787 華票(2820) 相關係數: 0.9518124579051498

中保科(9917)	國泰金(2882) 相關係數: 0.9628111220764365 中租-KY(5871) 相關係數: 0.9624276175736766 欣興(3037) 相關係數: 0.9552309031329286 開發金(2883) 相關係數: 0.9595490513544384 台新金(2887) 相關係數: 0.9757999860348402 南電(8046) 相關係數: 0.9521516646978737 景碩(3189) 相關係數: 0.9630957607412973 裕融(9941) 相關係數: 0.9741923294713162 潤泰全(2915) 相關係數: 0.9662027905463922 潤泰新(9945) 相關係數: 0.9816326773095064 聯華(1229) 相關係數: 0.9695562285165886 華票(2820) 相關係數: 0.9774865188714631 聯華食(1231) 相關係數: 0.9653537360895248 新產(2850) 相關係數: 0.985972189610123
永豐金(2890)	國泰金(2882) 相關係數: 0.9611616838247597 中信金(2891) 相關係數: 0.9876357620497689 兆豐金(2886) 相關係數: 0.9742710618961369 欣興(3037) 相關係數: 0.952122412497106 開發金(2883) 相關係數: 0.9916616437888479 第一金(2892) 相關係數: 0.9555715695127405 合庫金(5880) 相關係數: 0.9630838230507607 台新金(2887) 相關係數: 0.9585603372142375 裕融(9941) 相關係數: 0.9713581214299898 新光金(2888) 相關係數: 0.9585953674868262 潤泰新(9945) 相關係數: 0.9559800651454228 聯強(2347) 相關係數: 0.9548115601674267 長榮航(2618) 相關係數: 0.9665083668951364 華航(2610) 相關係數: 0.9524732836547619 聯華(1229) 相關係數: 0.9626807639374304 台中銀(2812) 相關係數: 0.9504984933403295 文曄(3036) 相關係數: 0.979727757897866 聯邦銀(2838) 相關係數: 0.971412506211948

	華票(2820) 相關係數: 0.9542821932886796 聯華食(1231) 相關係數: 0.9511778667515511 新產(2850) 相關係數: 0.9590429220004347
裕融(9941)	國泰金(2882) 相關係數: 0.9612824930164775 中信金(2891) 相關係數: 0.9510551822323731 兆豐金(2886) 相關係數: 0.9517857489334395 開發金(2883) 相關係數: 0.9736571585782928 台新金(2887) 相關係數: 0.9746901549428252 中保科(9917) 相關係數: 0.9741923294713162 永豐金(2890) 相關係數: 0.9713581214299898 潤泰全(2915) 相關係數: 0.9772116639973054 潤泰新(9945) 相關係數: 0.9829542213624077 聯華(1229) 相關係數: 0.9706634921737901 文曄(3036) 相關係數: 0.9654805727115088 新唐(4919) 相關係數: 0.9594514550462051 勝一(1773) 相關係數: 0.9597535532483065 華票(2820) 相關係數: 0.9634893714633308 聯華食(1231) 相關係數: 0.9795469070244046 新產(2850) 相關係數: 0.9738486844919129
新光金(2888)	國泰金(2882) 相關係數: 0.9520579994540149 中信金(2891) 相關係數: 0.9651225657068621 開發金(2883) 相關係數: 0.9666575621169958 永豐金(2890) 相關係數: 0.9585953674868263 華航(2610) 相關係數: 0.9531521635140778 華票(2820) 相關係數: 0.9555010193552184
南紡(1440)	中鋼(2002) 相關係數: 0.9599231646810329 元大金(2885) 相關係數: 0.9571835559228005 國票金(2889) 相關係數: 0.957844436219651 永豐餘(1907) 相關係數: 0.9517767577515258 中鴻(2014) 相關係數: 0.958376104486781 彩晶(6116) 相關係數: 0.9515546351503303 中石化(1314) 相關係數: 0.9613148638186966

	群益證(6005) 相關係數: 0.9534205778528161 新纖(1409) 相關係數: 0.9780629278571068 台聚(1304) 相關係數: 0.9539356182982851 統一證(2855) 相關係數: 0.9593531199955089 大統益(1232) 相關係數: 0.9536942515974345 聯成(1313) 相關係數: 0.9532374307372552 華紙(1905) 相關係數: 0.9707272108485879
潤泰全(2915)	開發金(2883) 相關係數: 0.9540185277008497 台新金(2887) 相關係數: 0.9744747445123585 中保科(9917) 相關係數: 0.9662027905463922 裕融(9941) 相關係數: 0.9772116639973055 潤泰新(9945) 相關係數: 0.9874827928806944 聯華(1229) 相關係數: 0.9597652231770104 新產(2850) 相關係數: 0.9662118602656846
光寶科(2301)	南亞(1303) 相關係數: 0.965735031874307 元大金(2885) 相關係數: 0.9528820109589137 華碩(2357) 相關係數: 0.9586658146981929 仁寶(2324) 相關係數: 0.9694626058487339 大聯大(3702) 相關係數: 0.9671285510110661 國票金(2889) 相關係數: 0.9555385044390746 豐興(2015) 相關係數: 0.9593729150632214 震旦行(2373) 相關係數: 0.9528636529989339 至上(8112) 相關係數: 0.9605006306736634
群創(3481)	聯詠(3034) 相關係數: 0.9592655665549698 友達(2409) 相關係數: 0.9876794480336354 彩晶(6116) 相關係數: 0.9638776808196372 天鈺(4961) 相關係數: 0.963735897183526
潤泰新(9945)	國泰金(2882) 相關係數: 0.9546298357488208 欣興(3037) 相關係數: 0.9635909129713248 開發金(2883) 相關係數: 0.9656787114172041 台新金(2887) 相關係數: 0.9811008498787515 景碩(3189) 相關係數: 0.9535722304888109

	中保科(9917) 相關係數: 0.9816326773095065 永豐金(2890) 相關係數: 0.9559800651454228 裕融(9941) 相關係數: 0.9829542213624076 潤泰全(2915) 相關係數: 0.9874827928806944 聯華(1229) 相關係數: 0.9816732107495669 台肥(1722) 相關係數: 0.9524841481310234 文曄(3036) 相關係數: 0.9525950685800845 華票(2820) 相關係數: 0.9663384198169983 聯華食(1231) 相關係數: 0.9581677750452621 新產(2850) 相關係數: 0.9811967852390693
聯強(2347)	中信金(2891) 相關係數: 0.9603285206728432 中華電(2412) 相關係數: 0.9586547307627991 兆豐金(2886) 相關係數: 0.9570863064912337 開發金(2883) 相關係數: 0.9511607049267911 永豐金(2890) 相關係數: 0.9548115601674269 長榮航(2618) 相關係數: 0.9522703268223138 華航(2610) 相關係數: 0.950870528812769 聯華(1229) 相關係數: 0.9511163485859458 聯邦銀(2838) 相關係數: 0.9603362628126394 崇越(5434) 相關係數: 0.9606468161108844 華立(3010) 相關係數: 0.954293273201114
仁寶(2324)	華碩(2357) 相關係數: 0.966289574021472 光寶科(2301) 相關係數: 0.9694626058487338 國票金(2889) 相關係數: 0.957037509359101 豐興(2015) 相關係數: 0.9553642151308938
長榮航(2618)	中信金(2891) 相關係數: 0.9559030321987256 兆豐金(2886) 相關係數: 0.9503347694196462 欣興(3037) 相關係數: 0.9561973533756232 開發金(2883) 相關係數: 0.9598832903351544 永豐金(2890) 相關係數: 0.9665083668951366 聯強(2347) 相關係數: 0.9522703268223139 華航(2610) 相關係數: 0.9820264549660008

大聯大(3702)	國泰金(2882) 相關係數: 0.9572088193646212 南亞(1303) 相關係數: 0.9523501243838509 元大金(2885) 相關係數: 0.9629137367193928 華碩(2357) 相關係數: 0.952151964985214 南電(8046) 相關係數: 0.950259416126821 華新(1605) 相關係數: 0.9527875007064405 光寶科(2301) 相關係數: 0.9671285510110661 豐興(2015) 相關係數: 0.9630553588313571 光罩(2338) 相關係數: 0.9511166238823543 華票(2820) 相關係數: 0.9532079167780299 震旦行(2373) 相關係數: 0.9506170609883696 至上(8112) 相關係數: 0.9619127215241364 聯華食(1231) 相關係數: 0.9559302164416426 中再保(2851) 相關係數: 0.9583799255691421
華航(2610)	開發金(2883) 相關係數: 0.954316786164388 永豐金(2890) 相關係數: 0.9524732836547619 新光金(2888) 相關係數: 0.9531521635140778 聯強(2347) 相關係數: 0.9508705288127689 長榮航(2618) 相關係數: 0.9820264549660009
大成鋼(2027)	中鋼(2002) 相關係數: 0.9651548363606441 長榮(2603) 相關係數: 0.96074923450154 陽明(2609) 相關係數: 0.9606641302140051 裕民(2606) 相關係數: 0.9584787723451436 慧洋-KY(2637) 相關係數: 0.9624978407612227 中鴻(2014) 相關係數: 0.9661456410980844 燁輝(2023) 相關係數: 0.9540701284158021 晶豪科(3006) 相關係數: 0.9591100022411018 新纖(1409) 相關係數: 0.9547190901660508 華紙(1905) 相關係數: 0.9606852929631581
台光電(2383)	欣興(3037) 相關係數: 0.9625275391189588 開發金(2883) 相關係數: 0.9514034535834389 台新金(2887) 相關係數: 0.9518949087695459

	華票(2820) 相關係數: 0.9550569649662404
聯華(1229)	國泰金(2882) 相關係數: 0.9577657989423313 中信金(2891) 相關係數: 0.952400547000593 欣興(3037) 相關係數: 0.9656095607484785 開發金(2883) 相關係數: 0.9698733756161053 合庫金(5880) 相關係數: 0.9573452566321472 台新金(2887) 相關係數: 0.9642477548784874 中保科(9917) 相關係數: 0.9695562285165886 永豐金(2890) 相關係數: 0.9626807639374302 裕融(9941) 相關係數: 0.9706634921737901 潤泰全(2915) 相關係數: 0.9597652231770105 潤泰新(9945) 相關係數: 0.9816732107495668 聯強(2347) 相關係數: 0.9511163485859458 台肥(1722) 相關係數: 0.9569196050475993 文曄(3036) 相關係數: 0.9621610855502034 崇越(5434) 相關係數: 0.9727628467372371 華票(2820) 相關係數: 0.9629312103649741 聯華食(1231) 相關係數: 0.9555845417528717 新產(2850) 相關係數: 0.9679799784971989
台中銀(2812)	中信金(2891) 相關係數: 0.9622256901213131 中華電(2412) 相關係數: 0.9555509620716862 兆豐金(2886) 相關係數: 0.9755113633289821 玉山金(2884) 相關係數: 0.9666849494063506 第一金(2892) 相關係數: 0.9846841803113153 合庫金(5880) 相關係數: 0.980531253780484 華南金(2880) 相關係數: 0.9694321431150125 永豐金(2890) 相關係數: 0.9504984933403295 文曄(3036) 相關係數: 0.9513345107209948 和潤企業(6592) 相關係數: 0.9611616242552613 聯邦銀(2838) 相關係數: 0.9782455844472743 勝一(1773) 相關係數: 0.957478633255859 王道銀行(2897) 相關係數: 0.972946860877779

國票金(2889)	南亞(1303) 相關係數: 0.9683254477169669 元大金(2885) 相關係數: 0.9810689209152605 台泥(1101) 相關係數: 0.9550719734224948 華碩(2357) 相關係數: 0.9800459353720404 南紡(1440) 相關係數: 0.957844436219651 光寶科(2301) 相關係數: 0.9555385044390746 仁寶(2324) 相關係數: 0.9570375093591011 大成(1210) 相關係數: 0.9709070055359158 中石化(1314) 相關係數: 0.952418295188722 晶豪科(3006) 相關係數: 0.9560818769441951 中華(2204) 相關係數: 0.967285665215932 群益證(6005) 相關係數: 0.9812992929431764 新纖(1409) 相關係數: 0.9717587286822162 台聚(1304) 相關係數: 0.9513831090077136 統一證(2855) 相關係數: 0.9742860044041806 大統益(1232) 相關係數: 0.9650773173967439 光罩(2338) 相關係數: 0.9574708969659405 凌陽(2401) 相關係數: 0.9580941936721161 宏全(9939) 相關係數: 0.9541320578738844 震旦行(2373) 相關係數: 0.951374167809996 華夏(1305) 相關係數: 0.967455752837111 三星(5007) 相關係數: 0.9586846656547963
慧洋-KY(2637)	富邦金(2881) 相關係數: 0.9508386167371249 長榮(2603) 相關係數: 0.9819159757362832 陽明(2609) 相關係數: 0.9773227389767264 萬海(2615) 相關係數: 0.9681761526031695 裕民(2606) 相關係數: 0.9714729865035581 大成鋼(2027) 相關係數: 0.9624978407612227 中鴻(2014) 相關係數: 0.9622407458018198 榮運(2607) 相關係數: 0.95947533116565 華紙(1905) 相關係數: 0.9545306854818444 潤弘(2597) 相關係數: 0.9511605198067232

永豐餘(1907)	南紡(1440) 相關係數: 0.9517767577515258 大成(1210) 相關係數: 0.9527550474137617 彩晶(6116) 相關係數: 0.9632102900434851 中石化(1314) 相關係數: 0.9627955799809607 群益證(6005) 相關係數: 0.9515768139954063 晶碩(6491) 相關係數: 0.9508581674453723 台聚(1304) 相關係數: 0.9547971428197031 統一證(2855) 相關係數: 0.9599245720219218 聯成(1313) 相關係數: 0.9709831620790007 國喬(1312) 相關係數: 0.9618328220776788
大成(1210)	南亞(1303) 相關係數: 0.9517026356714116 元大金(2885) 相關係數: 0.973724360850145 台泥(1101) 相關係數: 0.960239850523746 國票金(2889) 相關係數: 0.9709070055359157 永豐餘(1907) 相關係數: 0.9527550474137617 中石化(1314) 相關係數: 0.9538702006992643 中華(2204) 相關係數: 0.9703431907733546 群益證(6005) 相關係數: 0.968207872144897 新纖(1409) 相關係數: 0.9527340380735145 晶碩(6491) 相關係數: 0.9548135053446014 台聚(1304) 相關係數: 0.9554715891815938 統一證(2855) 相關係數: 0.9683291898398889 大統益(1232) 相關係數: 0.9641688488056341 聯成(1313) 相關係數: 0.9512353816400453 國喬(1312) 相關係數: 0.9521637786632885 卜蜂(1215) 相關係數: 0.951107768942644 宏全(9939) 相關係數: 0.9591614908826297 華夏(1305) 相關係數: 0.9570646757165479
中鴻(2014)	中鋼(2002) 相關係數: 0.977396766941827 長榮(2603) 相關係數: 0.9548662536214352 陽明(2609) 相關係數: 0.9691732131551071 裕民(2606) 相關係數: 0.9729332875463717

	南紡(1440) 相關係數: 0.958376104486781 大成鋼(2027) 相關係數: 0.9661456410980843 慧洋-KY(2637) 相關係數: 0.9622407458018198 燁輝(2023) 相關係數: 0.9529261030471964 新纖(1409) 相關係數: 0.9672175214669168 華紙(1905) 相關係數: 0.9807332658579663
彩晶(6116)	友達(2409) 相關係數: 0.961977943964396 南紡(1440) 相關係數: 0.9515546351503302 群創(3481) 相關係數: 0.9638776808196371 永豐餘(1907) 相關係數: 0.9632102900434852 中石化(1314) 相關係數: 0.960452917363937 統一證(2855) 相關係數: 0.9528648818316705 聯成(1313) 相關係數: 0.9587996431476161
文曄(3036)	國泰金(2882) 相關係數: 0.953296707672316 中信金(2891) 相關係數: 0.9700334639383417 兆豐金(2886) 相關係數: 0.9590031952029798 開發金(2883) 相關係數: 0.9838694876293925 合庫金(5880) 相關係數: 0.9594805758290744 台新金(2887) 相關係數: 0.9565293653410404 永豐金(2890) 相關係數: 0.9797277578978659 裕融(9941) 相關係數: 0.9654805727115088 潤泰新(9945) 相關係數: 0.9525950685800846 聯華(1229) 相關係數: 0.9621610855502035 台中銀(2812) 相關係數: 0.9513345107209947 聯邦銀(2838) 相關係數: 0.9670562025815204
燁輝(2023)	陽明(2609) 相關係數: 0.9627441726893972 萬海(2615) 相關係數: 0.9631272916684886 台玻(1802) 相關係數: 0.9541770856999371 大成鋼(2027) 相關係數: 0.9540701284158021 中鴻(2014) 相關係數: 0.9529261030471965 台驊投控(2636) 相關係數: 0.9556383357375202
新唐(4919)	台新金(2887) 相關係數: 0.9528412336447177

	景碩(3189) 相關係數: 0.9564105158356744 裕融(9941) 相關係數: 0.9594514550462051 強茂(2481) 相關係數: 0.9557426931252399 聯華食(1231) 相關係數: 0.9689586281204411
超豐(2441)	日月光投控(3711) 相關係數: 0.9502863376304082 南茂(8150) 相關係數: 0.9523527205280558 矽格(6257) 相關係數: 0.9749216578059482 盛群(6202) 相關係數: 0.9566315830442903 凌陽(2401) 相關係數: 0.9738745280312848 億光(2393) 相關係數: 0.9592276826159088 致新(8081) 相關係數: 0.9572651195454609