國 立 臺 北 大 學 資 訊 管 理 研 究 所

碩士論文

指導教授：方鄒昭聰　博士

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

Stock Price Trend Prediction Using Deep Learning and Natural Language Processing New Technology

- A Case Study of TSMC

研究生：夏鶴芸　撰

中　華　民　國　一　〇　九　年　八　月

國 立 臺 北 大 學

資訊管理研究所碩士班

108 學年度第 2 學期畢業生論文

研究生：夏鶴芸 撰

業經本委員會審議通過

題　　目：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

論文考試委員：

召 集 人 ＿＿＿＿＿＿

委　　員 ＿＿＿＿＿＿

委　　員 ＿＿＿＿＿＿

委　　員 ＿＿＿＿＿＿

指導教授 ＿＿＿＿＿＿

所　　長 ＿＿＿＿＿＿

論文口試及格日期：

中　華　民　國　一　〇　九　年　七　月　二　十　日

國立臺北大學 108 學年度第 2 學期碩士學位論文提要

論文題目：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
論文頁數：44 頁

所　組　別：資 訊 管 理 所　　（學號：710736110）
研　究　生：夏 鶴 芸　　指導教授：方 鄒 昭 聰

論文提要內容：

在過去預測股價的研究中，大多使用歷史交易資料及其他技術指標作為輸入特徵，然而股價走勢也會受到外部因素影響，包含相關指數的波動以及新聞或社群媒體提供的資訊。在使用的預測模型方面，過去大多利用 LSTM 或 RNN 對歷史資料進行分析，使用自然語言處理對新聞文本進行情緒分類。隨著 Google 提出 Transformer 模型和 BERT 模型後，國外已經開始研究使用這兩個模型預測股價走勢，但國內卻鮮少人利用。

本研究將探討新技術對預測股價走勢的影響，並以台積電為例，將台積電的股價資料與相關指數作為輸入資料，使用 BiLSTM 和 Transformer 預測股價走勢，並蒐集台積電相關新聞，利用股價漲跌與情緒詞典對新聞進行情緒標籤後，使用 BERT 進行新聞情緒分類。研究結果顯示，資料集加上相關指數資料對於預測股價走勢是有相當幫助，且成效最佳的是 Transformer 模型搭配股價與相關指數資料集。在新聞情緒分類方面，則是 BERT 模型搭配利用股價漲跌標籤的新聞資料集，得到更好的成效。

關鍵字：BERT、Transformer、台積電、股價走勢預測

ABSTRACT

Stock Price Trend Prediction Using Deep Learning and Natural Language Processing
New Technology - A Case Study of TSMC
by
Hsia, Ho-Yun
August 2020

ADVISOR(S): Dr. Fang-Tsou, Chao-Tsou
DEPARTMENT: Graduate Institute of Information Management
MAJOR: Information Management
DEGREE: Master of Business Administration

In the past study on stock price prediction, most of them used historical transaction data and other technical indicators as input features. However, stock price trends are also affected by external factors, including fluctuations in related indexes and information provided by news or social media. In the past, most of them used LSTM or RNN as predictive models to analyze historical data, and natural language processing was used to classify news texts. After Google proposed the Transformer model and the BERT model, others countries have begun to study the use of these two models to predict stock price trends, but few people use them in Taiwan.

This study will explore the impact of new technologies on predicting stock price trends, and will take TSMC as an example. Using TSMC's stock price data and related indexes as input data, BiLSTM and Transformer are used to predict stock price trends. After collecting TSMC-related news, using the stock price fluctuation and sentiment dictionary to label the news, and use BERT to classify the news sentiment. The result of the study show that using related index data is very helpful for predicting the stock price trend, and the best performance is the Transformer model with the stock price and related index data set. The study results show that the data set plus relevant index data is quite helpful for predicting the stock price trend. The best results are the Transformer model with stock prices and related index data sets. In terms of news sentiment classification, the BERT model is used with news data sets that use stock price fluctuation labels to achieve better results.

Keywords: BERT、Transformer Model、LSTM、BiLSTM、TSMC、Stock Price Trend Prediction

目次

圖次

表次

第壹章 緒論

第一節 研究背景與動機

在股票市場中，投資者大多以散戶為主，且隨著網路科技及社群網路的大眾化，投資者大多從網路新聞、財經網站以及社群網路文章中取得歷史與未來股市資訊，而這些資訊將會影響投資者的買賣策略，進而影響股價波動。但是對於大公司而言，影響股價波動不單只是過去歷史資料，還包括競爭對手、客戶、全球經濟、政治形勢、財政和貨幣政策等等許多外部因素，像是 2008 年的全球金融風暴、2018 年的中美貿易戰，以及今年 2020 年的新型冠狀病毒疫情，都相互作用連帶影響了股市波動。

長久以來預測股價一直是學術界和業界都十分重視的議題，過去已經有許多研究使用歷史交易數據及其他技術指標如移動平均線（Moving Average）、布林通道（Bollinger Band）或 KD 隨機指標（Stochastic Oscillator）等等資料，並利用迴歸分析（Regression Analysis）、整合移動平均自迴歸模型（Autoregressive Integrated Moving Average model, ARIMA Model）或支持向量機（Support Vector Machine, SVM）等等分析方法，來預測未來的股價趨勢。

然而，股價走勢也會受到許多外部因素影響，無法精準預測股價漲跌程度，這些外在因素大多都是經由股市或財經方面的新聞、網站或社群文章來瞭解公司的相關資訊。過去已經有研究利用文字探勘及語意分析，來判別這些新聞或消息的內容對公司帶來是正面或負面影響，進而探討新聞對股價漲跌的趨勢。

隨著深度學習和自然語言處理的技術突破，近年來國內的研究發現，大多利用 LSTM 或 RNN 分析歷史交易資料，以及使用自然語言處理對網路上財經新聞及社群網路的大量文字資料進行情緒分析為預測依據，以期提升股價走勢預測準確率，謝仁堡（2018）就採用語意分析、文本探勘分析新聞文章與股市價格的關係來預測未來股價走勢，並使用 LSTM 評估結果。

然而在 Google 提出 Transformer 模型和 BERT 模型後，在國外研究中發現，Stepka （2019）以通用電氣公司（General Electric Company）股價作為輸入資料，Jan Schmitz （2020）以 IBM 股價作為輸入資料，研究使用 Transformer 來預測股價走勢，但是國內並未看到有關 Transformer 針對股價趨勢的研究。使用 BERT 來對新聞進行情緒分析，也鮮少有關 BERT 針對股市新聞情緒分析的研究。

第二節 研究目的

基於上述研究動機，本研究目的為使用自然語言處理新技術 BERT 對台積電股市新聞進行正負面新聞分類，並利用深度學習新技術 Transformer 來預測台積電台灣股價走勢，探討新技術對預測股價走勢的影響。

依據研究目的，劃分以下三點：

1. 比較 BiLSTM 和 Transformer 模型預測台積電台灣股價走勢

2. 建立 BERT 新聞分類模型，分析影響台積電股價走勢的新聞

3. 結合兩邊的結果與真實的股價走勢進行比較

第貳章 文獻探討

第一節 深度學習

一、長短期記憶網絡（Long Short-Term Memory Network, LSTM）

長短期記憶網絡（LSTM）是循環神經網絡（Recurrent Neural Network, RNN）中一個特殊的類型。RNN 因為具有循環的結構，所以能夠記憶較短期與保留過去學習到的內部狀態，因此適用於處理序列變化的數據。但如果需要透過更多過去相關訊息來預測時，相關訊息和預測的內容之間的間隔越來越大，將產生梯度消失和梯度爆炸問題，造成 RNN 無法透過之前的訊息來進行預測，因此有學者 Hochreiter & Schmidhuber 在 1997 年提出 LSTM，它能夠記住長時間訊息來避免長期依賴（long-term dependency）的問題（Hochreiter and Schmidhuber 1997）。

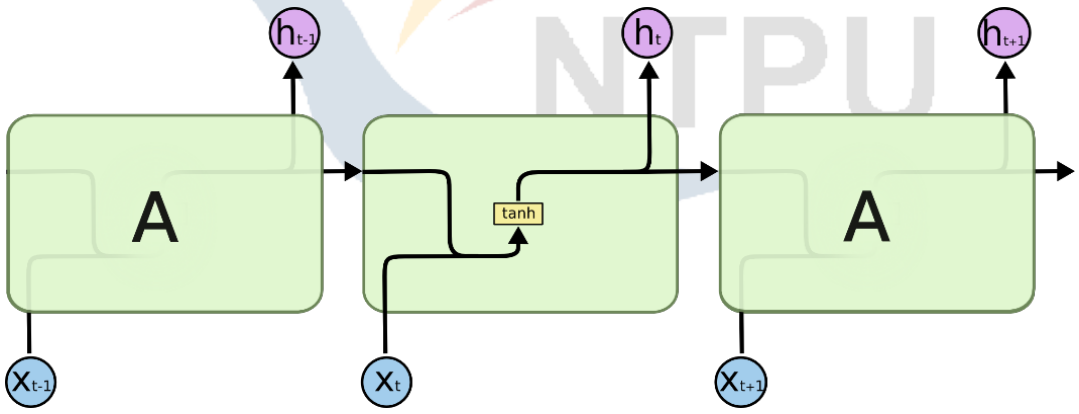


圖 1、普通 RNN 內部結構

資料來源：（http://colah.github.io/posts/2015-08-Understanding-LSTMs/）

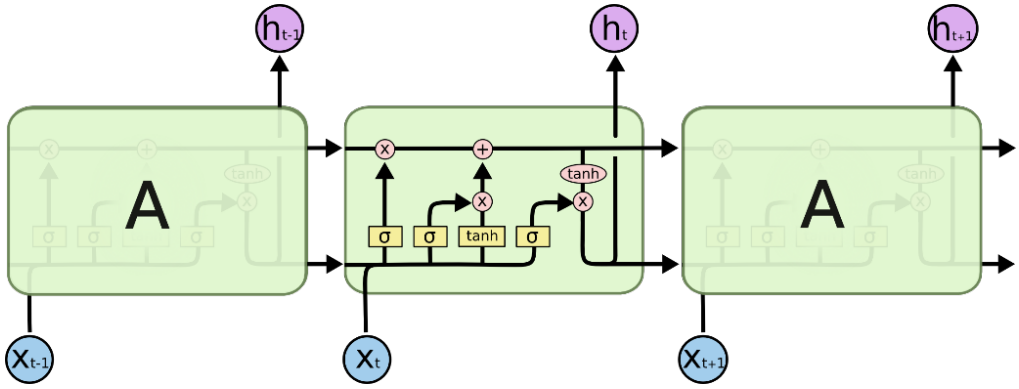


圖 2、LSTM 內部結構

資料來源：（http://colah.github.io/posts/2015-08-Understanding-LSTMs/）

普通 RNN 結構（圖 1）和 LSTM 結構（圖 2）的區別是普通 RNN 使用一個單一的 tanh 層，而 LSTM 則是使用四個相互作用的層。x（t）為當前狀態下數據的輸入，h（t-1）表示接收到的上一個節點的輸入。h（t）為當前節點狀態下的輸出，而 c（t）為傳遞到下一個節點的輸出。

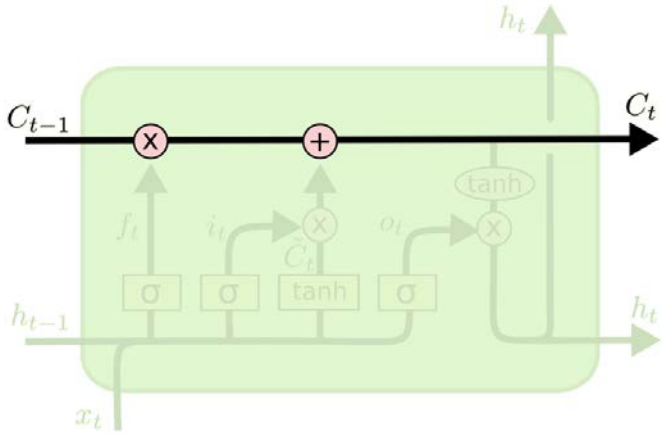
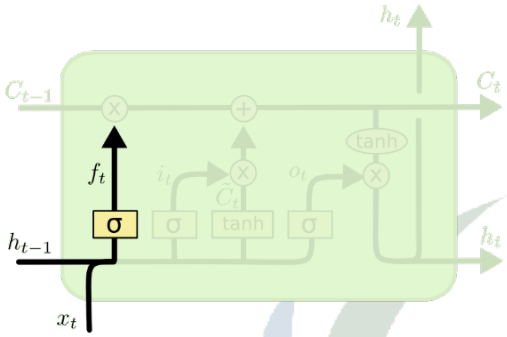


圖 3、傳送帶結構

資料來源：（http://colah.github.io/posts/2015-08-Understanding-LSTMs/）

LSTM 結構包含 cell state、遺忘門（forget gate layer）、傳入門（input gate layer）、輸出門（output gate layer）。LSTM 的核心在於 cell state（圖 3 中的黑色水平線），也就是可以讓訊息長時期被記憶保留的地方。cell state 的訊息傳輸就像經由一條傳送帶傳送，但是卻無法增加或刪除訊息，需要透過遺忘門、傳入門、輸出門這三個門結構（gates）來控制讓訊息被記憶。

(一)、遺忘門（forget gate layer）



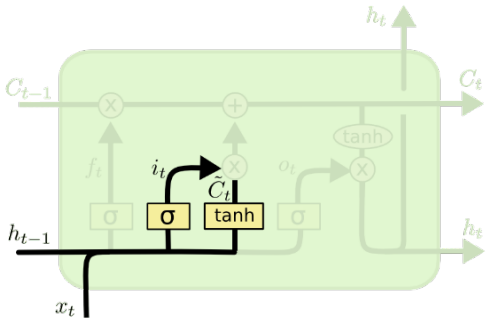
$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

圖 4、Forget gate layer

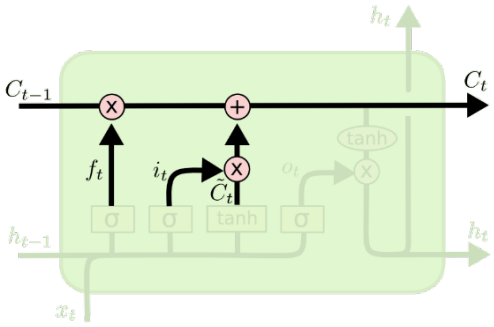
資料來源：（http://colah.github.io/posts/2015-08-Understanding-LSTMs/）

Forget gate layer 主要是決定要遺忘哪些訊息，藉由 h(t-1)與 x(t)經過 sigmoid 神經層來決定的，輸出內容為 0 到 1 之間的數值，0 代表不讓訊息通過而要被捨棄，1 則代表讓訊息通過而要保留下來。

(二)、傳入門（input gate layer）



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$



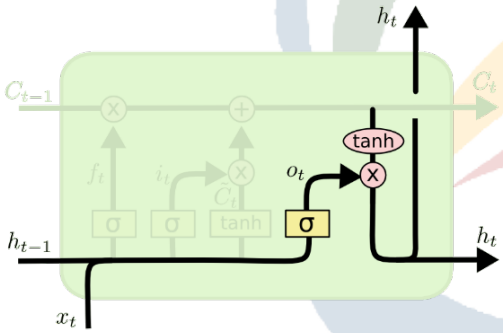
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

圖 5、Input gate layer 更新 cell state

資料來源：（http://colah.github.io/posts/2015-08-Understanding-LSTMs/）

Input gate layer 主要是決定在 cell state 中加入多少新的訊息，先藉由 x（t）與 h（t-1）經過 sigmoid 神經層來決定哪些訊息需要更新，再經過 tanh 產生新的訊息，最後對 cell state 進行更新。

(三)、輸出門（output gate layer）



$$o_t = \sigma\left(W_o\left[h_{t-1}, x_t\right] + b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

圖 6、Output gate layer

資料來源：（http://colah.github.io/posts/2015-08-Understanding-LSTMs/）

Output gate layer 是決定要輸出的訊息，藉由 x（t）與 h（t-1）經過 sigmoid 神經層來決定要被輸出的訊息，再通過 tanh 層來把數值轉換到-1 和 1 之間，最後將 sigmoid 值與 tanh 層輸出的權重相乘，最後可以得到輸出的訊息。

LSTM 因為有長期記憶的特性，可以避免遺忘重要特徵和整個序列，同時又可以像 RNN 一樣保留短期記憶，因此適用於具有時間順序的序列數據，例如語音分析、情感分析、語音識別和財務分析。

二、雙向 LSTM（Bidirectional LSTMs, BiLSTM）

雙向 LSTM 是 LSTM 模型的擴展，它是在輸入序列上訓練兩個 LSTM，而不是一個 LSTM。在輸入序列上第一個是使用原數據（即從左到右），第二個是數據的相反形式（即從右到左），透過兩次的輸入數據可以改善學習的長期依賴性，因此可以提高模型的準確性。Siami-Namini, S., Tavakoli, N., & Namin（2019）比較 LSTM 與 BiLSTM 應用在預測時間序列的成效，結果顯示 BiLSTM 模型提供更好的預測。

三、注意力機制（Attention mechanism）

注意力機制（Attention mechanism）由 Bahdanau, Cho & Bengio 在 2014 年提出的，是使用基於 RNN 或 LTSM 的 Encoder-Decoder 架構的機制，目前已經應用在許多類型的深度學習領域中，像是機器翻譯、語音識別、圖像標註等（Bahdanau, Cho et al. 2014）。近年來，開始將 Attention mechanism 應用在時間序列預測的工作，來改進 RNN 或 LSTM 模型，使結果和效率大幅提升。（Li, Zhu et al. 2019）

(一)、Encoder-Decoder 架構

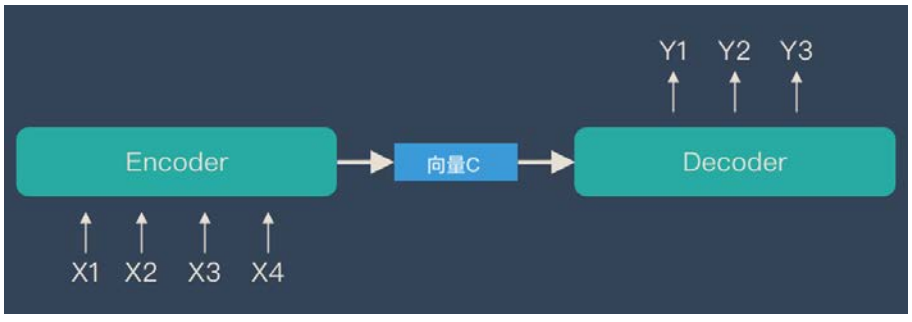


圖 7、Encoder-Decoder 架構

資料來源：（https://easyai.tech/ai-definition/encoder-decoder-seq2seq/）

傳統 Encoder-Decoder 架構的 RNN 或 LTSM 模型雖然在許多領域中取得不錯的結果，但有一個問題是不論輸入或輸出序列的內容長短，Encoder 和 Decoder 中間只有一個固定長度的「向量 C」來傳遞訊息（如圖 7），而 Encoder 和 Decoder 就會受限於固定長度向量來表示，因此當輸入序列的內容越長時，就越難保留全部的訊息，可能會失去一些重要訊息，使得學習效果很差，而造成模型的結果變差。（Bahdanau, Cho et al. 2014）

(二)、Encoder-Decoder 架構 + 注意力機制（Attention mechanism）

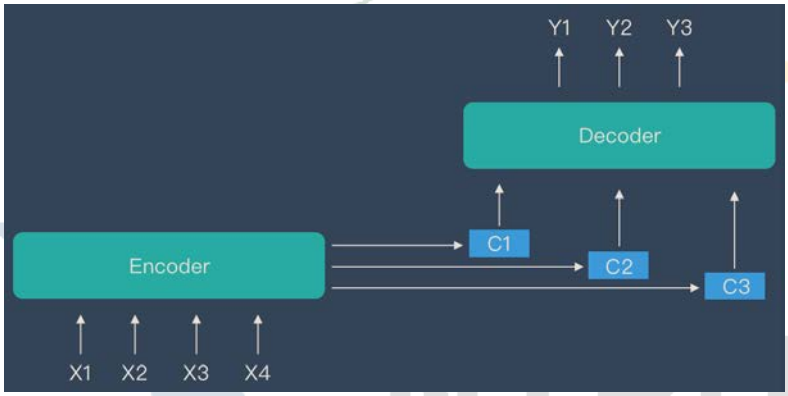


圖 8、Encoder-Decoder 架構 + Attention 機制

資料來源：（https://easyai.tech/ai-definition/encoder-decoder-seq2seq/）

Attention 機制是改變了傳統 Encoder-Decoder 架構都依賴於一個固定長度向量 C 來傳遞訊息的限制，也就是為了解決因輸入序列太長而遺漏重要訊息的問題。Encoder 和 Decoder 之間不再是固定長度向量 C，而是 Encoder 中每一次中間產生的向量結果，最終形成的一個向量的序列 C1、C2、C3（如圖 8）來傳送，因此就能將全部重要訊息皆傳遞給 Decoder。

四、Transformer 模型

Transformer 模型是由 Google 在 2017 年 6 月發表的論文「Attention Is All You Need」中提出，是一個完全基於 Attention 機制的 Encoder-Decoder 架構的模型，完全取代傳統 CNN、RNN 或 LTSM。而 RNN 的運作方式存在一個問題，就是無法有效地平行運算，如圖 9 所示，RNN 需經過 4 個時間點依序看過 [a1, a2, a3, a4] 後，才能取得序列中最後一個元素的輸出 b4。因此參考了 Attention 機制，提出了自注意力機制（Self-Attention mechanism），這個機制不只跟 RNN 一樣可以處理序列數據，還可以平行運算。

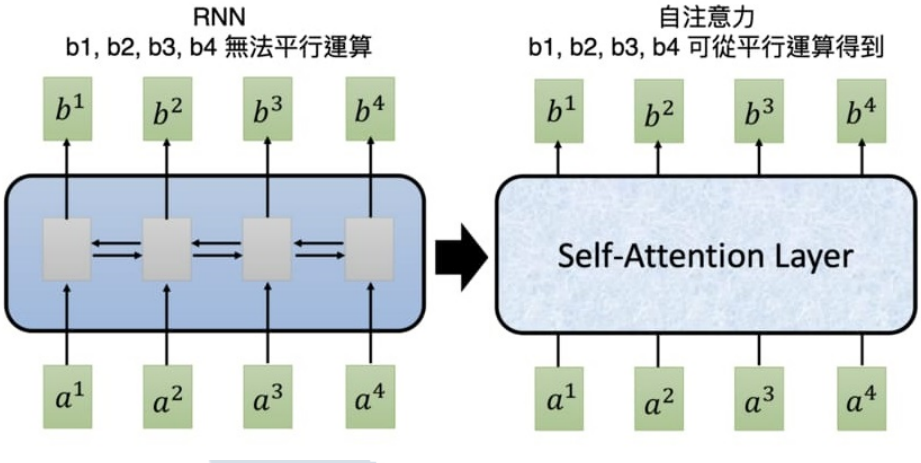


圖 9、RNN 跟 Self-Attention 的比較圖

資料來源：（https://easyai.tech/ai-definition/encoder-decoder-seq2seq/）

Transformer 的結構是由 Self-Attention 和 Feed Forward Neural Network 所組成，如圖 10 所示，左邊為 Encoder，右邊為 Decoder。Transformer 中簡化的 Encoder-Decoder 版本如圖 11 所示。

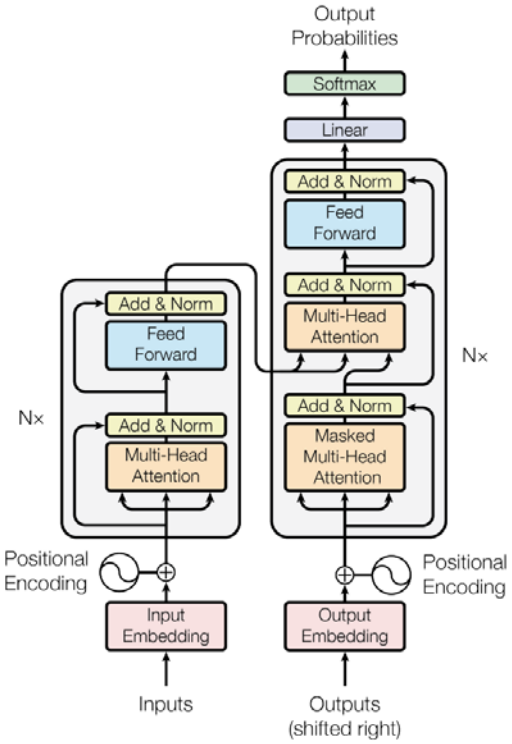


圖 10、Transformer 模型架構

資料來源：(Vaswani, Shazeer et al. （2017）. Attention Is All You Need.）

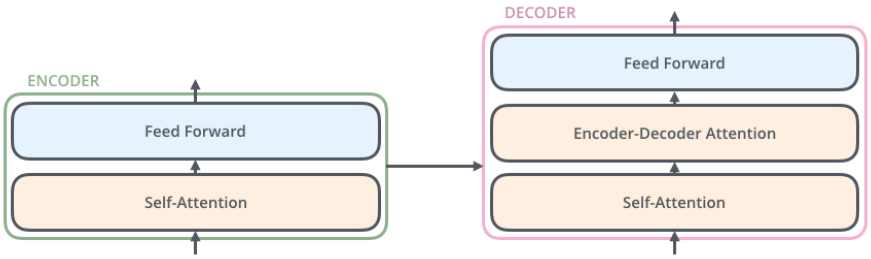


圖 11、Transformer 的 Encoder-Decoder 架構簡化版

資料來源：(http://jalammar.github.io/illustrated-transformer/)

在 Transformer 中，Encoder 跟 Decoder 各自有 Self-Attention 和 Feed Forward。Encoder 的輸入序列會先經過 Self-Attention 處理加權後，得到特徵向量並傳送給 Feed Forward。但 Decoder 不同的是多了一個 Encoder-Decoder Attention，它是接收 Encoder 產生的特徵向量與 Decoder 的 Self-Attention 產生特徵向量總和後，傳送給 Feed Forward。

第二節 BERT

BERT（Bidirectional Encoder Representations from Transformers）是在 2018 年 10 月由 Google AI 團隊發表的論文「BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding」中提出，並公開 BERT 的 TensorFlow 代碼。同年 11 月，再次發布 BERT 的多語言模型和中文模型，而透過維基百科中文語料庫來做過訓練的 BERT 中文模型，已經掌握中文的基本規律，因此只需要微調少量訓練資料，就能讓 BERT 處理所需要的任務。

BERT 使用 Transformer 一種學習文本中單詞之間上下文關係的 Attention 機制。Transformer 包括兩個獨立的機制：一個是讀取文本輸入的編碼器 encoder，另一個是生成任務預測的解碼器 decoder （Vaswani, Shazeer et al. 2017）。由於 BERT 的目標是生成語言模型，所以只需要 encoder 機制。BERT 可以用於各種 NLP 任務，只需在核心模型中添加一層，例如：在分類任務中如情感分析，只需要在 Transformer 的輸出之上加一個分類層。

BERT 是第一個深度雙向的、無監督的語言系統，無監督代表只使用純文本語料庫進行預訓練（Devlin, Chang et al. 2018）。預訓練的詞嵌入向量（word embeddings）表達可以是上下文無關（context-free）的，也可以是上下文相關（contextual）的，而且上下文相關的表示可以是單向的或雙向的。舉例來說：

- 上下文無關的模型，比如 word2vec （Mikolov, Sutskever et al. 2013）或 GloVe（Pennington, Socher et al. 2014），會為詞彙中的每個單詞生成一個詞嵌入向量（word embeddings）。例如，「bank」這個詞在 bank account（銀行帳戶）和 river bank（河岸）中將具有相同的表示。

- 上下文單向模型只能根據左側（或右側）的單詞來生成每個單詞的表示。例如，在句子“I accessed the bank account”中，「bank」的單向表示為「I accessed the」，而不是 account。

- 上下文雙向模型（BERT）會使用每個單詞左側和右側的前後文來表示。例如，在句子“I accessed the bank account”中，「bank」單詞的雙向表示為「I accessed the ... account」。

BERT 與 OpenAI GPT、ELMo 上下文預訓練方法相比，模型結構的差異如圖 12 所示，OpenAI GPT 是使用從左到右的 Transformer，ELMo 則是使用串聯獨立訓練的從左到右 LSTM 和從右到左 LSTM 來生成任務的功能，BERT 是使用雙向 Transformer，且在所有層中共同依賴於左右上下文，因此 BERT 是深度雙向模型，OpenAI GPT 是單向模型，ELMo 是淺雙向模型。

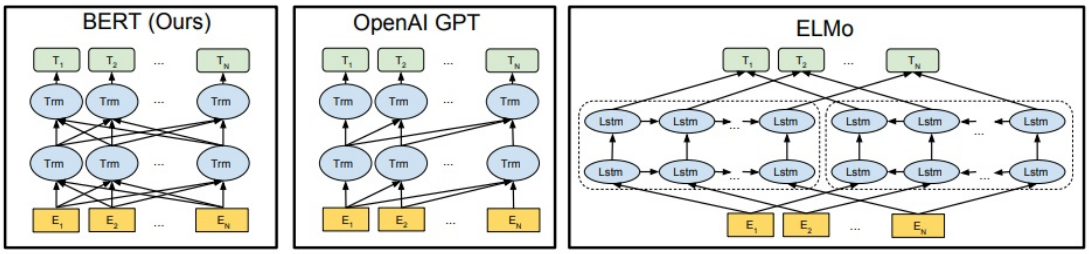


圖 12、BERT、OpenAI GPT、ELMo 模型的結構

資料來源：（https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html）

第參章 研究方法

第一節 研究架構

本研究的提出應用深度學習與自然語處理研新技術預測股價走勢的架構如圖 13 所示：

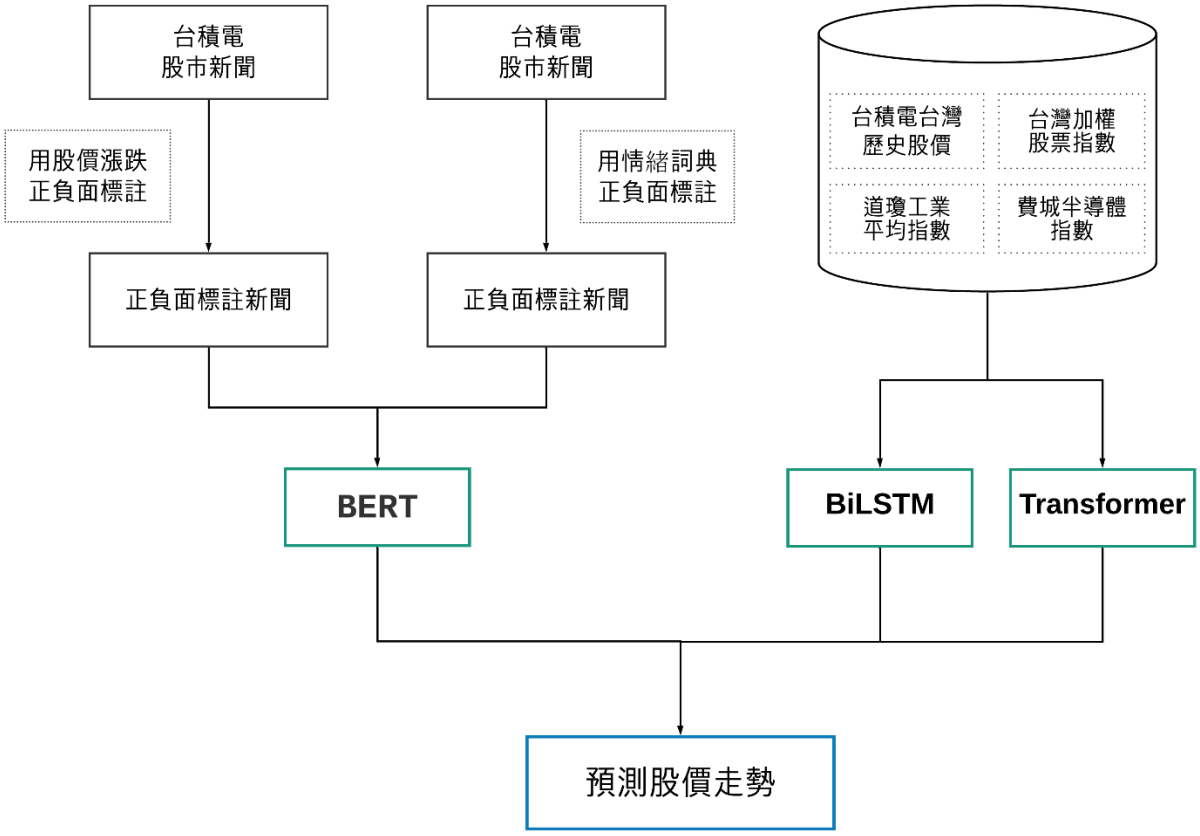


圖 13、研究架構

資料來源：本研究

本研究使用自然語言新技術 BERT 對台積電股市新聞進行正負面新聞分類，使用深度學習新技術 Transformer 與 BiLSTM 來預測台積電台灣的股價走勢，輸入的資料包含台積電的台灣歷史股價、台灣加權股票指數、道瓊工業平均指數以及費城半導體指數，最後將兩邊的結果與真實的股價走勢進行比較。

第二節 資料蒐集

本研究以台積電為研究對象，使用 Yahoo 股市的台灣與美國歷史股價，採用 2015 年 6 月 1 日至 2020 年 5 月 31 日的每日開盤價、收盤價、最高價、最低價、調整後收盤價、成交量作為數據集，為了更瞭解影響台積電股價的原因，需要從不同的方面和角度來描述股票，因此除了台積電的歷史股價外，還要整合不同類型的數據作為輸入特徵。

在過去有研究結果顯示美國股市與台灣股市的關聯性中，道瓊工業指數是最具影響力，其次是那斯達克指數或費城半導體指數，劉照群（2008）研究美國費城半導體指數、台灣股市、台積電股價的關聯性，結果顯示三者存在著長期均衡關係，且費城半導體指數影響台積電股價比台股更為強烈。

對於基本面分析，本研究使用台灣股市與美國股市中會影響台積電的指數作為輸入特徵之一，包括台灣加權股票指數（TSEC weighted index，TAIEX）、道瓊工業平均指數（Dow Jones Industrial Average，DJIA）、費城半導體指數（PHLX Semiconductor，SOX）及那斯達克指數（NASDAQ）。另外搜集台積電從 2014 年 7 月 1 日至 2020 年 5 月 31 日所有關於台積電的每日新聞進行情緒分析（正面的、負面的、中性的）作為特徵之一。

對於技術面分析，投資者遵循的技術指標也是作為輸入特徵之一，包括 7 和 21 天移動平均線（Moving Average, MA）、指數平均數指標（Exponential Moving Average, EXPMA / EMA）、動量指標（Momentum, MTM）、布林通道（Bollinger Bands）、指數平滑異同移動平均線（Moving Average Convergence / Divergence, MACD），為 1970 年 Richard Donchian & J.M.Hurst 在「The Profit Magic of Stock Transaction Timing」這一本書中所提到的方法。

一、 台積電股票價格及相關指數

本研究使用 Python 套件 fix_yahoo_finance 來搜集台積電的台灣歷史股價（股票代碼為 2330.TW）與美國歷史股價（股票代碼為 TSM）、台灣加權股票指數（股票代碼為^TWII）、道瓊工業平均指數（股票代碼為^DJI）、費城半導體指數（股票代碼為^SOX）及那斯達克指數（股票代碼為^IXIC），時間為 2015 年 6 月 1 日至 2020 年 5 月 31 日，股價資料整合如表 2-1，只取前三欄的資料。每個股票蒐集的資訊內容皆包含「Date」、「Open」、「High」、「Low」、「Close」、「Adj Close」、「Volum」，每個欄位描述如表 2-2。

表 2-1、股價資料表-以台積電台灣股價為例

Date	open	high	low	close	adj close	volum
2014/7/1	126.5	128.5	126	128	127.25	33022.54
2014/7/2	129.5	132.5	129	132.5	130.75	51623.75
2014/7/3	131.5	135	130.5	134.5	132.75	40283.58

資料來源：本研究

表 2-2、股票資料庫欄位表

欄位	描述
Date	股價日期
Open	當日開盤股價
High	當日最高股價
Low	當日最低股價
Close	當日收盤價
Adj Close	調整後的收盤價
Volume	成交量

資料來源：本研究

二、 鉅亨網之台積電股市新聞

本研究使用 python 中的 requests 與 json 套件，來爬取鉅亨網（https://www.cnyes.com） 上的台積電相關股市新聞，時間為 2015 年 6 月 1 日至 2020 年 5 月 31 日，每個新聞蒐集的資訊內容皆包含「Title」、「Content」、「TM」、「Url」，每個欄位描述如下表 2-3。

表 2-3、股市新聞資料庫欄位表

欄位	描述
Title	新聞標題
Content	新聞內文
TM	新聞發佈時間
Url	新聞網址

資料來源：本研究

第三節 資料處理

一、 自相關分析

自相關性則是指一個時間序列的兩個不同時間點的數值是否相關聯。本研究想利用自相關分析來瞭解股票價格在各個時間點的數值是否具有相關聯，若自相關性為 0，則表示各個時間點的數值不相關，因此就無法根據過去的股價來預測未來股價。

本研究將台積電的台灣歷史股價與美國歷史股價、台灣加權股票指數、道瓊工業平均指數及費城半導體指數及那斯達克指數進行自相關分析。

自相關分析結果為圖 14，x 軸代表現在與過去相距的天數，最左邊相關性數值最高，而越往右邊相距的天數越多，則相關性數值越低，但 y 軸上顯示的相關性數值為-1 到 1 之間，代表各特徵的現在與過去歷史資料之間具有相關性，因此台積電股價及相關指數是可以根據過去的股價來預測未來的股價。

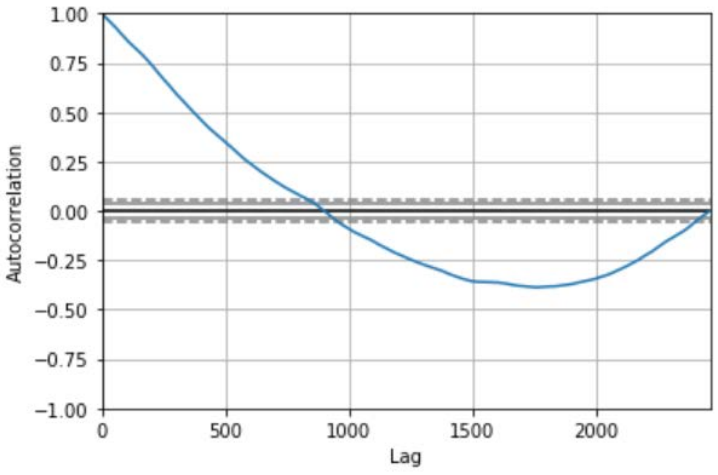


圖 14、自相關分析（Autocorrelation）

資料來源：本研究

二、 相關係數矩陣分析

相關係數表示兩組數值向量之間的關聯性，相關係數的值介於 1 與-1 之間，越靠近 1 代表正相關程度越高，出現的數值最大為 1，表示兩個數列呈現完全正相關，相反越靠近-1 則代表負相關程度越高，出現數值最小為-1，則表示兩個數列完全負相關，兩者幾乎完全相反，而越靠近 0 則代表兩數值之間的關係越微弱。

本研究將台積電的台灣歷史股價與美國歷史股價、台灣加權股票指數、道瓊工業平均指數及費城半導體指數及那斯達克指數進行相關係數矩陣分析，以瞭解股價與指數之間關係的緊密程度。

相關係數矩陣結果為圖 15，顯示相關係數為 0 到 1 之間，代表各特徵之間為正相關，因此台積電股價與相關指數之間是具有關聯性，互相會受到影響。台積電台灣股價與台積電美國股價、那斯達克指數的相關性過高，為了避免模型過度模型擬合，導致預測準確結果不佳，因此本研究將排除這兩個輸入資料。

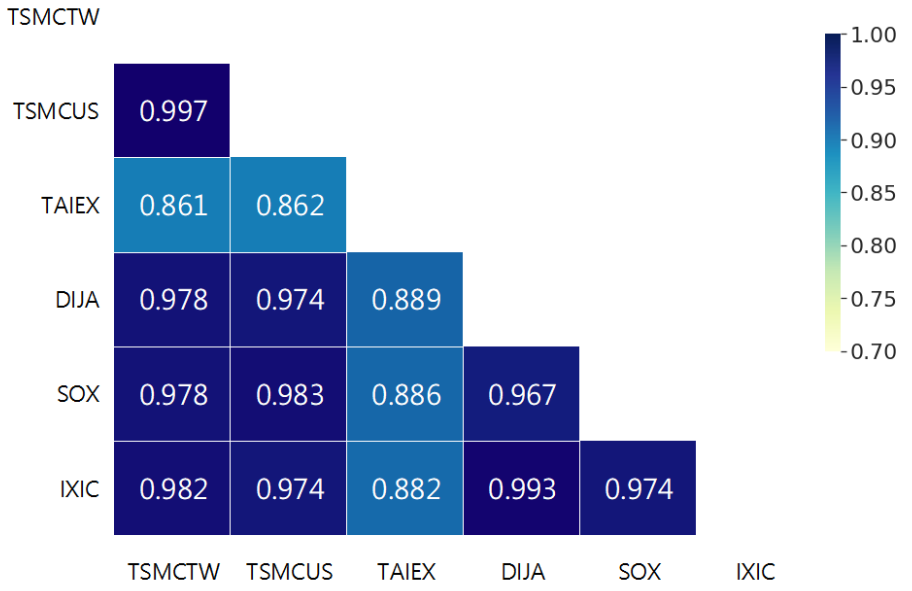


圖 15、相關矩陣（Correlation matrix）

資料來源：本研究

三、 CKIP Tagger

常見的斷詞工具，通常都會使用結巴（jieba）、CKIP，而 CKIP 是由台灣中央研究院中文詞知識庫小組所開發的中文分詞系統，在中文的自然語言處理中，是斷詞最精準的工具。之前使用 CKIP 都需要到官方網站試用，或是要寄信取得授權，直到 2019 年 9 月，更新版本後，將這套工具 CKIP Tagger 開放原始碼在 Github（https://github.com/ckiplab/ckiptagger）上，並且可以根據需求來調整原始碼。從官方提供的資料顯示，在成效方面與結巴（jieba）比較，精確度（Precision）從 90.51%提升至 97.49%，而準確度（Accuracy）從 89.10%提升至 94.59%，整體成效提升許多。

本研究蒐集的台積電相關股市新聞為中文繁體，因此使用 CKIP Tagger 來達到更好的中文斷詞效果。如表 3-1 為中文斷詞結果範例。

表 3-1、中文斷詞結果範例

台股今 （19） 日由台積電領軍下帶量上漲，加權指數收在 10873.19 點，單週上漲 48 點週線順利收紅，成交量回升至 1225.67 億元，上市公司市值單週增加 1466.61 億元，總市值共 32.85 兆。。
['台股', '今', '19', '日', '由', '台積電', '領軍下', '帶量', '上', '漲', '加權', '指數', '收', '在', '1087319', '點', '單週', '上', '漲', '48', '點', 週線', '順利', '收紅', '成交量', '回升', '至', '12 2567', '億元', '上市公司', '市值', '單週', '增加', '146661', '億元', '總', '市值', '共', '3285', '兆', '。']

資料來源：本研究自行整理

四、 新聞情緒標籤

本研究將使用兩種方式對新聞進行情緒標籤，一種是依據股價漲跌幅度作標註，一種是利用情緒詞典作標註。第一種方式是依據股價漲跌幅度作標註，計算的方式為新聞當日的收盤價與上一交易日的收盤價之差，計算出的結果，若大於 0%，則標註為正面，小於 0%，則標註為負面，而等於 0%，則標註為中立。

$$\text{公式：}\frac{\text{（當日收盤價 − 上一交易日的收盤價）}}{\text{上一交易日的收盤價}} \times 100\%$$

第二種方式是利用情緒詞典作標註。文本情緒分析的最常見的分析方法就是使用情緒詞典，在進行情緒分析時，需要有匯集不同情緒和情感的詞，並標記每個詞為正負詞性的詞典，來作為分析的依據。針對在不同領域，所建立的情緒詞典，更能提升情緒分析時的準確度。

在中文情緒詞典中，詞主要分為正面詞和負面詞，台灣大學自然語言處理實驗室所建立的意見詞詞典（NTUSD），裡面包含了正面情緒詞大約有 2800 筆以及負面情緒詞大約有 8000 筆（陳韋帆, 2018）。新聞文字是可以利用情緒分析來判斷文章想要表達的語氣和情緒，因此本研究使用兩種一般情緒詞典（How-net、NTUSD），加上一種金融領域詞典（iMFinanceSD），並賦予每個詞一個情緒分數，正面詞情緒分數為 1，負面詞情緒分數為 -1。

作法是先將新聞文本使用斷詞工具 CKIP Tagger 斷詞後，利用情緒詞典對新聞文章有出現在詞典裡的正負詞標記情緒分數，若標示為 1，則為正面詞，反之標示為 -1，則為負面詞，而標示為 0，則不代表正或負面情緒。最後計算每則新聞的正或負面情緒詞個數總計並相減，正面情緒詞個數-負面情緒詞個數>0，這則新聞標記為正面新聞（Positive），反之<0，則標記為負面新聞（Negative），而剛好等於 0，則標記為中性新聞（Unrelated）。

第四節 資料分析

一、 BERT 模型流程

經過兩種方式將 2015 年 6 月 1 日至 2020 年 5 月 31 日的每則台積電股市新聞標記正面（Positive）、負面（Negative）或中性（Unrelated）的情緒標籤後，產生兩種資料集，並各別將資料集分成訓練集為 2015 年 6 月 1 日至 2018 年 5 月 31 日與測試集為 2018 年 6 月 1 日至 2020 年 5 月 31 日。將訓練集資料放進 BERT 中，再進行 Fine-tune，最後再將測試集資料放入進行訓練，最後評估成效並比較。

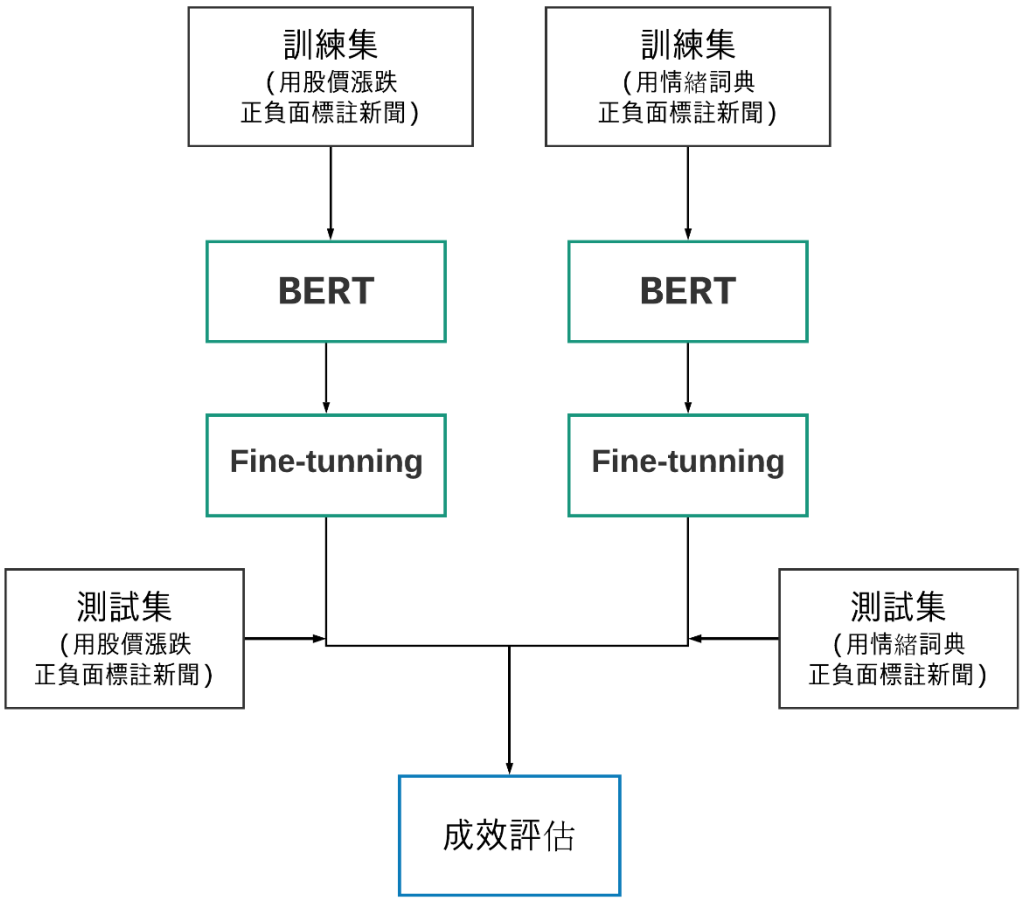


圖 16、BERT 模型流程圖

資料來源：本研究

二、 Transformer 模型

搜集 2015 年 6 月 1 日至 2020 年 5 月 31 日台積電的台灣歷史股價、台灣加權股票指數、道瓊工業平均指數以及費城半導體指數作為輸入資料，並將資料集分成訓練集（70%）為 2015 年 6 月 1 日至 2018 年 11 月 30 日，驗證集（15%）為 2018 年 12 月 1 日至 2019 年 8 月 31 日，測試集（15%）為 2019 年 9 月 1 日至 2020 年 5 月 31 日，放 Transformer、BiLSTM 訓練，最後預測出後一天股價走勢，且評估 Transformer 和 BiLSTM 的成效並比較。

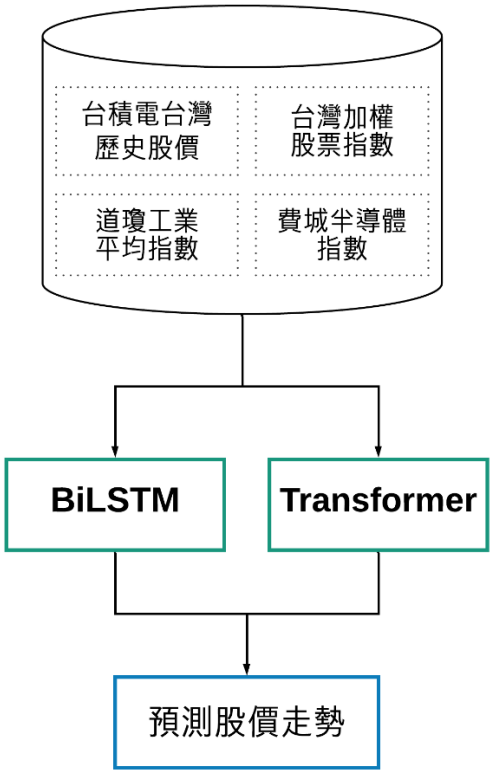


圖 17、Transformer 模型流程圖

資料來源：本研究

第肆章 研究結果

一、 模型成效分析

BiLSTM（圖 18）和 Transformer（圖 19）模型有相似的結構，比較不一樣的是 Transformer 多了一層 Time2vec，它是一種時間的向量表示形式（Kazemi and Goel 2019），可以容易地加入到模型結構中。Transformer 模型主要是利用位置的概念提供單字在句子中的位置表示，因此加入 Time2vec 是提供時間序列關係的概念，來改善模型效能。

Model: "model"

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	[(None, 128, 6)]	0	
bidirectional (Bidirectional)	(None, 128, 256)	138240	input_1[0][0]
bidirectional_1 (Bidirectional)	(None, 128, 256)	394240	bidirectional[0][0]
bidirectional_2 (Bidirectional)	(None, 128, 128)	164352	bidirectional_1[0][0]
global_average_pooling1d (Globa	(None, 128)	0	bidirectional_2[0][0]
global_max_pooling1d (GlobalMax	(None, 128)	0	bidirectional_2[0][0]
concatenate (Concatenate)	(None, 256)	0	global_average_pooling1d[0][0] global_max_pooling1d[0][0]
dense (Dense)	(None, 64)	16448	concatenate[0][0]
dense_1 (Dense)	(None, 1)	65	dense[0][0]

Total params: 713,345
Trainable params: 713,345
Non-trainable params: 0

圖 18、BiLSTM 模型結構

資料來源：本研究

```
Model: "model"
__________________________________________________________________________________________________
Layer (type)                    Output Shape         Param #     Connected to
==================================================================================================
input_2 (InputLayer)            [(None, 128, 5)]     0
__________________________________________________________________________________________________
time2_vector_1 (Time2Vector)    (None, 128, 2)       512         input_2[0][0]
__________________________________________________________________________________________________
concatenate (Concatenate)       (None, 128, 7)       0           input_2[0][0]
                                                                 time2_vector_1[0][0]
__________________________________________________________________________________________________
transformer_encoder_3 (Transfor (None, 128, 7)       99114       concatenate[0][0]
                                                                 concatenate[0][0]
                                                                 concatenate[0][0]
__________________________________________________________________________________________________
transformer_encoder_4 (Transfor (None, 128, 7)       99114       transformer_encoder_3[0][0]
                                                                 transformer_encoder_3[0][0]
                                                                 transformer_encoder_3[0][0]
__________________________________________________________________________________________________
transformer_encoder_5 (Transfor (None, 128, 7)       99114       transformer_encoder_4[0][0]
                                                                 transformer_encoder_4[0][0]
                                                                 transformer_encoder_4[0][0]
__________________________________________________________________________________________________
global_average_pooling1d (Globa (None, 128)          0           transformer_encoder_5[0][0]
__________________________________________________________________________________________________
dropout (Dropout)               (None, 128)          0           global_average_pooling1d[0][0]
__________________________________________________________________________________________________
dense (Dense)                   (None, 64)           8256        dropout[0][0]
__________________________________________________________________________________________________
dropout_1 (Dropout)             (None, 64)           0           dense[0][0]
__________________________________________________________________________________________________
dense_1 (Dense)                 (None, 1)            65          dropout_1[0][0]
==================================================================================================
Total params: 306,175
Trainable params: 306,175
Non-trainable params: 0
```

圖 19、Transformer 模型結構

資料來源：本研究

在參數的部分，BiLSTM 和 Transformer 模型的輸入資料向量皆為（32, 128, 6），表示輸入將接受 32 個序列（batch_size = 32），每個序列的長度為 128 天（seq_len = 128），每個序列日都有 6 個特徵（開盤價、最高價、最低價、收盤價、調整後收盤價、交易量）。另外使用 MSE（均方誤差）、平均絕對誤差（MAE）和平均絕對百分誤差（MAPE）做為損失函數，MSE 是預測值與真實值的誤差均值，而 MAE 和 MAPE 是用來衡量一個模型預測結果的好壞。

本研究將輸入資料分為兩種形式分別對 BiLSTM、Transformer 進行預測，第一種形式資料只單純為台積電股價資料（開盤價、最高價、最低價、收盤價、調整後收盤價和交易量），第二種則加上台積電相關指數（台灣加權股票指數、道瓊工業平均指數和費城半導體指數），因此會產生 BiLSTM、BiLSTM + 相關指數、Transformer、Transformer + 相關指數四個預測結果，比較表如下表 3-2。

表 3-2、四個模型的成效比較

		BiLSTM	BiLSTM+ 相關指數	Transformer	Transformer+ 相關指數
訓練集	MSE	0.0055	0.0054	0.0054	0.0052
	MAE	0.0545	0.0540	0.0542	0.0535
	MAPE	11.4596	11.4265	11.3660	11.3421
驗證集	MSE	0.0068	0.0067	0.0067	0.0062
	MAE	0.0650	0.0641	0.0643	0.0624
	MAPE	13.0684	12.923	12.8955	12.8436
測試集	MSE	0.0152	0.0148	0.0149	0.0142
	MAE	0.0898	0.0882	0.0876	0.0854
	MAPE	20.4863	20.4693	20.4577	20.4423

二、 新聞情緒分類分析

本研究利用兩種方式對新聞進行情緒標籤，再利用 BERT 模型進行訓練，最後評估成效並比較。在成效評估方面，將使用精確度（Precision）、準確度（Accuracy）、召回率（Recall）及 F-score 作為評估的依據，並根據表 3-3 的四個指標來計算。

表 3-3、評估指標

	實際為正面新聞	實際為負面新聞
預測為正面新聞	TP （True Positive）	FP （False Positive）
預測為負面新聞	FN （False Negative）	TN （True Negative）

資料來源：本研究

1. 正確正面（True Positive）：當預測為正面新聞，實際也為正面新聞。
2. 正確負面（True Negative）：當預測為負面新聞，實際也為負面新聞。
3. 錯誤負面（False Negative）：當預測為負面新聞，實際為正面新聞。
4. 錯誤正面（False Positive）：當預測為正面新聞，實際為負面新聞。

表 3-2 的四個指標，精確度、準確度、召回率及 F-score 的計算公式如下：

(1). 精確度（Precision）：預測為正面新聞中，實際為正面新聞的比率。

公式：Precision = TP /（TP + FP）

(2). 準確度（Accuracy）：預測所有新聞中，判斷正確的比率。

公式：Accuracy =（TP + TN）/（TP + FP + TN + FN）

(3). 召回率（Recall）：實際為正面新聞，能夠預測出為正面新聞的比率。

公式：Recall = TP /（TP + FN）

(4). F 值（F-score）：可以權衡綜合 Precision 和 Recall 兩個指標的結果。

公式：$F-Score = 2 \times Precision \times Recall \ /（Precision + Recall）$

本研究針對兩種方式情緒標籤的新聞資料集進行成效評估，結果如表 3-4。

表 3-4、成效評估結果

	股價漲跌 標籤的新聞資料集	情緒詞典 標籤的新聞資料集
精確度 （Precision）	0.625（62.5%）	0.586（58.6%）
準確度 （Accuracy）	0.641（64.1%）	0.599（59.9%）
召回率 （Recall）	0.746（74.6%）	0.635（63.5%）
F 值 （F-score）	0.681（68.1%）	0.609（60.9%）

資料來源：本研究

三、 股價走勢分析

從 Transformer 模型輸出的結果中，選出台積電 2015 年 6 月 1 日至 2020 年 5 月 31 日間，隨機找五天來做預測，並查看前一天的新聞標籤是否有與漲跌有關，最後選出五天為 2018 年 2 月 6 日、2018 年 7 月 20 日、2018 年 10 月 31 日、2020 年 1 月 30 日、2020 年 3 月 19 日。

1. 2018/02/06

Date	real close	predicted close	股價走勢	新聞標籤
2018/02/05	253			Negative（正確）
2018/02/06	239	240.62	跌（正確）	

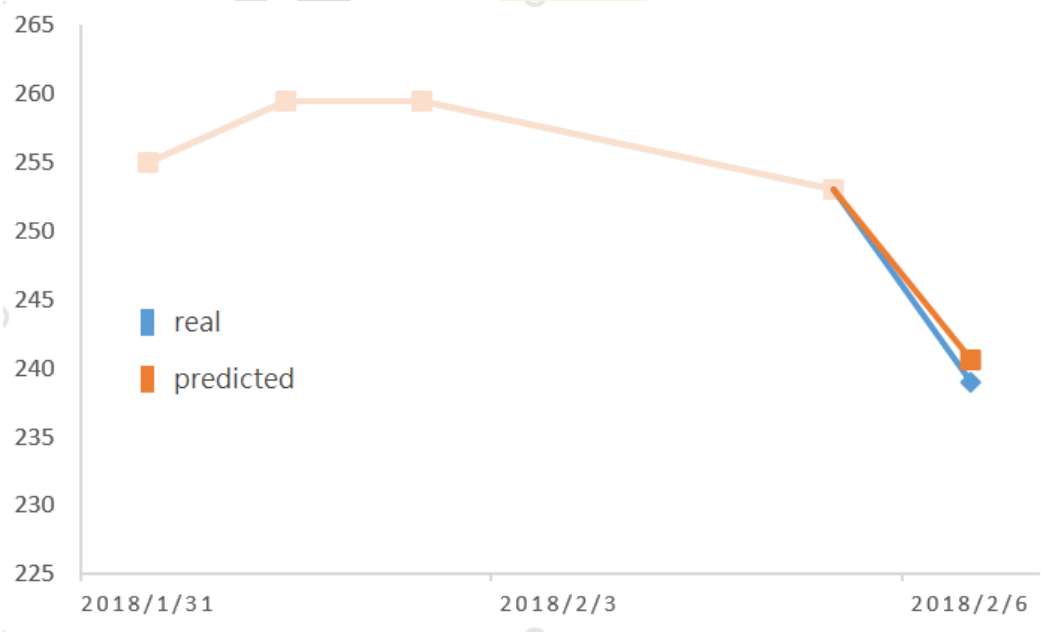


圖 20、2018/02/06 的真實和預測股價

資料來源：本研究

表 3-5、2018/02/05 的新聞

前一天 2018/02/05 的新聞標註：negative

台股早盤暴跌逾 260 點下殺 10855 點

美股暴跌跌破月線，台股今 （5） 日也難逃重挫，大盤指數開盤 10970.3 點，重挫 155.93 點，也跟跌跌破月線，台積電開盤挫低至 250.5 元，跌幅超過 3%，恐慌性賣壓不止，加權指數開低走低，最深跌到 10855.43 點、跌點逾 250 點，短線先尋求季線支撐。

美股四大指數上週五皆重挫，其中道瓊指數重挫 665.75 點、跌幅為 2.54%，連動夜盤台指期下跌 143 點，跌幅達 1.29%，收在 10971 點，創期貨夜盤上線以來最大單日跌點及跌幅。

日本及韓國股市今日也開低走低，早盤跌幅超過 2%，跌破季線。

蘋概三王指標股台積電今天最低被打到 250.5 元，陷入月線保衛戰；鴻海跳空最低跌到 92.5 元，跳空跌破所有均線；大立光最低撂壓到 3775 元，跌幅超過 4%。

[0, 0, -1, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, -1 0, 0, -1, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, -1, 0, 0, -1, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, -1, 0, 0, 0, -1, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, -1, 0, 0, -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, 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, -1, 0, 0, -1, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 1, -1, 0, -1, 0, 0, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0]

資料來源：鉅亨網、本研究

2. 2018/07/20

Date	real close	predicted close	股價走勢	新聞標籤
2018/07/19	224.5			Positive（正確）
2018/07/20	237.5	230.085	漲（正確）	

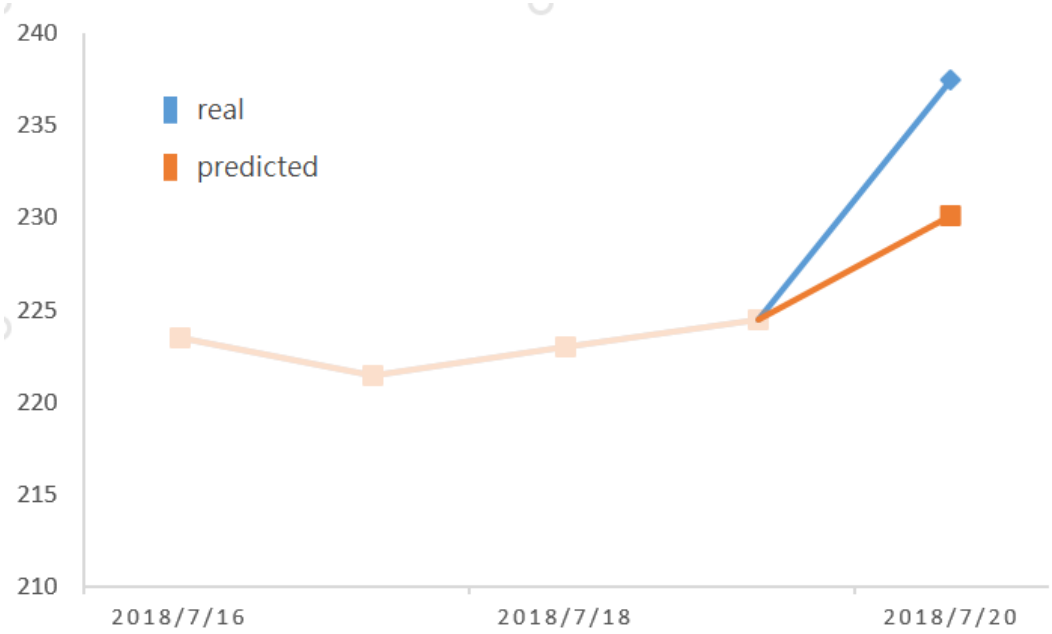


圖 21、2018/07/20 的真實和預測股價

資料來源：本研究

表 3-6、2018/7/19 的新聞

前一天 2018/7/19 的新聞標註：positive 台積電看好四大成長動能 物聯網、汽車與 HPC 正成長 手機持平 針對台積電 （2330） 所提到四大驅動半導體產業成長的動能，今年預估，智慧型手機貢獻持平到小下滑，物聯網與汽車營收貢獻均可年增二成以上，HPC 則是剛開始。台積電提到，就智慧型手機部分，營收佔比最大，預估今年是持平到微幅下滑。物聯網和汽車電子方面，目前各佔營收 6% 到 7%，預估今年營收可望年增二成以上，至於 HPC 部

分，目前剛開始，所以成長也會大。

[0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, -1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]

資料來源：鉅亨網、本研究

3. 2018/10/31

Date	real close	predicted close	股價走勢	新聞標籤
2018/10/30	223			Positive（正確）
2018/10/31	234	259.02	漲（正確）	

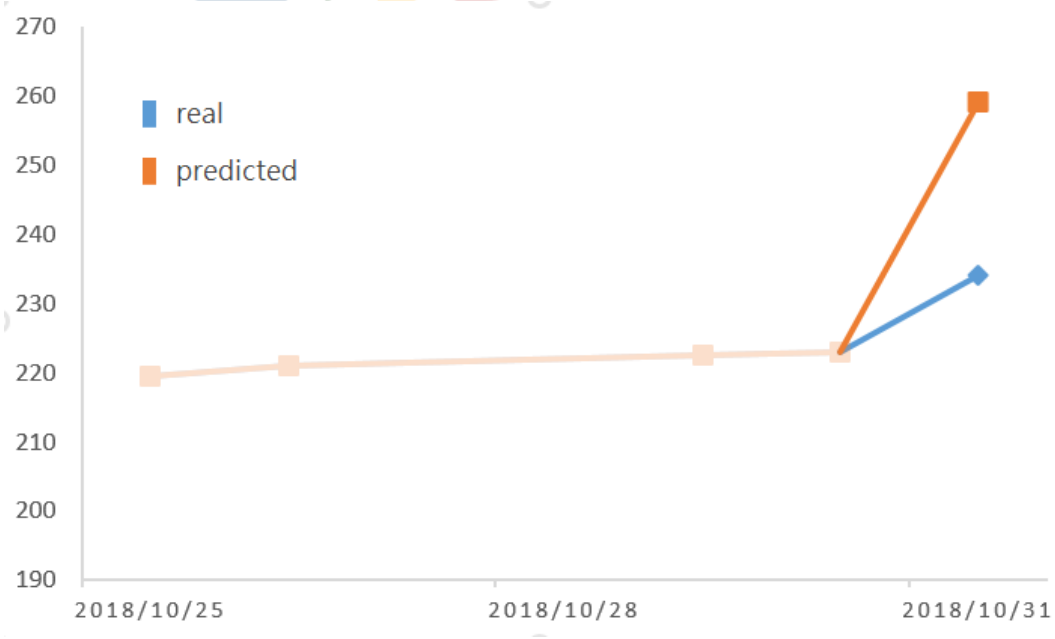


圖 22、2018/10/31 的真實和預測股價

資料來源：本研究

表 3-7、2018/10/30 的新聞

前一天 2018/10/30 的新聞標註：positive

台積電專利申請持續奪冠 鴻海年減 55%落居第七

經濟部智慧財產局今（30）日公布第 3 季智慧財產權趨勢，其中在專利申請部分，發明專利及設計專利皆小幅成長，新型專利則呈現衰退。本國法人發明專利申請則由晶圓代工龍頭台積電 （2330-TW） 持續領先，鴻海 （2317-TW） 申請件數則銳減，排名掉到第七，但總體而言法人申請件數較去年同期有所成長；外國法人專利申請則由高通（QCOM-US） 居於首位。

智慧局數據表示，第 3 季發明專利申請總量為 18328 件，年減 2%，其中發明專利年增 1%，連續 7 季正成長，設計專利微增 0.1%；新型專利的部分則呈現衰退，年減 11%。

智慧局數據也顯示，本國法人發明專利申請的部分，總申請件數較去年有所成長，其中台積電以 303 件續居首位，成長率達 35%；宏達電 （2458-TW） 及廣達 （2382-TW） 也分別有 43 件與 36 件，呈現 73% 與 125% 的高成長率；鴻海申請件數則為 51 件，較去年同期減少了 55%，從 Q2 的第 5 落至第 7。

智慧局數據也指出，外國法人專利申請中，高通以 234 件申請數領先阿里巴巴的 132 件，及東芝記憶體的 115 件，成長率更高達 73%，其前 3 季累積件數更大幅成長 1 倍以上，發明佈局在外國法人之中最為積極。

[0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 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, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1, 0, 0, 1, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, -1, 1, 0, 0, 0, 0, 0, 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, -1, 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, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, -1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1]

資料來源：鉅亨網、本研究

4. 2020/01/30

Date	real close	predicted close	股價走勢	新聞標籤
2020/01/20	333			Positive（錯誤）
2020/01/30	316.5	336.4	漲（錯誤）	

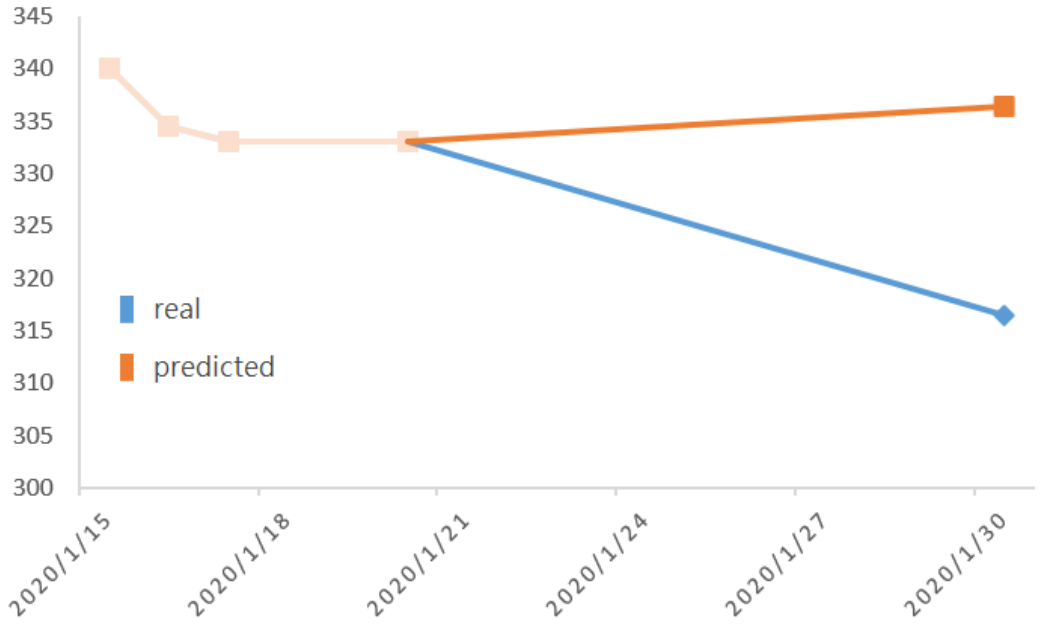


圖 23、2020/01/30 的真實和預測股價

資料來源：本研究

表 3-8、2020/1/20 的新聞

前一筆資料的日期 2020/1/20 的新聞標註：positive

金豬年封關三大法人買超 61 億元 加碼群創、大賣中石化

台股今 （20） 日迎接豬年封關日，終場小漲 28.42 點，收在 12118.71 點，三大法人合計買超 61.16 億元，其中，外資、自營商分別買超 51.36、12.05 億元，投信則賣超 2.25 億元，金豬年最後一個交易日，外資、投信持續聚焦面板族群，合計買超群創近 20000 張，反觀中石化則遭大賣逾 50000 張。

期貨部分，外資今日減碼多單 592 口，多單未平倉口數達 31789 口，投信減碼 518 口空單，空單未平倉口數來到 22642 口，自營商則加碼 919 口多單，合計多單未平倉口數 2162 口。

外資今日買超前十名，群創 17907 張居冠、中信金 13810 張、友達 10055 張、晶電 9597 張、鴻海 8865 張、欣興 8000 張、元大金 7583 張、聯電 5698 張、永豐金 4847 張、仁寶 4813 張。外資今日賣超前十名，中石化遭賣 50981 張居冠、台積電 3628 張、京元電 3053 張、世界 2694 張、遠東新 2360 張、和碩 2298 張、富邦 VIX 2277 張、南茂 2250 張、英業達 1704 張、瑞儀 1495 張。

投信前十大買超標的，群創 1917 張居冠、和鑫 1582 張、泰鼎 - KY 1173 張，其他依序為聯茂、威剛、國產、中美晶、欣興、精元、技嘉。投信賣超前十大標的，京元電 2753 張、頎邦 1020 張，其他依序為台燿、臻鼎 - KY、華航、台積電、中信金、開發金、矽創、中鋼。

台股今日為金豬年封關日，金豬年以來，加權指數大漲 22%，一度創下台股近 29 年多來新高 12197 點紀錄，法人認為，只要農曆春節期間未出現黑天鵝事件，開紅盤行情仍可期，而隨著台積電前景樂觀帶動下，金鼠年後加權指數可望再挑戰改寫歷史新高。

[0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 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, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, -1, 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, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 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, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 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, -1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, -1, 0, 0, 1]

資料來源：鉅亨網、本研究

5. 2020/03/19

Date	real close	predicted close	股價走勢	新聞標籤
2020/03/18	260			Negative（正確）
2020/03/19	248	258.3	跌（正確）	

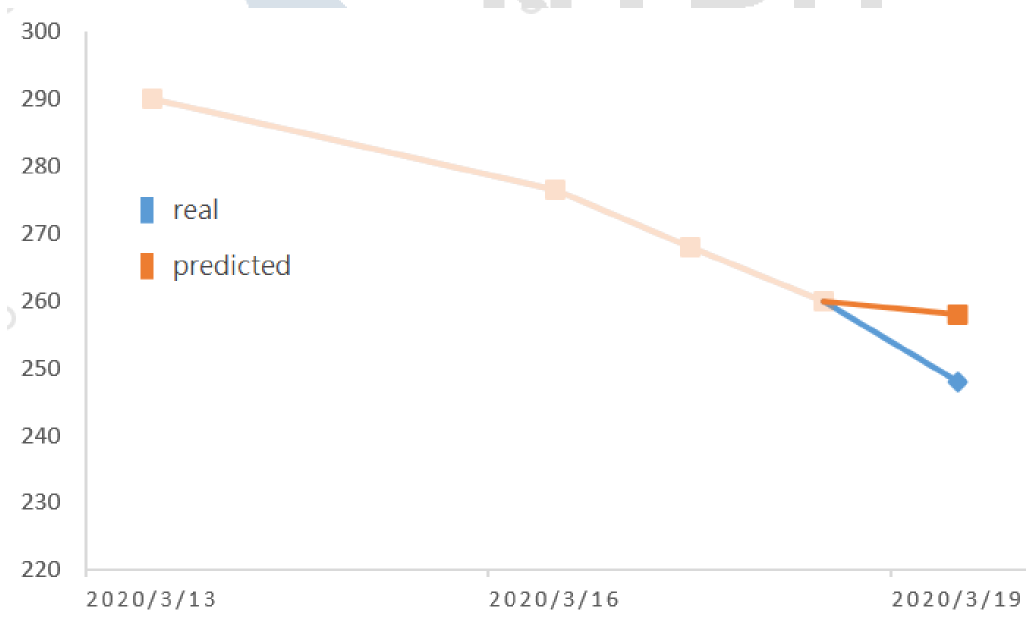


圖 24、2020/03/19 的真實和預測股價

資料來源：本研究

表 3-9、2020/3/18 的新聞

前一天 2020/3/18 的新聞標註：negative

美股破底反彈 台指期自 44 個月低點強彈近 300 點

台指期夜盤 18 日美股期重創下，最低來到 8550 點，創 2016 年 7 月 7 日以來 44 個月的低點，美股早盤開盤後一度大跌逾 1300 點，不過隨後買盤進駐收斂跌幅，台指期夜盤也同步向上，強拉近 300 點。

美股開盤重挫逾千點，盤中一度跌破 20000 點大關，不過隨即展開強勁反彈，台指期也從底部向上拉升，一度站上 8843 點，與開盤前的低點 8550 點相比，拉升 293 點，晚間 10 點左右在 8700-8800 點附近震盪，不過，仍失守 9000 點。

另外，台積電一名員工確診感染武漢肺炎，美股開盤後，ADR 旋即跳空重挫逾 9%，盤中也隨美股走揚，跌幅暫時收斂至 6% 左右。

[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, 0, -1, 0, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, -1, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0, 0, 0, -1, -1, 0, 0, 0, 0, 0, 0, -1, 0, 0, 0, 0, 0]

資料來源：鉅亨網、本研究

第伍章 結論與建議

一、 研究結論

根據研究目的以下有三個結論：

1. 比較 BiLSTM 和 Transformer 模型預測台積電台灣股價走勢

本研究將輸入資料分成兩種形式，分別台積電股價資料與台積電股價資料加上相關指數，最後產生四個預測結果。從成效比較表 3-2 結果顯示，表現最佳為 Transformer + 相關指數，其次為 BiLSTM + 相關指數和 Transformer 兩者不相上下，最差為 BiLSTM，由此得知，加上相關指數對於預測股價走勢是有相當幫助的。

2. 建立 BERT 新聞分類模型，分析影響台積電股價走勢的新聞

本研究利用兩種方式對新聞進行情緒標籤，分別是股價漲跌與情緒詞典，根據表 3-3 研究結果發現，利用股價漲跌標籤的新聞資料集，成效上精確度、準確度、召回率皆高於利用情緒詞典標籤的新聞資料集，由此得知，使用股價漲跌對新聞進行情緒標籤是較可行的方式。

3. 結合兩邊的結果與真實的股價走勢進行比較

本研究從 2015 年 6 月 1 日至 2020 年 5 月 31 日間，隨機找五天（2018 年 2 月 6 日、2018 年 7 月 20 日、2018 年 10 月 31 日、2020 年 1 月 30 日、2020 年 3 月 19 日）來預測股價走勢，並查看前一天的新聞標籤是否有與漲跌有關。

根據研究結果發現，資料集以台積電台灣股價為主，因此有些日期會因為台股休市而缺少股價，造成時間序列不連續有缺失，因此無法利用過去歷

史資料準確預測股價走勢，像是 2020 年 1 月 30 日這一天，因為遇到過年期間，所以前一天有開市的日期為 2020 年 1 月 20 日，造成股價和新聞的預測皆為錯誤。

二、 研究建議與限制

1. 根據結論一的結果（表 3-2），注意到測試集資料的結果與訓練集、驗證集相比，MSE、MAE、MAPE 皆為偏高，才發現測試集的期間為 2019 年 9 月 1 日至 2020 年 5 月 31 日，在這期間剛好遇到新冠狀病毒的突發狀況，因此現實中股市的波動很大，無法藉由過去歷史資料來預測。

2. 本研究預測股價走勢的結果與真實走勢的方向誤差不大，但是走勢斜率卻誤差較大，且因為台灣股市的漲跌幅限制是 10%，有些預測的走勢卻已經超出 10%的漲跌幅度，如圖 22 為 2018 年 10 月 31 日的預測結果，雖然走勢方向是正確的，但是上漲幅度為 16%，因此建議未來預測時，可以加入漲跌幅度作為輸入資料之一，使預測準確度提升。

3. 本研究利用股價漲跌與情緒詞典兩種方式對新聞進行情緒標籤，兩者皆有缺失的地方。使用股價漲跌的缺點是同一天的會有許多新聞，不是每一篇都為正面或負面，因此使用股價漲跌只能代表一天的所有新聞為正面或負面居多。另外使用情緒詞典時，每則新聞可能會包含許多家公司的資訊，因次建議可以只針對說明台積電資訊的段落來進行情緒標註，以提升每則新聞的情緒標籤的準確度。

4. 本研究在使用有加入相關指數的資料集，預測成效較佳，因此建議未來在搜集台積電新聞時，除了關鍵字為台積電，也可以加上台積電相關指數（台灣加權股票指數、道瓊工業平均指數和費城半導體指數）為關鍵字，來蒐集更多有關於台積電的新聞。

參考文獻

一、英文文獻

1. Bahdanau, D., Cho, K., & Bengio, Y.（2014）. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
2. Boris B. （2019）. Using the latest advancements in deep learning to predict stock price movements. Retrieved from https://towardsdatascience.com/aifortrading-2edd6fac689d
3. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. （2018）. Bert: Pre-training of deep bidirectional transformers for language understanding.
4. Devlin, J., et al. （2018）. "Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. "
5. Di Persio, L., & Honchar, O. （2016）. Artificial neural networks architectures for stock price prediction: Comparisons and applications. International journal of circuits, systems and signal processing, 10, 403-413.
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., & Bengio, Y.（2014）. Generative adversarial nets. Advances in Neural Information Processing Systems.
7. Gunduz, H., Yaslan, Y., & Cataltepe, Z. （2017）. Intraday prediction of Borsa Istanbul using convolutional neural networks and feature correlations. Knowledge-Based Systems, 137, 138-148.
8. Hessel, M., Modayil, J., Van Hasselt, H., Schaul, T., Ostrovski, G., Dabney, W., & Silver, D. （2018）. Rainbow: Combining improvements in deep reinforcement learning. In Thirty-Second AAAI Conference on Artificial Intelligence.
9. Hochreiter, S., & Schmidhuber, J. （1997）. Long short-term memory. Neural

computation, 9（8）, 1735-1780.

10. Hurst, J. M. （1970）. The Profit Magic of Stock Transaction Timing. Prentice-Hall.
11. Isaac Godfried. （2019）. Attention for time series forecasting and classification. Retrieved from https://towardsdatascience.com/attention-for-time-series-classification-and-forecasting-261723e0006d
12. Jason Brownlee .（2017）. Attention in Long Short-Term Memory Recurrent Neural Networks. Retrieved from Long Short-Term Memory Networks. https://machinelearningmastery.com/attention-long-short-term-memory-recurrent-neural-networks/
13. Kazemi, S. M., Goel, R., Eghbali, S., Ramanan, J., Sahota, J., Thakur, S., & Brubaker, M. （2019）. Time2vec: Learning a vector representation of time. arXiv preprint arXiv:1907.05321.
14. Kim, S., & Kang, M. （2019）. Financial series prediction using Attention LSTM. arXiv preprint arXiv:1902.10877.
15. Koshiyama, A., Firoozye, N., & Treleaven, P. （2019）. Generative Adversarial Networks for Financial Trading Strategies Fine-Tuning and Combination.
16. Koutnik, J., Greff, K., Gomez, F., & Schmidhuber, J. （2014）. A clockwork rnn.
17. Li, Y., Zhu, Z., Kong, D., Han, H., & Zhao, Y. （2019）. EA-LSTM: Evolutionary attention-based LSTM for time series prediction. Knowledge-Based Systems.
18. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J.（2013）. Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems （pp. 3111-3119）.
19. Olah, C. （2015）. Understanding lstm networks.
20. Pennington, J., Socher, R., & Manning, C. （2014, October）. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing （EMNLP） （pp. 1532-1543）.

21. Siami-Namini, S., Tavakoli, N., & Namin, A. S. （2019, December）. The performance of LSTM and BiLSTM in forecasting time series. In 2019 IEEE International Conference on Big Data （Big Data） （pp. 3285-3292）. IEEE.
22. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I.（2017）. Attention is all you need. In Advances in neural information processing systems （pp. 5998-6008）.
23. Vlad, G. A., Tanase, M. A., Onose, C., & Cercel, D. C. （2019, November）. Sentence-Level Propaganda Detection in News Articles with Transfer Learning and BERT-BiLSTM-Capsule Model. In Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda （pp. 148-154）.
24. Will Koehrsen.（2018）. Automated Machine Learning Hyperparameter Tuning in Python. Retrieved from https://towardsdatascience.com/automated-machine-learning-hyperparameter-tuning-in-python-dfda59b72f8a

二、中文文獻

1. 陳韋帆,& 古倫維. （2018）. 中文情感語意分析套件 CSentiPackage 發展與應用. 圖書館學與資訊科學.
2. 謝仁堡.（2018）. 語意分析和長短期記憶用於新聞預測股市未來漲跌走勢. 國立臺中科技大學資訊管理系碩士班碩士論文。取自 https://hdl.handle.net/11296/r8v564

三、網路資源

1. Jey Zhang. （2017）. 理解 LSTM/RNN 中的 Attention 機制. 取自 http://www.jeyzhang.com/understand-attention-in-rnn.html
2. Zhang Yi. （2018）. Auto Machine Learning 筆記-Bayesian Optimization. 取自 http://codewithzhangyi.com/summary/
3. 李宏毅. （2019）. 淺談神經機器翻譯&用 Transformer 與 TensorFlow 2 英翻中. 取自 https://leemeng.tw/neural-machine-translation-with-transformer-and-tensorflow2.html
4. 彭兆卿. （2017）. [強化學習]區分 Model-free 和 Model-based 方法. CSDN 官方博客。取自 https://blog.csdn.net/ppp8300885/article/details/78524235

著作權聲明

論文題目：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

論文頁數：44 頁

系所組別：資訊管理研究所

研究生：夏鶴芸

指導教授：方鄒昭聰 博士

畢業年月：109 年 8 月