國立台北商業大學
資訊與決策科技研究所
碩士學位論文

基於 BERT 模型的社群媒體情緒分析對台灣股市長期趨勢的影響探討

Exploring the Impact of BERT-based Sentiment Analysis of Social Media on Long-term Trends of Taiwan Stock

研究生：王柏淞

指導教授：張瑞雄 教授 王亦凡 教授

中華民國一一二年六月

摘 要

情緒分析是一種能夠自動化分析文本裡蘊含的情緒之技術。而 BERT 模型是目前自然語言處理領域前幾熱門的模型之一，以深度雙向編碼器為基礎，能夠進行高效而準確的自然語言理解。本研究主要目的在探討 BERT 模型的社群媒體情緒分析對台灣股市長期趨勢的影響。

研究內容為從社群媒體上擷取每日大量的文章，利用 BERT 模型進行情緒分析，再將分析結果整理成一整年的社群媒體情緒趨勢。最後，利用波段法之分析，將其與台灣股市一整年的趨勢進行圖表比對期間之關聯度，以探討社群媒體情緒分析是否能夠對股市趨勢有所預測和影響，希望此研究能夠為了解社群媒體對股市的影響提供新的角度和思路。

關鍵詞：社群媒體、情緒分析、BERT、股市

Abstract

Sentiment analysis is a technique capable of automatically analyzing the emotions contained in text. BERT is one of the leading models in the natural language processing field, based on a deep bidirectional encoder, which enables efficient and accurate natural language understanding. The main purpose of this study is to explore the impact of sentiment analysis using the BERT model on long-term trends in the Taiwan stock market.

The research involves crawling a large number of articles daily from social media, performing sentiment analysis using the BERT model, and then compiling the analysis results into a year-long social media sentiment trend. Finally, by comparing the charts and analyzing the correlations, such as the band method, with the annual trend of the Taiwan stock market, this study aims to investigate whether sentiment analysis from social media can predict and influence stock market trends. It is hoped that this research will provide new insights and ideas for understanding the impact of social media on the stock market.

Keywords: Social Media, Sentiment Analysis, BERT, Stock

誌謝

時光飛逝，兩年的研究生生涯也即將抵達盡頭，「痛苦會過去，美麗會留下下。」可以說是這一路走來最真實的寫照。謝謝廖文華教授兼所長帶我進入資訊這塊美好的領域，盡管因為個人身體問題而害怕連累教授，但還是感謝教授一路的照顧。謝謝我的指導教授—張瑞雄校長對我的悉心教導，教授因為轉戰更大的領域而非常忙碌，但還是會抽出時間跟我開會及回我訊息。一開始的選題其實挺迷茫，但經過跟教授的頻繁討論後順利確定了方向，實在是萬分的感謝。謝謝王亦凡教授提供我很多開創性的建議，並且在人生的道路上給予許多指導，用極遠的目光帶我瞭解許多事物。當然還要感謝一起奮鬥的劭齊、其庭、辰秧、允謙…等，艱辛的路途中有同好結伴而行是非常幸運的一件事。

最後，感恩資訊與決策科技研究所的所有同學、教授、助教，以及女朋友和我的爸爸、媽媽、哥哥，在這些日子中提供了很多的慰藉與勉勵，讓我能平安畢業。

國立臺北商業大學 無違反學術倫理聲明書
National Taipei University of Business
Statement of Academic Ethics

立聲明書人 王柏淞 (學號：11066005) 於資訊與決策科技研究研究所，撰寫 基於 BERT 模型的社群媒體情緒分析對台灣股市長期趨勢的影響探討/ Exploring the Impact of BERT-based Sentiment Analysis of Social Media on Long-term Trends of Taiwan Stock 期間，業經指導教授指示：絕不可剽竊、抄襲及剪貼他人之論述，並已獲知剽竊、抄襲及剪貼論文之定義。因此，凡引述他人之觀點及圖表，本人皆在論文詳實註明出處，絕未涉及抄襲、剽竊及剪貼等違反學術倫理之情事。如有違反，本人除願意負起法律責任，並無條件同意由教育部及國立臺北商業大學註銷本人之碩士學位，絕無異議。

I declare that the texts, citations, charts, figures, illustrations, and pictures contained in the thesis titled Exploring the Impact of BERT-based Sentiment Analysis of Social Media on Long-term Trends of Taiwan Stock have no plagiarism of others' works nor violations of academic ethics. I accept legal liability for plagiarism and violations of academic ethics and completely agree the revocation of the Master's Degree granted to me by The Ministry of Education (Taiwan) and National Taipei University of Business if any of the said misconduct is confirmed true.

聲 明 人(Student)：王柏淞 (親筆簽名 Signature)

中 華 民 國 112 年(Y) 7 月(M) 8 日(D)

*本聲明書之影本請裝訂於紙本內頁、掃瞄後電子檔加入電子論文內頁（目錄之前頁）。

目 錄

表目錄

圖目錄

第一章 緒論

隨著互聯網和移動設備的普及，社群媒體已成為人們獲取信息、分享觀點和交流社交的重要平台。社群媒體上的訊息往往反映了人們對時事和話題的關注程度，以及對這些話題的情感態度。在此背景下，社群媒體情感分析逐漸成為研究的熱點之一。社群媒體情感分析可以對文本內容進行情感分類，從而提供有關人們對某個話題或事件的情感傾向，這對於了解大眾對於股市的看法和情感是非常重要的。

股市是一個受多種因素影響的複雜系統。在過去，研究者們主要是從股市基本面、技術面和市場心理等方面來進行股市預測。然而，隨著社群媒體數據的大量湧現，越來越多的研究者開始關注社群媒體情感對股市趨勢的影響。因此，探討社群媒體情感對股市趨勢的影響，能夠為股市投資者提供重要的參考資訊，同時對於研究社群媒體的影響力也具有一定的學術價值。

BERT 模型是目前自然語言處理領域最熱門的模型之一，其擁有深度雙向編碼器的特性，能夠進行高效而準確的自然語言理解，對於社群媒體情感分析具有重要的應用價值。本研究旨在利用 BERT 模型進行社群媒體情感分析，探討社群媒體情感分析對股市趨勢的影響。透過這一研究，能夠進一步拓展社群媒體情感分析的應用範圍，同時也能夠為股市投資者提供新的投資策略和思路。

第二章 文獻探討

2.1 社群媒體

社群媒體作為一種新興的傳播媒介，在現代社會中發揮著越來越重要的作用。它們包括微博、Twitter、Facebook 等，這些平台不僅讓用戶能夠快速地分享信息、交流意見，也為企業、政府、學術界等提供了更多的溝通和互動機會。社群媒體上的信息量巨大，涉及的主題也非常廣泛，從個人生活到社會事件，從商業活動到政治選舉，都有廣泛的應用。

社群媒體上的信息以文本為主，因此自然語言處理技術對於社群媒體的分析和應用具有重要的意義。通過對社群媒體文本的處理和分析，可以獲得許多有價值的信息，如情感分析、主題分析、用戶行為分析等，進而應用於各個領域，如商業、政治、媒體等。

2.2 情緒分析

情緒分析是自然語言處理中的一個重要應用領域，它的技術實現需要運用自然語言處理技術和機器學習技術，通常包括文本預處理、特徵提取、情感詞典構建、情感分類等步驟。情感詞典是情緒分析的核心，它是由一系列詞語和詞語的情感極性標記構成的，情感極性標記可以是正面、負面或中性。通過匹配文本中的詞語和情感詞典中的詞語，可以對文本進行情感分類。近年來，深度學習技術的發展，特別是基於深度神經網絡的情感分類方法，已經成為情緒分析中的熱門研究方向。

除了傳統的機器學習方法，近年來，深度學習技術的發展也為情緒分析帶來了新的思路和方法。深度學習技術通常基於深度神經網絡，可以有效處理自然語言處理中的問題，包括文本分類、情感分析等。其中，基於深度神經網絡的情感

分析方法已經成為情緒分析中的熱門研究方向。例如有研究提出了一種基於卷積神經網絡的情感分類方法(Kim, 2014)，該方法將情感分類視為一個文本分類問題，透過學習詞語的特徵來進行情感分類。有些研究則是設計了一種基於目標注意力機制的細粒度情感分析(Huang & Cheng, 2018)，該方法利用目標注意力機制來讓模型專注在目標中更相關的部分，這會提升情感分析的準確率。

情緒分析技術在各個領域都有著廣泛的應用價值，它可以幫助人們更好地了解世界、理解人們的情感和行為，對於商業決策、社會分析等方面都有重要的作用。隨著深度學習技術的發展，情緒分析技術也在不斷地進步和發展，未來將會有更多的應用場景和發展方向。

2.3 BERT 模型

BERT（Bidirectional Encoder Representations from Transformers）是 google 於 2018 年提出的一種基於 Transformer 模型的語言模型。相較於傳統的語言模型只能單向預測下一個單詞，BERT 採用雙向 Transformer 編碼器，能夠同時考慮上下文的內容，從而更好地理解句子的含義。

BERT 模型的訓練過程採用了大量的未標註文本，通過 Masked Language Model（MLM）和 Next Sentence Prediction（NSP）等兩種任務進行訓練。MLM 任務要求模型在輸入文本中隨機遮蔽一些單詞，然後預測被遮蔽的單詞。NSP 任務要求模型判斷兩個句子是否連貫。

BERT 模型在多項自然語言處理任務上取得了優秀的表現，包括問答系統、情感分析、語言翻譯等。例如 2018 年 Bert 模型在 SQuAD1.1 和 GLUE 等公開測試集上取得了卓越的表現(Devlin et al., 2018)。隔年，BERT 於情感分析任務上的應用也得到了比傳統情感分析方法更好的結果(Yang et al., 2019)。

在近年的研究中，學者們不斷優化 BERT 模型的性能和效率，例如經過針對性調整後的 RoBERTa 模型，透過進一步優化 BERT 的預訓練方法和訓練數據，提高了模型的性能和泛化能力(Liu et al., 2019)。DistilBERT 模型則是針對 BERT 模型

在部署上的問題進行優化，將模型壓縮為原模型的 40%，提高了模型的運行速度和效率(Sanh et al., 2020)。ALBERT 模型則是探索了 BERT 模型在多任務學習上的應用，並且通過對 BERT 模型進行精簡優化，提高了模型的性能和泛化能力(Lan et al., 2020)。

2.4 情緒分析與股票相關的應用

在過去幾年，研究者們將社交媒體和情感連結，並在股票的價格預測領域進行了許多有趣的研究。如通過情感分析探討金融新聞對股價的影響(Li et al., 2014)，實驗結果顯示在個股、行業板塊和指數層面上，採用情感分析的模型在驗證集和獨立測試集上均優於詞袋模型。或是嘗試建立一個利用社交媒體情感的股票價格走勢預測模型(Nguyen, Shirai, & Velcin, 2015)，該研究將特定企業主題的情感納入股票預測，而不僅僅是關注整體情緒。其研究結果顯示，在一年內交易的 18 支股票準確率高於平均值，比僅使用歷史價格的模型提高 2.07%。

2016 年，有研究將 Twitter 數據的情感分析應用在股市的預測(Pagolu, Reddy, Panda, & Majhi, 2016)。研究結果顯示股票價格的上升和下跌與推文中的公眾情感之間存在很強的關聯。2020 年，提出了一種基於深度學習的股票市場預測模型，考慮了投資者的情感傾向並結合經驗模態分解（EMD）和修訂後的長短時記憶（LSTM）技術(Jin, Yang, & Liu, 2020)。實驗結果證明，這一方法有效提高了預測準確性，並且減少了時間延遲。

時間來到近兩年，一種基於多種數據源和投資者情緒的股票價格預測方法被提出(Wu, Liu, Zhou, & Weng, 2022)，其被命名為 S_I_LSTM，實驗結果顯示，預測的股票收盤價比單一數據源更接近真實收盤價，平均絕對誤差可達 2.386835，優於傳統方法。

同年，還有一個名為 HiSA-SMFM 的股市預測模型出現(Gupta, Madan, Singh, & Singh, 2022)，其透過歷史數據和情感數據通過應用 LSTM 來有效預測股票價格。整合這些因素後，可以更準確地預測股票價格。

有研究則是基於加權文本內容和金融異常(Qiu, Song, & Chen, 2022)，開發了一個新的情感指數來預測股票趨勢。實驗結果顯示，修改過的情感指數可以有效地提高對股票趨勢預測的預測能力。

同樣使用 LSTM 的研究中，還有像是基於教學和學習優化（TLBO）模型和長短期記憶情感分析的股票價格預測方法(Swathi, Kasiviswanath, & Rao, 2022)，使用推特數據。實驗結果顯示，TLBO-LSTM 模型在預測股票價格方面具有出色表現，準確率高達 94.73%。

第三章 研究方法

在本研究的流程中，使用的語言主要為 Python，開發環境則是以 Google 所提供的雲端開發平台 Colab 和專門為 Python 所設計的 PyCharm 兩個平台為主。研究旨在使用前文回顧中情緒分析效果最佳的 BERT 模型對社交媒體上的金融相關文章進行分析，從而獲取每日的平均情緒，再將平均分數透過移動平均法整理成情緒趨勢，研究架構如圖 3.1 所示。

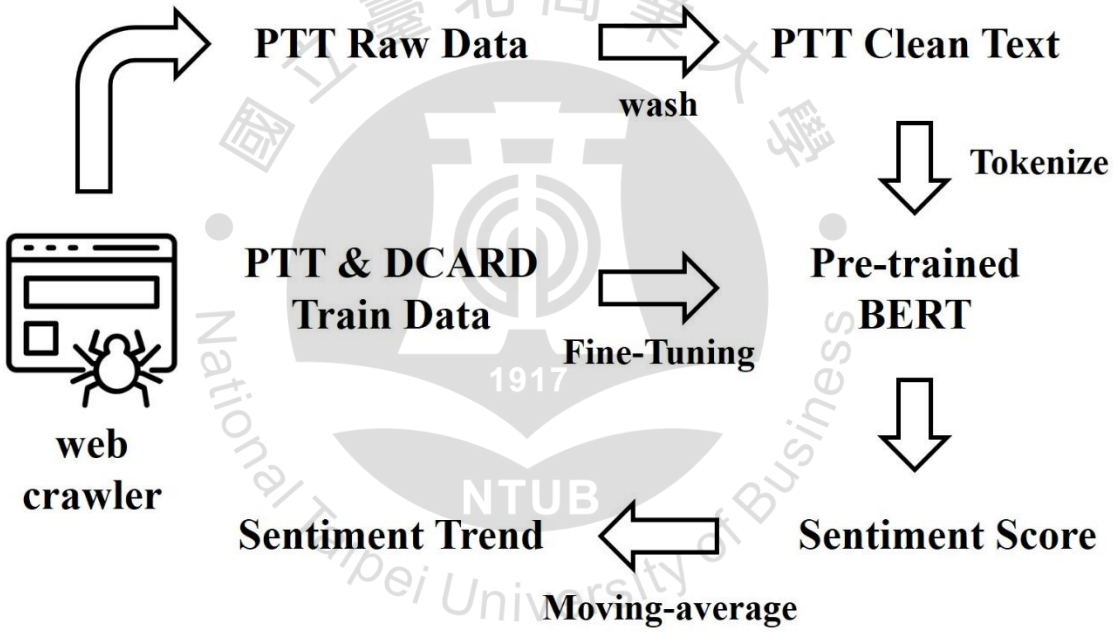


圖 3.1 研究架構

3.1 資料集介紹

本研究的目標是分析 2021 年到 2022 年之間的社群媒體情緒和台灣股市的關聯，而目前比較廣為人知的大型社群媒體包含論壇有：Dcard、Ptt、Facebook、巴哈姆特…等，但經過統整後發現，做為目標的社群媒體需具有：

1. 能夠完整檢閱過去資料(即每一篇文章)。

2. 與股市或金融相關的討論區。

3. 較為踴躍的發文人數。

由於研究本身時間的關係，目標文章的年份已是距今至少一年以上的資料，而原本作為選定目標的 Dcard 在往前檢索超過三個月後就無法獲得文章的詳細清單，只能透過搜索關鍵字及標題來查找文章，且平台本身也不支持日期式搜索，所以在設計完相對應的整套爬蟲程式卻發現無法往前遍歷超出期限之文章後，只能忍痛排除。而作為備案的 Ptt 不僅有股票版，且股票版上的文章每天至少有 10 篇以上，且能夠從當前月份一路往回查閱至創版時期，所以本研究主要以 Ptt 股票版的文章作為分析對象，輔以少數的 Dcard 文章作為訓練模型的輸入。

Ptt 股票版總共有「標的」、「新聞」、「心得」、「請益」、「投顧」、「問卷」、「其他」、「公告」、「閒聊」等九種類別。本研究選取其中的「心得」、「請益」和「閒聊」等三種較有個人主觀情緒的文章作為情緒分析的目標，並透過 Python 套件 selenium 配合 request 以及 PyPtt 進行抓取，日期限定為 2021 年 2 月 17 日（台股 2021 年開盤日）到 2022 年 1 月 27 日（台股 2021 年封關日）內之所有種類符合的股票版文章，最後共爬取了 13096 筆資料。

原始資料包含下列欄位：index、title、content、date、push_number。

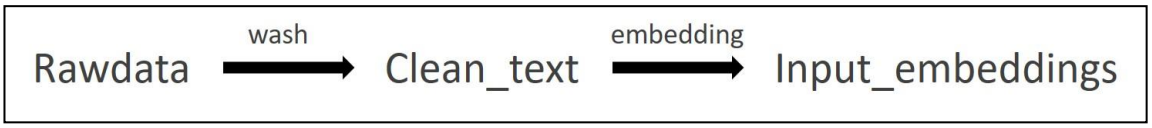
	index	title	content	date	push_number
0	53136.0	[請益] 永豐債委託	\n\n想請問有沒有人用過委託下單\n\n剛才看可下單餘額突然少了10%資產\n\n很少遇到...	Wed Feb 17 01:20:45 2021	21
1	53139.0	[請益] 全職交易，然後抱長線??	\n\n 小弟有個問題，全職交易如果是短線操作我是可以理解\n\n可能每個星期或每個...	Wed Feb 17 03:16:02 2021	41
2	53140.0	Re: [心得] 還來的判決終究會降臨於主動投資者嗎？	\n※ 引述《x77 (肛之戀精術士)》之銘言：\n: "親愛的上帝，請賜給我雅量從容的接受...	Wed Feb 17 05:17:36 2021	65
3	53141.0	[請益] "本多終勝"的"多",最低的額度是要多少錢?	\n問題：\n在股版問問題，最常見的答案不外乎就是歐印台GG、本多終勝、無腦多ETF那幾個答...	Wed Feb 17 05:22:11 2021	10
4	53143.0	Re: [請益] "本多終勝"的"多",最低的額度是要多少錢?	\n我這一年操作的心得\n\n多的定義在於你買的標的物價格\n\n而且現在有零股\n\n即使...	Wed Feb 17 06:23:05 2021	NaN
5	53148.0	Re: [請益] "本多終勝"的"多",最低的額度是要多少錢?	\n本多忠勝不是這意思\n\n是你可以隨便凹\n\n股價跌\n\n你就用力買拉起來再賣\n\...	Wed Feb 17 08:15:40 2021	NaN
6	53150.0	[閒聊] 2021/02/17 盤中閒聊	\n==========110/02/17台股資訊重點整理，供股民做投資參考====...	Wed Feb 17 08:30:01 2021	99
7	53153.0	[請益] 買大型基金 理論上是不是無敵	\n大家好像都有個共識\n\n股票是大型法人在玩的 散戶只是韭菜\n\n韭菜被割是活該\n\...	Wed Feb 17 09:11:18 2021	21
8	53158.0	Re: [請益] "本多終勝"的"多",最低的額度是要多少錢?	\n※ 引述《pipiboygay (喜歡男人的男生)》之銘言：\n: 問題：\n: 在股版...	Wed Feb 17 10:00:03 2021	11
9	53159.0	Re: [請益] "本多終勝"的"多",最低的額度是要多少錢?	\n※ 引述《pipiboygay (喜歡男人的男生)》之銘言：\n: 問題：\n: 在股版...	Wed Feb 17 10:08:23 2021	9

圖 3.2 原始資料

3.2 資料前處理

本研究對資料會進行兩次前處理，第一次為清洗，目的是將爬下來的 Ptt 股

票版文章去除多餘部分後留下純文字。第二次為轉換，目的是透過 hugging face



提供的套件將純文字轉換成 Bert 模型能閱讀的型態。

圖 3.3 清洗過程

3.2.1 清洗資料

由於 Ptt 的文章機制問題，常會有使用者藉由引述別人文章來做開頭或結尾，而 Ptt 的文章每一句都是以空行分開，引述部分則是用字元冒號或「※」作為開頭，除了引述外還會有時不時插入的版規注意事項以及空行，因此本研究設計了一個微型 function 來進行清洗的動作，如圖 3.4 所示。

```
def preprocessPttText(text):
    prohibited_characters = (
            ":", "--", "※", "http", "ーー", "Sent from", "發文前請先詳閱", "根據板規1-4-2",
            " 1.問漲跌，或持股分析", "1.問漲跌，或持股分析", "請使用 [標的] 分類", "請使用[標的]分類")
    # 檢測起始值無法檢測到空字串 所以依據 冒號開頭、網址、分隔線、空字串 進行排除
    clean_line = [line.strip() for line in text.split("\n") if not line.startswith(prohibited_characters) and line != ""]
    output_string = "\n".join(clean_line)
    return output_string
```

圖 3.4 清洗文章的 function（僅針對股票版文章）

圖 3.5 與圖 3.6 為文章清洗前後的實際對比。

```
Re: [請益] 融券虧損,攤平跟認賠的抉擇

: 方案1:就再去融券2張去攤平，(50+180)/3=77元左右
: 大概等於每張的成本是77元，那如果有跌下來的話，比如說跌到80元的話
: 那我就等於(80-77)*3=9元 9*1000股=9000(元)
: 相較於不攤平直接認賠,80-56=24 24*1000=24000(元)
: 我可以少賠24000-9000=15000元，但這缺點是花費的成本較高，要用18萬來補洞。

: 方案2:
: 不攤平，當學經驗認賠回補。將剛剛方案1要拿來攤平的錢18萬元拿來做別檔股票
: 將18萬攤平的錢加上融券認賠後拿回3萬元，這20萬元當作股本重新開始
: 拿這筆股本趁現在萬六點的時機，好好選擇股票，從此不再融券、融資，只做現股
: 學會有多少錢的成本就做多少錢的股票，有賺再將之前的虧損慢慢補回
: 請問經驗豐富的版友，如果今天是大家遇到這情況，大家會怎麼做呢??
: 還是大家有其他更好補救的方法呢???真心感謝大家的耐心，祝大家新牛年都賺大錢

那如果把攤平的錢拿去買現股(或是融資)做多 跟融券對做 觀察一段時間

然後把虧損的那方停損出場 獲利的部分看情況續抱

請問大家覺得

這樣的做法有甚麼盲點或是不洽當的地方嗎??

還是有其他更好的做法可以提供參考的嗎??

--
發文前請先詳閱[請益]分類發文規範，未依規範發文將受處份。
根據板規1-4-2.請益可提及標的，但是只要屬以下如：
 1.問漲跌，或持股分析。  2.預測個股或產業未來性。  3.或有推薦意圖。
請使用 [標的] 分類並依正確格式發文，違者1-4-2砍文處分。
------------------------ 以上宣導事項請勿刪除 違者4-1 ----------------------

--
```

圖 3.5 清洗前的文本

```
那如果把攤平的錢拿去買現股(或是融資)做多 跟融券對做 觀察一段時間
然後把虧損的那方停損出場 獲利的部分看情況續抱
請問大家覺得
這樣的做法有甚麼盲點或是不洽當的地方嗎??
還是有其他更好的做法可以提供參考的嗎??
```

圖 3.6 清洗後的文本

3.2.2 轉換資料

獲得純文字並且正確空行的文本後，無論是要訓練 BERT 模型，還是輸出情緒判斷的值，都是無法直接將之丟入模型的，需要將其轉換成 BERT 模型所能讀

懂的形式。而丟入的值又會依據模型有無訓練過而分成兩大類。

3.2.2.1 未訓練過的 BERT 模型輸入

首先需要將這段文本進行分詞，並且選擇與所使用的模型相符的分詞器。由於 BERT 模型使用的是 WordPiece 詞彙袋，因此分詞的結果可能是一個單一字元，而非完整詞語。接著，在將分詞後的文本轉換為 input_ids 之前，需要在文本的開頭加入[CLS]，在文本的結尾加入[SEP]，這樣模型才能知道輸入文本的開始和結束位置。最後還需要映射成對應的字向量，如以下文本：

"今天的天氣真好，來去散散步吧！"

經過分詞並加入判斷開頭結尾的標籤，而變成：

['[CLS]', '今', '天', '的', '天', '氣', '真', '好', '，', '來', '去', '散', '散', '步', '吧', '!', '[SEP]']

再經過模型特有的向量映射變成 input_ids，開頭結尾則用 101 及 102 代表：

tensor([101, 791, 1921, 4638, 1921, 3706, 4696, 1962, 1557, 106, 2523, 6900, 1394, 1343, 3141, 3141, 3635, 1469, 6624, 6662, 511, 102])

第二步，我們需要設計 attention mask 告訴模型哪些位置是 padding 的。在 BERT 中，padding 指的是為了讓所有輸入的句子長度相同，所添加的特殊標記，在文章判讀中很常會將長度訂為 512 個字元。attention mask 通常是一個由 0 和 1 組成的矩陣，其中 1 表示這個位置是輸入文本，0 表示這個位置是 padding。若是以上述句子為例，則會轉換為：

tensor([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1])

最後，在 BERT 模型中 token_type_ids 主要用來區分兩個不同的句子。BERT 模型是基於 Transformer 的神經網絡模型，將輸入的文本序列轉換為數字序列（即 tokens）。然而，對於一些自然語言處理的任務，需要處理的文本不止一個句子，而是包含多個句子或段落，如問答、文本蘊涵等任務。在這些任務中，BERT 模型需要知道每個 token 屬於哪個句子或段落。為了解決這個問題，BERT 模型引入

了 token_type_ids 標識不同句子的 tokens，通過將不同句子的 tokens 賦予不同的 token_type_ids，模型就能夠分別處理每個句子或段落的信息，以便更好地進行任務處理。以上述例子作為轉換的話，可得：

tensor([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0])

總體來說，未訓練過的 BERT 模型的輸入需要三個要素：input_ids、attention mask 和 token_type_ids。而這些要素的設計必須與所使用的 BERT 模型相符，才能使模型正確地進行文本分析和任務執行。

3.2.2.2 預訓練過的 BERT 模型

當使用預訓練的 BERT 模型時，只需要提供輸入的文本，也就是 input_ids，就可以進行模型的預測，不需要提供 token_type_ids 和 attention mask。input_ids 是一個整數序列，其中每個整數代表輸入文本中的一個詞彙在 BERT 的詞彙表中的索引。例如，在一個由 14 個詞彙組成的句子中，input_ids 可能如下所示：[101, 784, 4638, 4636, 4495, 2523, 4696, 1962, 8024, 3341, 3341, 3918, 3918, 7213, 138, 102]。在這個序列中，101 代表[CLS]符號，102 代表[SEP]符號，而其他的整數則代表輸入句子中每個詞彙的索引。當 BERT 模型收到 input_ids 時，它會將每個詞彙轉換為詞向量，並通過多層神經網路進行計算，從而生成一個代表整個句子含義的向量。這個向量可以用來進行下游任務，例如文本分類或命名實體識別等。由於 BERT 模型是基於 Transformer 的，因此它不需要額外的 attention mask 來指示模型將注意力集中在哪些詞彙上。Transformer 中的注意力機制可以自動學習要將注意力集中在哪些詞彙上，因此不需要額外的 attention mask。同時，由於預訓練模型是基於兩個任務（Masked Language Model 和 Next Sentence Prediction）進行訓練的，因此模型在預訓練過程中已經學會了如何將輸入的文本分成不同的句子，因此也不需要提供 token_type_ids。

以 Ptt 文本為例進行轉換可得：

```
有空去看看YouTube 上FSD的影片
這邊隨便講幾個優勢
1.電動車領域領先Audi應該有3～5年
2.自動駕駛沒有對手
3.汽車生產工藝顛覆傳統，真的要屌打傳統車廠
好了!
股價應該可以再翻個8倍!

input_ids:
 tensor([  101,  3300,  4958,  1343,  4692,  4692,  8487,   677,   148, 10117,
         4638,  2512,  4275,  6857,  6920,  7401,   912,  6341,  2407,   943,
         1032,  1248,   122,   119,  7442,  1240,  6722,  7526,  1818,  7526,
         1044,  9976,  2746,  6283,  3300, 12222,  2399,   123,   119,  5632,
         1240,  7690,  7691,  3760,  3300,  2205,  2797,   124,   119,  3749,
         6722,  4495,  4496,  2339,  5971,  7545,  6208,  1001,  5186,  8024,
         4696,  4638,  6206,  2239,  2802,  1001,  5186,  6722,  2449,  1962,
          749,  8013,  5500,  1019,  2746,  6283,  1377,   809,  1086,  5436,
          943,   129,   945,  8013,   102])
```

圖 3.7 轉換成 input_ids 的示範文本

3.3 預訓練的 BERT 模型

完成了資料的前處理與轉換後，接著就需要一個模型產出所謂的情緒分數，但正常的流程需要大量資料跟 GPU 來進行訓練，對此 huggingface 提供了許多已經預訓練過的模型，包含針對各式各樣不同任務的。

本研究使用的是基於 BERT 的預訓練情緒分析模型，將正負情緒分成 5 個等級，並且 5 分是最為正向，1 分是最為負面。透過爬取額外 2000 篇 Ptt 文本及 300 篇 Dcard 文本並進行人工的 1-5 分情緒標籤註記再進行模型評估。可以得到以五個等級的情感精度來說，較為良好的結果。

表 3.1 模型評估

指標	1分	2分	3分	4分	5分	平均
正確率 (Accuracy)	72.5%	76.2%	81.5%	85.1%	78.4%	78.7%
精確率 (Precision)	61.3%	57.8%	74.1%	85.6%	81.2%	72.0%
召回率 (Recall)	55.7%	54.2%	71.3%	87.1%	78.8%	69.4%
F1值	58.3%	55.5%	72.5%	86.2%	80.0%	70.5%

3.4 整理情緒趨勢

準備好輸入資料及模型後，可以開始產出情緒分數。由前述可知，分數分為5個等級，一分最負面。接著設計一個迴圈將所有的文章依序丟入模型產出分數，程式碼如圖 3.8 所示。

```
for data in Clean_Ptt.iterrows():
    try:
        train_text = cc.convert(data[1]["content"]) # 取出每筆文章的內容然後轉成簡體
        if len(train_text) > 512:
            train_text = train_text[:512]
        batch = tokenizer(train_text, return_tensors="pt")
        batch = torch.tensor(batch['input_ids'])
        with torch.no_grad():
            outputs = model(batch)
            predictions = F.softmax(outputs.logits, dim=1)
            labels = int(torch.argmax(predictions, dim=1))+1
        Clean_Ptt.loc[data[0],"sentiment"] = labels
        print(data[0])
    except:
        print(data[0])
        pass

Clean_Ptt.to_excel("Finally.xlsx", index=False)
```

圖 3.8 輸入模型產出分數

完成上述步驟後，會發現每一天發布的文章數量不一樣，自然也會有數量不等的情緒分數，若是直接整理成趨勢會非常難整理。因此本研究先將同天的分數作平均，得到每日的平均分數後，再進行移動平均法從而畫出趨勢圖，如圖 3.9 所示。

```python
def moving_average(data, window_size):
    # :param data: 一維數組，表示要計算移動平均值的數據
    # :param window_size: 整數，表示計算平均值的窗口大小
    # :return: 一維數組，表示計算出的移動平均值序列
    ma = []
    for i in range(len(data)):
        if i < window_size:
            ma.append(sum(data[0:i+1]) / (i+1)) # 如果還沒有達到窗口大小，則使用前向填充法
        else:
            window = data[i-window_size+1:i+1] # 計算窗口內的平均值
            ma.append(sum(window) / window_size)
    return ma
```

圖 3.9 移動平均法

移動平均法是一種用來分析時間序列數據的統計方法，通過平滑數據以減少隨機波動的影響，從而揭示數據的趨勢和週期性變化。移動平均法的基本思想是在時間序列上移動一個固定窗口大小的平均值，用平均值來代替原始數據，從而平滑數據。移動平均法可以針對不同的窗口大小進行計算，通常窗口大小越大，數據就越平滑，反之窗口大小越小，平滑後的數據就越接近原始數據。

移動平均法有兩種常見的方法：簡單移動平均法和加權移動平均法。簡單移動平均法是一種等權重的方法，將最近的 n 個數據點加起來並除以 n，得到平均值，再將平均值與最新的數據點一起作為新的平均值。加權移動平均法則是將最近的 n 個數據點分別乘以不同的權重，然後將加權後的數據點加起來除以權重之和，得到平均值。移動平均法在時間序列數據的分析和預測中有廣泛應用，例如股票價格預測、天氣預報、經濟指標預測等等。移動平均法的優點是簡單易懂，容易實現，可以有效地平滑數據，減少隨機波動的影響，但是移動平均法也有一些缺點，例如無法捕捉數據的短期變化，以及需要選擇合適的窗口大小和權重，否則可能會出現過度平滑或不足平滑的情況。

第四章 研究結果

將研究得出的情緒趨勢圖與台股大盤 K 線圖做比較之後，發現每個月份的準確率不盡相同。有的極為相似，有的只有部分相似，有的則是幾乎相反或沒有關聯，而這也導致很難用傳統的統計方法來檢查關聯，也因此設計了波段法來檢視關聯性。

波段法具體而言，是將每個連續五天或以上股價均呈上漲或下跌的股票走勢視為一個波段，同時，使用股票價格和交易量計算每個波段的股市趨勢。在使用相關性分析評估每個波段的情感趨勢和股市趨勢之間的關聯性。而研究結果顯示，過去一年共 11 個波段中有 8 個波段的情感趨勢和股市趨勢呈現波動一致的現象，波段比來到 0.727，即當股市趨勢向上時，社群媒體情感趨勢也連續大幅的向上或向下；當社群媒體情感趨勢向下時，股市趨勢也同理。這個結果表明，社群媒體情感趨勢和股市趨勢會互相產生影響，且在較為長期且明顯的情緒波動時，股價的漲跌也會相對突出。

其次，有的月份社群媒體情感趨勢和股市趨勢之間相關性較強，而有些月份則相關性較弱或不存在。這可能是由於許多因素的複雜交互作用造成的，例如：

1. 網民們可能受到各種因素的影響，如新聞事件、政治環境、經濟指標等。當股市大幅波動時，這些因素同時也會對社群媒體情感趨勢產生直接或間接的影響。
2. 股民對股市波動的反應可能隨時間而變化。當股市處於下跌趨勢時，人們可能會對此感到擔憂和焦慮，但隨著時間的推移，股民會開始對股市下跌產生麻痺、適應甚至是高興的情緒。
3. 股市和社群媒體之間的影響可能存在時間滯後效應。即使情感趨勢與股市趨勢之間存在關聯性，但情感趨勢的變化可能需要一定的時間才能影響股市趨勢的變化。

圖 4.1 為月份中較為相關的比較圖示例。

圖 4.1 五月份股市比較圖

而我們也可以觀察到，代表性的重大事件對情緒變化和股市之間的關聯性有顯著影響。以五月份的新冠肺炎不明感染源風暴為例，當疫情惡化並擴大至全球範圍時，市場出現了普遍的恐慌情緒。這種恐慌情緒迅速蔓延至台股市場，導致「512 股災」的發生。在此次股災中，加權指數跌至 15,165.27 點，兩個交易日內總共下跌約 2,000 點，創下驚人的 1,400 點單日跌幅。

而情緒反應到股市的延遲也是一個值得關注的因素。在大型事件發生後，人們的情緒可能並非立即對股市產生影響，而是在一段時間後逐漸顯現，亦或是股市跟事件的綜合衰敗導致股民的恐慌更加劇。在五月份新冠肺炎不明感染源風暴

的例子中，隨著疫情消息的擴散，恐慌情緒逐漸累積並在市場中形成壓力。最終，這種情緒壓力在五月份中段達到高峰，使得股市出現大幅下跌。

此現象也代表重大事件會增強情緒變化與股市之間的關聯性，而情緒反應到股市的延遲及之後兩者相加的餘波也是一個不容忽視的影響。

圖 4.2 三至六月情緒股市比較圖

從下圖 4.3 中的紅線可看出，情緒漲跌與股市並不能完全相符，且持續的幅度通常也以情緒圖較為短暫。

圖 4.3 七至十月情緒股市比較圖

下圖 4.4 中可發現，情緒的連續起伏相較於股票，易有滯後效應且幅度通常較為誇張、短暫。

圖 4.4 年末兩月情緒股市比較圖

第五章 討論

在本章中，將本研究與第 2.4 節中提到的相關文獻進行比較。比較的重點是 BERT 模型在情緒分析中的準確性和它在股市預測中的後續應用，以及可能影響情緒趨勢和股市趨勢之間相關性的因素。

首先，本研究的研究方法為透過網路爬蟲收集社群媒體的文章，並將其變成情緒數據。例如，Li 等人（2014）和 Nguyen 等人（2015）也探索了情感分析在股票價格預測中的應用，但他們使用的模型為 LSTM 且數據集也不盡相同。而本研究則是利用 BERT 模型進行情緒分析，並專注於社交媒體的情緒長期趨勢。

其次，本研究利用 BERT 模型對社交媒體數據進行情感分析，與先前提到的其他幾項研究類似。例如，Pagolu 等人（2016 年）也在 Twitter 數據上應用情緒分析進行股票市場預測。然而，他們關注的是股票價格波動與推文中公眾情緒之間的關係，而本研究則更關注特定股票版面的整體情緒趨勢及其與股票市場走勢的相關性。

與 Nguyen 等人（2015）的研究相比，本研究的方法著重於顯示出情緒趨勢和股市波動之間的相關程度而非預測。Nguyen 等人的研究與只使用歷史價格數據的模型相比，測試股票的準確率平均提高了 2.07%。而本研究中，透過波段法則是能發現在 11 個波段中的 8 個，情緒趨勢和股票市場趨勢表現出一致的波動，表明本研究的方法可以有效地捕捉到情緒和股票市場趨勢間的關聯。

本研究與以往文獻的另一個顯著區別是使用波段法來考察相關性。這種技術將股市走勢歸納為連續五天或更長時間的持續上升或下降運動，使我們能夠比傳統統計方法更有效地評估情緒趨勢和股市長期走勢之間的關系。這種方法使本研究有別於以往的研究，以往的研究大多採用傳統的相關分析方法。

本研究還強調了重大事件對情緒變化和股市走勢之間關系的影響，例如 5 月

份的 COVID-19 爆發的例子。這一觀察表明，本研究的方法對重大事件對情緒和股市走勢的影響很敏感，能為投資者和市場分析師及後續相關研究者提供寶貴的見解。

總而言之，本研究提出的方法與先前討論的相關文獻相比，顯示出不同的側重點及若干優勢。不論是使用 BERT 模型進行情緒分析，使用波段法進行相關分析，以及考慮重大事件對情緒和股市趨勢之間關係的影響，都有助於形成一種更有效的方法來分析社交媒體情緒和股市長期走勢之間的關係。雖然仍有需要改進和進一步研究的地方，但本研究為這一領域的未來工作提供了一個堅實的基礎。

第六章 結論與未來展望

本研究利用波段法檢視社群媒體情感趨勢和股市趨勢之間的關係，結果顯示兩者之間存在相關性。然而，在不同月份的關聯性卻不盡相同。其次，人們對股市波動的反應可能隨時間而變化，除此之外還有許多因素可能交互作用導致社群媒體情感趨勢和股市趨勢之間的相關性在不同月份有所不同。因此，本研究建議未來研究可以從以下幾個方面進一步探討：

1. 可以考慮將更多的因素納入情感趨勢分析中，例如政治、經濟、社會等因素。這樣可以更全面地了解社群媒體情感趨勢和股市趨勢之間的關係。

2. 可以進一步研究情感趨勢和股市趨勢之間的時間滯後效應。例如，可以分析情感趨勢對股市趨勢的影響需要多長時間才會產生，以及股市趨勢對情感趨勢的影響需要多長時間才會反映在社群媒體上。

3. 可以考慮使用更多種類的模型和方法來進行分析，例如深度學習、神經網絡等方法。這些方法可能可以更準確地捕捉社群媒體情感趨勢和股市趨勢之間的複雜關係。

4. 可以進一步探討股民對股市波動的反應和情感趨勢之間的關係。例如，可以分析不同股民對股市波動的反應和情感趨勢之間的關係，以及不同情感趨勢對不同股民的影響。這樣可以更全面地了解社群媒體情感趨勢和股市趨勢之間的關係。

因此，未來的研究希望可以從多個方面進一步探討社群媒體對股市的影響，以提高預測股市趨勢的準確性，同時借助更多跨領域的研究者力量，以此完成更加準確的趨勢及預測。

參考文獻

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805.*

Gupta, I., Madan, T. K., Singh, S., & Singh, A. K. (2022). HiSA-SMFM: historical and sentiment analysis based stock market forecasting model. *arXiv preprint arXiv:2203.08143.*

Jin, Z., Yang, Y., & Liu, Y. (2020). Stock closing price prediction based on sentiment analysis and LSTM. *Neural Computing and Applications, 32,* 9713-9729.

Kim, Y. (2014). Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. https://doi.org/10.3115/v1/d14-1181

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). ALBERT: A lite BERT for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942.*

Li, X., Xie, H., Chen, L., Wang, J., & Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems, 69,* 14-23.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692.*

Nguyen, T. H., Shirai, K., & Velcin, J. (2015). Sentiment analysis on social media for stock movement prediction. *Expert Systems with Applications, 42(24),* 9603-9611.

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016, October). Sentiment analysis

of Twitter data for predicting stock market movements. In *2016 international conference on signal processing, communication, power and embedded system (SCOPES)* (pp. 1345-1350). IEEE

Po-Chih Huang, Pu-Jen Cheng. (2018). Target Attention Network for Targeted Sentiment Analysis. http://tdr.lib.ntu.edu.tw/jspui/handle/123456789/71297

Qiu, Y., Song, Z., & Chen, Z. (2022). Short-term stock trends prediction based on sentiment analysis and machine learning. *Soft Computing, 26(5),* 2209-2224.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.*

Swathi, T., Kasiviswanath, N., & Rao, A. A. (2022). An optimal deep learning-based LSTM for stock price prediction using twitter sentiment analysis. *Applied Intelligence, 52(12),* 13675-13688.

Wu, S., Liu, Y., Zou, Z., & Weng, T. H. (2022). S_I_LSTM: stock price prediction based on multiple data sources and sentiment analysis. *Connection Science, 34(1),* 44-62.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J. G., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems (pp. 5754-5764).*