長庚大學工學院

學生校外實習期末報告

基於NLP技術的金融情緒分析與股票趨勢預測

實 習 期 間：自 112 年 07 月 03 日 至 112 年 08 月 25 日

輔 導 老 師 ：張哲維

學 生：B0928023 盧于璇

實 習 單 位 ： 人工智慧學士學位學程

實 習 機 構 ： 麥錫森智能科技股份有限公司

中 華 民 國 112 年 08 月 31 日

致謝詞

光陰荏苒，八週的實習歷程已經劃下句點。在這段期間，我獲益良多，心中充滿了感恩與滿足。

這次實習能夠美滿結束，讓我特別感激麥錫森智能科技股份有限公司的賴昭榮董事長。他在我們的實習期間充當了我們的技術導師，每次開會時都給予我們寶貴的建議和指引。當我們遇到技術上面的問題時，除了線上的技術教導，他更特地額外將時間空出，一對一的線上討論讓我們得以向他請教技術上的疑難。同樣地，我要特別感謝金融顧問Steven，在開會時他會與我們分享了許多金融領域的知識，並且糾正了我們錯誤的金融觀念，讓我們在進行實驗分析時，可以有更明確的方向。

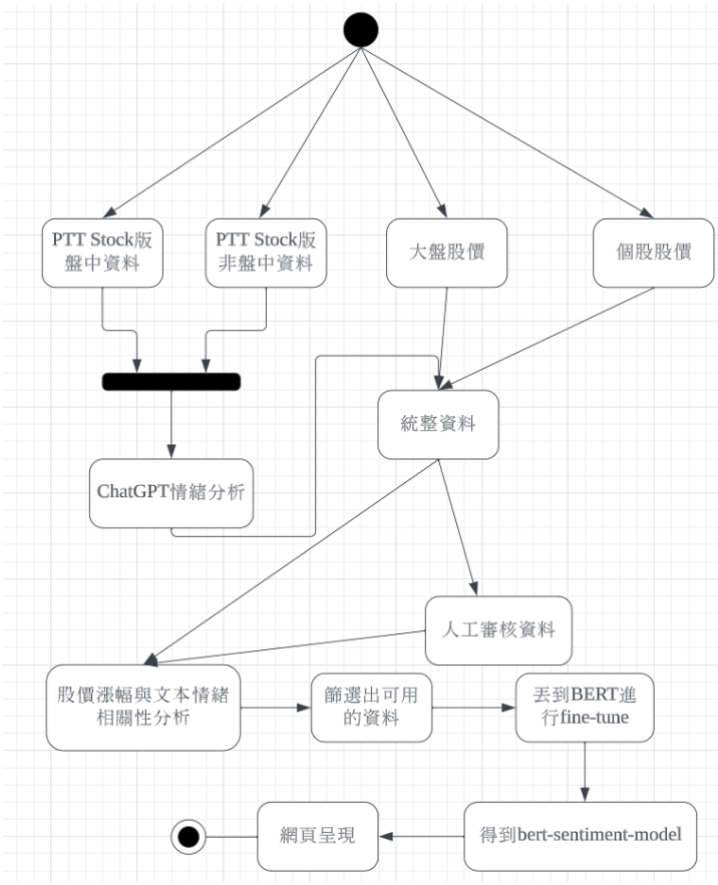
除了公司內部的幫助，我也要由衷感謝長庚大學資訊工程學系的張哲維教授。他提供了許多有價值的點子與建議，讓我們能夠不斷改進每個細節。我對於他的支持感激不盡。

此外，我也要感謝與我一同參與此實習專案的冠諭、君熙、詩晴。與他們的合作相當愉快，我們一同合作，各司其職。每次私下討論問題時，我們總能夠集思廣益，迅速解決問題，效率非常高。遇到挫折時，也會互相鼓勵扶持對方。

最後，我要感謝紹丞學長、沛錡學姊，謝謝他們在去年剛研究這個題目的時候，留下很多他們當初的記錄文件以及實作流程，讓我們在交接的時候可以快速進入狀況，順利進行這次的實習。除此之外，謝謝紹丞學長以及沛錡學姊在我們遇到實作上的問題或是對於實驗分析沒有想法的時候無私地提供我們過往解決方法的經驗以及想法，更撥空時間讓我們進行線上會議去詢問他們，由衷感謝。

摘要

本次實習，我們的目標主要是想要利用大型語言技術來分析重大的金融資訊去計算市場情緒指數，並透過情緒指數，去分析未來的個股趨勢，為此搭建了一個系統去驗證資料的可靠性與實用性。



圖一、系統架構圖

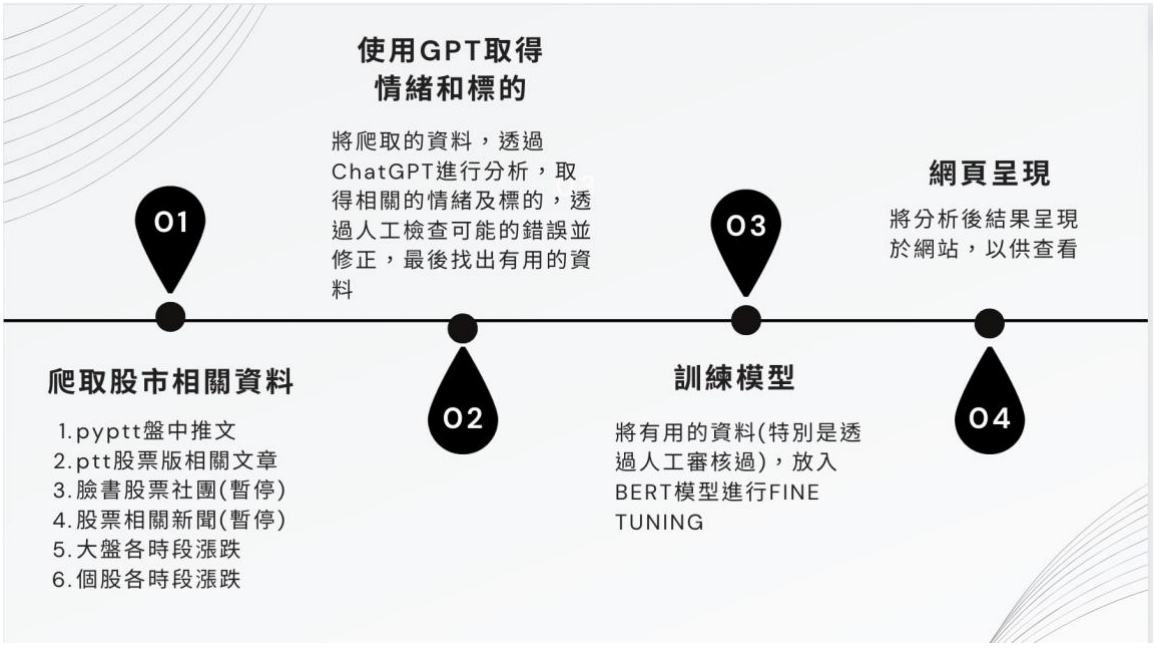
由於本次實習時間有限，以及任務的困難度較高，我們最終沒有達成原定目標，但在過程中解決了許多問題也發現了許多未來可能會遇到問題，並將其整理成文件交接，讓公司後續能進行參考快速入手。

在這系統我主要負責的是，PTT Stock版非盤中資料的爬蟲、臉書股票社團文章爬蟲、利用OpenAI API進行文本情緒分析、過濾資料及標記資料、文本情緒與股價漲跌的相關性分析、股價資訊爬蟲，以上我會在後續一一詳述。

目錄

實習目標

本次實習最一開始的目標，是將經過AI分析後的市場趨勢、股價漲跌預測，以及情緒指數等資訊整理彙總，並展現於前端介面上，同時也希望可以多收集不同平台的資料（例如:臉書股票社團、FinMind），如此一來，使用者在進行股票交易時，將能夠獲得更多有價值的數據參考。但後來經討論後，由於實習時間只有兩個月，於是將目標縮小至以分析PTT Stock版的發文及推文的情緒指數為主，在實習最後我們有得到發文及推文的情緒指數，但是無論是在資料搜集上，還是資料之處理都有太多的問題，因此花了較多時間在資料標記以及資料處理的問題，因此沒有做太多股價漲跌及情緒的相關性實驗，及完整的模型訓練。最後也因時間問題最終進行了收尾，並把實習過程的程式及文件打包，交接給公司。



圖二、系統流程圖

工作內容

以下是我所負責的工作內容：

- **利用多線程爬蟲方式獲取PTT Stock版非盤中發文及推文**:

  撰寫及修改多線程爬蟲程式，透過篩選條件過濾盤中閒聊的發文及推文，以順利抓取PTT Stock版上非盤中的所有資料。時間範圍為2022/10至2023/07，發文資料量共約1萬6000筆，推文資料量約80萬筆。

圖三、爬蟲流程圖

- **股價資訊爬蟲**：

  從台灣證券交易所的網站上，獲取特定股票在指定日期的交易資訊。

	日期	成交股數	成交金額	開盤價	最高價	最低價	收盤價	漲跌價差	成交筆數	漲幅%
0	112/07/03	15,118,041	8,743,824,984	578.00	580.00	576.00	579.0	+3.00	15,659	NaN
1	112/07/04	17,777,363	10,361,905,547	585.00	585.00	580.00	585.0	+6.00	18,848	1.04
2	112/07/05	15,553,503	9,060,750,346	589.00	589.00	579.00	582.0	-3.00	16,504	-0.51
3	112/07/06	32,069,711	18,234,491,768	573.00	574.00	565.00	565.0	-17.00	60,108	-2.92
4	112/07/07	19,858,943	11,244,712,238	565.00	572.00	563.00	565.0	0.00	21,855	0.00
5	112/07/10	18,996,089	10,794,393,087	567.00	573.00	565.00	565.0	0.00	17,792	0.00
6	112/07/11	18,566,571	10,665,245,909	574.00	577.00	570.00	577.0	+12.00	17,795	2.12
7	112/07/12	16,220,006	9,338,600,582	574.00	578.00	572.00	578.0	+1.00	14,426	0.17
8	112/07/13	26,878,397	15,790,086,719	587.00	590.00	585.00	585.0	+7.00	32,540	1.21
9	112/07/14	24,381,177	14,372,679,526	589.00	591.00	587.00	591.0	+6.00	26,789	1.03
10	112/07/17	14,311,753	8,428,551,949	588.00	591.00	587.00	591.0	0.00	18,226	0.00
11	112/07/18	22,022,410	12,866,196,578	587.00	588.00	580.00	581.0	-10.00	31,079	-1.69
12	112/07/19	24,909,638	14,517,904,258	584.00	587.00	579.00	581.0	0.00	20,069	0.00
13	112/07/20	15,676,734	9,102,325,307	580.00	584.00	578.00	579.0	-2.00	19,382	-0.34
14	112/07/21	52,275,128	29,291,080,481	560.00	563.00	557.00	560.0	-19.00	86,810	-3.28
15	112/07/24	27,656,825	15,469,188,055	557.00	563.00	557.00	558.0	-2.00	30,269	-0.36
16	112/07/25	22,767,807	12,899,067,927	561.00	569.00	561.00	569.0	+11.00	21,119	1.97
17	112/07/26	15,147,698	8,586,520,667	569.00	571.00	563.00	566.0	-3.00	21,679	-0.53
18	112/07/27	13,004,888	7,396,939,592	570.00	570.00	566.00	569.0	+3.00	12,672	0.53
19	112/07/28	19,009,675	10,814,265,557	569.00	573.00	565.00	567.0	-2.00	19,313	-0.35
20	112/07/31	28,409,339	16,044,638,425	575.00	575.00	560.00	565.0	-2.00	28,425	-0.35

圖四、股價資訊爬蟲結果(以台積電七月份為例)

- **利用OpenAI API進行文本情緒分析**：

  撰寫及修改情緒分析的程式碼，順利將PTT Stock版非盤中資料的爬蟲結果丟到ChatGPT進行情緒及標的的分析。時間範圍2023/03至2023/07，發文資料量共8000筆、推文資料量約40萬筆。

圖五、文本情緒分析流程圖

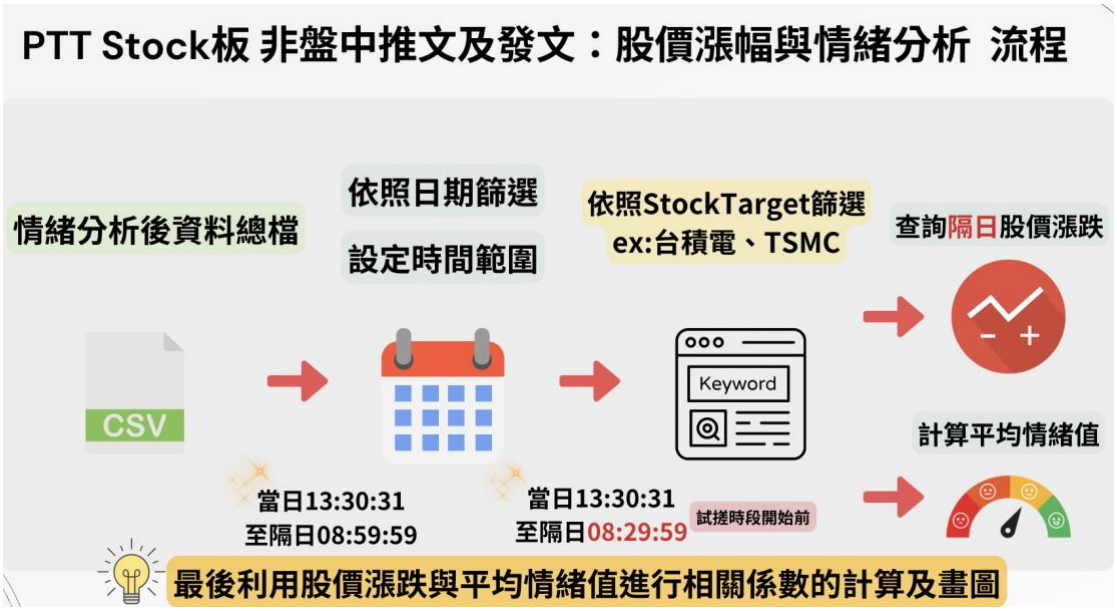
- **過濾資料及標記資料**：

人為過濾ChatGPT所分析出來的情緒及標的是否正確，若不正確的話，會進行人為標記資料。目的是為了不讓丟進BERT模型的有用資料過少，導致準確率不高的風險。

圖六、過濾資料及標記資料

- **文本情緒與股價漲跌的相關性分析**：

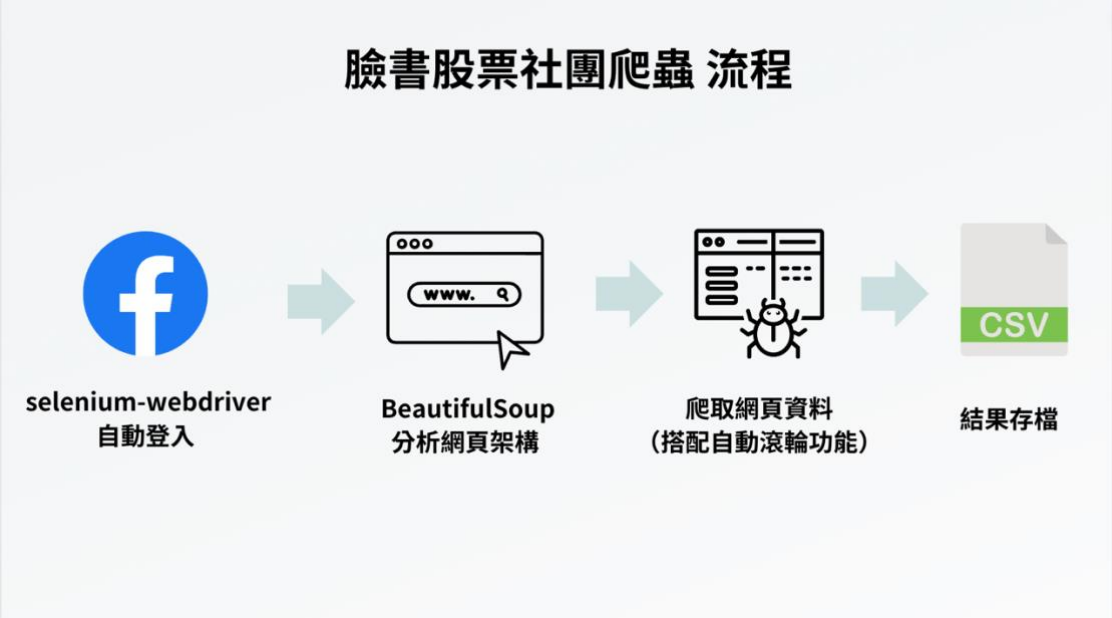
  將ChatGPT情緒分析完的文本結果（包含情緒及標的），依照個股的標的進行篩選，去對應（當）隔天的股價漲跌，計算相關係數值以及繪製相關係數圖。



圖七、文本情緒與股價漲跌的相關性分析流程圖

- **臉書股票社團文章的爬蟲**：

  欲透過收集不同平台的金融相關文章，使資料集的足夠，因此嘗試了臉書股票社團文章的爬蟲。透過Selenium 的爬蟲方法及搭配自動滾輪功能，最終有順利爬下500篇社團文章的貼文內容，但因最後以PTT Stock版上的文章為主，因此這部分後來先暫緩。



圖八、臉書股票社團文章的爬蟲流程

- **文件撰寫與程式整理**：

  將套件的使用以及使用步驟寫成GitHub README。

所學專業知識及技術

在專業知識方面，我在使用Pandas套件進行資料分析和處理方面積累了豐富的經驗。從最初的簡單操作到後來能夠從PTT中利用多線程爬取的方式成功地爬取相關資料，以及使用OpenAI API進行情緒分析。此外我透過閱讀學長姐提供的文件和程式碼，能夠更快速地找到和解決程式碼中的錯誤，使我在實習任務中更加得心應手。

除了在人工智慧系學到的相關專業知識外，我還在金融領域獲得了豐富的知識。儘管我之前對於股票一無所知，但通過這段時間的學習，我現在不僅熟悉股票市場的開收盤時間，還能夠分辨股票市場的上漲（紅色）和下跌（綠色）。除此之外，透過分析PTT股票版的發文及推文內容，讓我認識了很多上市公司名稱及代號以及一些股票的相關術語，這對我來說是一個巨大的成長。在實習的最後階段，我甚至開始思考如何結

合股票市場的情況來分析文章情緒與股價漲跌之間的關係，這使我在股票知識方面有了更深入的理解。

從技術方面來看，由於我撰寫了不只一種網頁爬蟲程式（PTT股票版非盤中閒聊資料、臉書股票社團、台灣證券交易所的網站），我對於分析網頁結構以及成功爬取數據的過程比過往更加熟練。此外，我也學會了如何對資料進行過濾，並運用DataFrame格式的數據進行各種技巧性操作。另外，我也利用GPT-3.5-turbo(0614)模型進行了文本情緒分析以及股票相關的問題探討。考慮到我之前沒有使用過OpenAI API的經驗，這也為我提供了一個很好的學習機會。

綜合來看，這段實習期間讓我在專業知識、金融領域以及技術方面都取得了顯著的進步。這些經驗將對我未來的學習和職業發展帶來寶貴的影響。

實習成果及貢獻

實習期間的成果與貢獻主要體現在兩個方面：文本情緒指數分析及股價漲跌相關性研究。

在文本情緒指數分析方面，我負責了PPT Stock 版資料收集（透過爬蟲）以及使用GPT-3.5-turbo（0614）版本的模型對 PPT Stock 版非盤中資料進行發文和推文的情緒分析（圖九）。除此之外，我也撰寫了台灣證券交易所的網站爬蟲，以獲取特定股票在指定日期的交易資訊。

	Date	Content	Sentiment	Stocktarget
0	2023-04-01 00:02:10	作者rayccccc (\GG/ \GG/ \GG/)看板Stock標題[新聞] 合一傷口醫...	positive	4743
1	2023-04-01 00:31:01	作者salonchen (康德)看板Stock標題[新聞] 北極星藥業-KY 4 / 6起得融資...	neutral	NaN
2	2023-04-01 01:16:26	作者madeinheaven ()看板Stock標題[新聞] 大陸網信辦對美光在華銷售產品 ...	negative	美光
3	2023-04-01 04:28:37	作者kyleger (打噴)看板Stock標題[請益] Re: [情報] 3432 台端 0...	negative	3432 台端
4	2023-04-01 05:46:18	作者afflic (afflic)看板Stock標題Re: [請益] 怎麼解決不敢ALL I...	negative	NaN
...	...	...	...	...
1411	2023-04-30 20:20:37	作者vflbdk (眼淚在笑)看板Stock標題[新聞] 上櫃ESG30指數小學堂活動5/2...	positive	NaN
1412	2023-04-30 21:11:41	作者twobosu ()看板Stock標題Re: [新聞] 雷虎簽下美雷達大廠MOU T-4...	neutral	NaN
1413	2023-04-30 21:35:17	作者awei181 (awei)看板Stock標題[新聞] 隨病毒門診爆量 安特羅疫苗7月底...	positive	6564-TW
1414	2023-04-30 22:43:22	作者thinksilver (跟白色的沉思)看板Stock標題[新聞] 軍工等5大題股 謝金...	positive	台泥
1415	2023-04-30 23:10:04	作者humbler (默人H)看板Stock標題[新聞] 不如美元定存！金控現金殖利率跌破3...	negative	NaN

	Post_Index	Tag	Userid	Content	Updatetime	Sentiment	Stocktarget
0	2308	→	shinewonder	這比軍火商來台的新聞還大條 雷虎真的要噴發了	2023-05-01 00:00:00	positive	NaN
1	2304	噓	soliboy	這篇洗地文到底要洗多久 每個平台都發一篇	2023-05-01 00:00:00	neutral	NaN
2	2299	推	lucky466	32額娃定存就是讚	2023-05-01 00:04:00	positive	NaN
3	2301	→	sa1989	http://i.imgur.com/kCORo9d.jpg	2023-05-01 00:07:00	neutral	NaN
4	2300	推	jeff79723	飯店真的不看好，觀光基本面只看好王品	2023-05-01 00:08:00	neutral	王品
...	...	...	...	...	...	...	...
127832	515	噓	tksq	看起來就是很會畫虎爛 哈哈	2023-05-31 23:58:00	negative	NaN
127833	480	推	Jerry469	AMD跌爛	2023-05-31 23:58:00	negative	AMD
127834	535	→	HSGJR77	https://youtu.be/8slG1EGbalw 不害怕	2023-05-31 23:58:00	neutral	NaN
127835	480	推	duckwei	過關噴的只有債券吧	2023-05-31 23:58:00	neutral	NaN
127836	480	→	Jerry469	彥州救救我QQ	2023-05-31 23:58:00	neutral	彥州

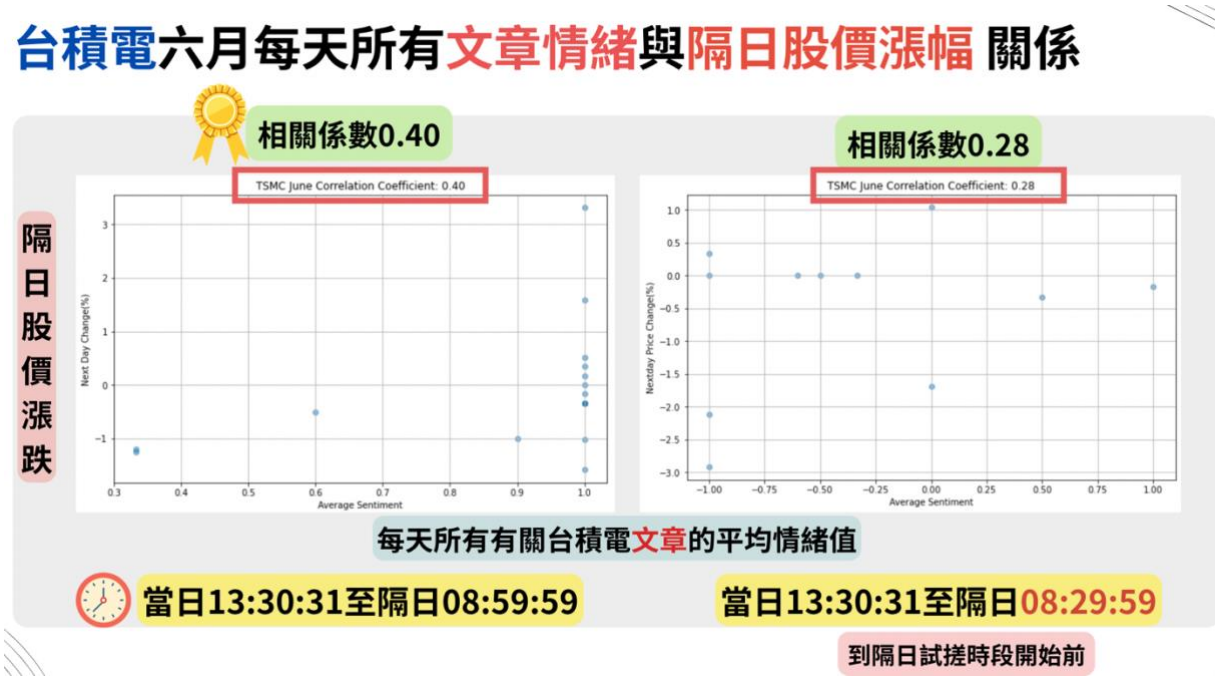
圖九、利用GPT-3.5-turbo（0614）情緒分析的結果

```
def sentiment_prompt_all(self, stock_market_chat):
    return f"""

    Analyze the sentiment of the following series of stock market chats.
    Each chats accompanied by a corresponding number,
    and the entire collection of the series chats is enclosed within triple backticks.
    請辨認 chat 的情緒是 "positive" or "negative" or "neutral".
    Your answer is a single word, either "positive" or "negative" or "neutral".
    如果 stock market target is not mentioned in the chat 請辨認情緒為 "neutral"
    otherwise, please list it in chinese, stock code or code name.如果沒有提及 stock code or stock name
    請將答案設為空字串""。

    Provide above answers in just JSON format which cloud convert to Pandas DataFrame
    directly with keys for the series chats separated by angle brackets,
    respectively:
    Sentiment, Stocktarget.

    Follow the format:
    {{
        "the corresponding number": {{"Sentiment": ..., "Stocktarget": ...}},
        "the corresponding number": {{"Sentiment": ..., "Stocktarget": ...}},
        "the corresponding number": {{"Sentiment": ..., "Stocktarget": ...}},
        ...
    }}

    Stock_market_chat text: {stock_market_chat}
    Please note, it is essential to adhere to the rules specified by the JSON formats and values.
    """
```

圖十、利用GPT-3.5-turbo（0614）情緒分析的結果

在文本情緒與股價漲跌相關性研究方面，我採用了兩種方法來探討這種相關性。首先，我將熱門台股特定時段的文本情緒與股價漲跌幅進行對應，並計算相關係數。結果顯示，在當天市場情緒與隔日股價趨勢之間呈現正相關，相關係數落在0.18~0.49之

間。這表明了市場情緒可能會影響隔日股價的變化。其次，我也對試搓時段的發文和推文情緒與股價漲跌之間的關係進行了研究，結果也顯示出相關性較高（圖十）。



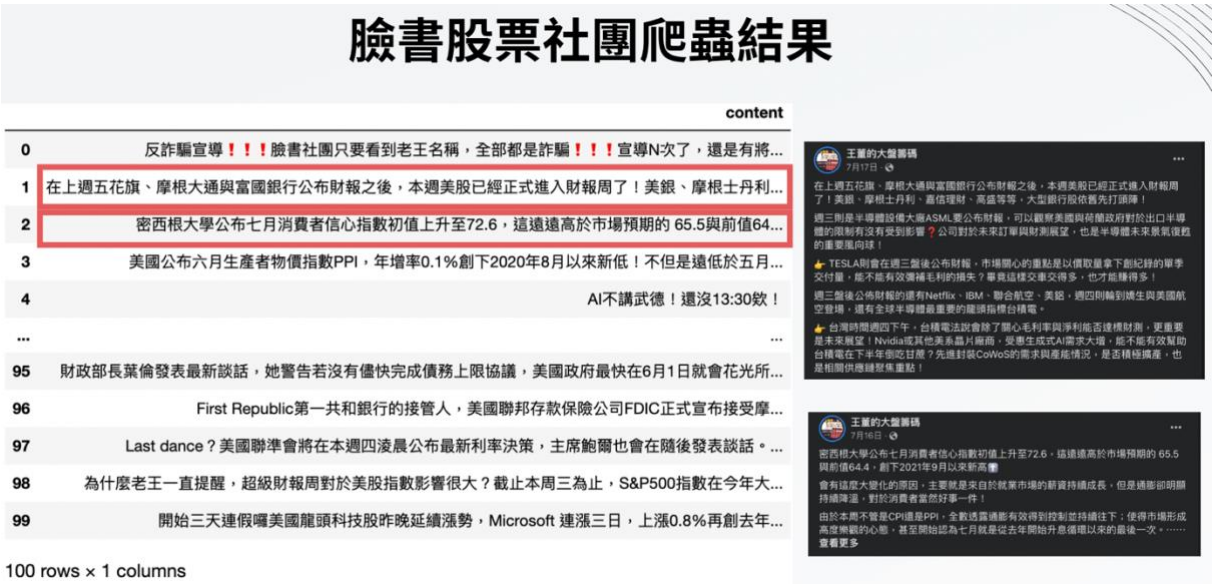
圖十、發文文章情緒與隔日股價漲幅關係（以台積電六月份為例）

然而，在進行分析時，我遇到了一些挑戰。部分股票的文章數量較少或過於集中在某幾天，可能導致分析結果不準確（圖十一）。為了解決這個問題，我進行了人工審核，檢查是否有缺失的標的或分析錯誤的情緒，並進行手動修改。

圖十一、發文文章情緒與隔日股價漲幅關係（以長榮海運五月份為例）

在實習的後幾週，我也花了大量時間進行逐行審核，雖然只能審核6月和7月的部分資料，但這些挑戰也將成為未來需要處理的重要方向。

此外，在初期，我嘗試從不同平台收集資料，特別是臉書股票社團。然而，臉書的網頁架構複雜，經常更改，對爬蟲造成了一些困難。最終，我使用Selenium進行臉書爬蟲，並撰寫自動滾輪的函式，以爬取連續頁面的文章內容（圖十二）。



圖十二、臉書股票社團爬蟲結果

總之，我在實習期間充分運用了爬蟲技能、情感分析模型、股價資訊和統計方法，成功地探討了文本情緒與股價漲跌之間的相關性。這些成果不僅豐富了我的專業知識，也為未來的研究和實踐提供了有價值的參考。

學習心得

在這段實習旅程中，我始終保持著謹慎的態度。在正式開始之前，得知我們的實習內容與科技金融有關，我既興奮又不安。興奮源於我對金融領域的好奇，我曾因此雙主

修學校的工商管理學系，希望在大學時期培養商業和金融相關知識。然而，我對於股票一無所知，甚至連開收盤的時間都不了解，更別提趨勢圖和K線圖了。因此，在實習初期，我感到有些措手不及。每次會議中提到股市相關的知識或術語，我都感到一頭霧水，結束後不得不上網查詢或向同學詢問賴博士提到的關鍵字。在實習初期，我真的遇到了不少挫折。

然而，隨著兩個月的實習經驗，我每天閱讀PPT股票版上的發文和推文，對於上市公司的暱稱、代碼甚至是討論術語，我竟然能倒背如流或是馬上知道術語的意思，這是我的一大進步！在實習過程中，我的主要成果之一是分析股價漲跌的相關性。儘管在這方面的概念和方法仍有待改進，但我還是有很大的成就感。相較於實習開始時，我已經能夠思考分析的方法，思考如何選擇特定時段的文本資料進行分析，以及如何將情緒指數與股價漲跌相關聯。儘管我使用的方法可能存在缺陷，例如我採用了情緒值的平均值這個方法會有一些細節會忽略到，但在整個實習過程中，我已經逐漸具備了思考並解決問題的能力。然而，也正因為對股市知識的不熟悉，在進行實驗時可能存在一些條件不周全的情況，例如美股開盤、試搓時段和期貨等，這些都是我未來需要繼續改進的地方。總之，儘管在分析過程中遇到了一些問題，但至少我已經得出了文本情緒和股價漲跌之間存在正相關的結論，這也為我的實驗研究提供了結果。

這兩個月的實習經驗讓我受益匪淺。我對科技金融充滿興趣，而幸運的是，實習內容也與此相關。儘管一開始對金融知識毫無概念，但在每週的進度會議上，賴博士總會給予我建議，告訴我哪些地方需要更周詳思考，這使我在實際分析時有了明確的方向，也給我很大的自由發揮空間。

最後，我相信若整個系統能夠成功實現，將會是一個極具潛力的系統，可能帶來顯著的獲益。儘管受限於時間，我們無法完成整個計劃，但我期待未來有機會再次參與類似的合作項目。

這段實習不僅充實了我的專業知識，也培養了我的分析思維和解決問題的能力，使我更有信心迎接未來的挑戰。

結論

儘管最終未能實現最初設定的目標，但在開發過程中，我獲得了豐富的專業知識和技術，不僅提升了實作技能，也加深了對金融知識的理解。這對我未來的成長來說是巨大的一步。儘管我目前對投資股票尚無經驗，但這次的實習經歷讓我相信，在未來實際投資時能少走一些冤枉路。再次感謝張哲維教授為我們尋找了這次實習機會，讓我有機會接觸一直以來感興趣卻未曾接觸的領域。同時，感謝賴昭榮董事長讓我們能夠遠程進行實習。過去兩個月的學習和經驗無疑對我來說是寶貴的財富，也讓我看到了人工智慧與其他產業合作的無限可能性。我相信我所學的知識將迅速在實際應用中發揮作用。

這段實習經歷為我帶來了深刻的啟示，不僅讓我提升了技能，更拓展了我的視野。我期待將來能將所學知識應用於更廣泛的領域，為自己的未來鋪平道路。

建議事項

或許我們可以打造一個股市資訊交流的專用平台，此平台專注於整理重要的財經新聞與股票資訊。這個平台可以提供使用者一個交流股市資訊的地方，讓他們快速獲取關

鍵的股市資訊。使用者可以在每則消息上，類似於Facebook的表情反應，以此迅速收集情緒資料，這樣就可以不用去擔心ChatGPT分析情緒是不是有缺漏或是有誤，我們也可以直接讀取該使用者所發出的這條訊息的原始情緒是什麼，就不用再花時間去手動審核，也可以省去很多時間。

這個平台的核心價值在於為投資者提供即時且有價值的資訊，讓他們能夠更好地做出決策。這種情感反應功能將有助於我們了解投資者對於不同消息的情緒反應，並能夠更準確地分析市場情緒的波動。透過使用者的表情給予，我們可以迅速量化情緒的變化，從而更好地把握市場的情況。

這個平台的設計層面也應該注重簡潔易用，以確保使用者能夠快速找到他們需要的資訊。同時，它也可以是一個社群平台，讓投資者可以在此互相交流意見和洞察。透過這種交流，使用者能夠更深入地理解市場趨勢，從而作出更明智的投資決策。

總之，這個股票資訊交流平台的建議有助於為投資者提供更便捷和準確的股市資訊，同時也能夠幫助我們更好地掌握市場情緒變化，以提升投資決策的準確性和效率。

參考文獻

[1]. The Fourth Workshop on Financial Technology and Natural Language Processing,80-150.

[2]. TweetFinSent: A Dataset of Stock Sentiments on Twitter,1-11.

一一二學年校外實習成果報告　基於NLP技術的金融情緒分析與股票趨勢預測　　人工四　盧于璇　撰