|  |  |
| --- | --- |
| **附件一** | 國立臺灣科技大學  資訊管理研究所  研究方法期末報告 |
|  | |

|  |
| --- |
| **金融社群之情緒分析探討台積電股票及台灣加權指數之關係**  **English(Times New Roman, 22pt)** |
| 研究生：黃雅婄  學號：M11209202 |
| 指導教授：XXX博士  中華民國 年 月 |

摘要

數位科技的飛速進步及通訊設備的便利性，使得全球人們能透過網際網路在社群媒體上進行無地域、無時差限制的即時交流，社群媒體不僅是討論人文、社會和經濟發展議題的重要傳播平台，同時也是金融理財討論的重要場域。

過去人們只能透過新聞、報章雜誌或公開資訊觀測站提供的重大訊息來了解股市現況，如今投資者們能夠自由在金融理財社群中分享股票操作經驗並討論市場趨勢，因此除基本分析和技術分析等方式外，消息面對經濟、股價帶來的影響也在科技發展迅速的背景下日益劇增，金融訊息傳遞與接收管道的改變也影響了大眾投資行為，從新聞、網路社群各方得知的消息都會影響投資人對股票的預期心理和操作態度。

本研究旨在透過大型語言模型和生成式AI等新興技術對文本進行自然語言處理（NLP），對社群媒體中台積電和台灣加權指數的評論和貼文進行情緒分析（Sentiment Analysis），並結合焦點小組來深度分析投資者對於社群媒體內容所產生的情緒影響程度和投資人行為，進而調整模型中情緒字詞的權重、以提升分析結果的準確度，針對社群中正面、中立和負面情緒詞彙，製作出金融市場之社群情緒詞典。

透過社群情緒分數，分析台積電與台灣加權指數股價與社群情緒之關聯性，最後做出股市預測，以幫助投資人即獲取最新的股票資訊，提升其投資決策的準確性和效益。

關鍵字： 台積電、台灣加權指數、社群媒體、情緒分析（Sentiment Analysis）、焦點小組（focus group）、大型語言模型（LLM）

**目錄**

[摘要 II](#_Toc167106625)

[圖目錄 V](#_Toc167106626)

[表目錄 VI](#_Toc167106627)

[第一章 緒論 1](#_Toc167106628)

[1.1 研究背景與動機 1](#_Toc167106629)

[1.2 研究目的 3](#_Toc167106630)

[1.3 研究架構 4](#_Toc167106631)

[第二章 文獻探討 5](#_Toc167106632)

[2.1 台積電與台股加權股價指數 5](#_Toc167106633)

[2.2 社群平台之情感分析 6](#_Toc167106634)

[2.2.1 情感分析 6](#_Toc167106635)

[利用投資人情緒預測股指的整合學習方法 7](#_Toc167106636)

[分析社群媒體中投資人情緒對股票報酬的影響：生存分析方法 7](#_Toc167106637)

[利用社會情緒和互動數據進行股票走勢預測：整合 NLP 和貝葉斯框架 7](#_Toc167106638)

[2.3 文字探勘與生成式AI的幫助大型語言模型 7](#_Toc167106639)

[第三章 研究方法 9](#_Toc167106640)

[3.1 實驗設計與架構 9](#_Toc167106641)

[3.2 研究假設限制 9](#_Toc167106642)

[3.3 內容分析法 9](#_Toc167106643)

[3.4 金融社群情緒字典與檢核 9](#_Toc167106644)

[3.5 情緒分析模型建置 9](#_Toc167106645)

[筆記 10](#_Toc167106646)

[字典權重設定 11](#_Toc167106647)

[其他修改建議 12](#_Toc167106648)

[參考文獻 13](#_Toc167106649)

圖目錄

[圖 1.1. 研究架構流程圖 4](#_Toc167035811)

表目錄

[表 1.1. 台灣加權指數市值佔比排行前10股票 5](#_Toc167057842)

第一章 緒論

1.1 研究背景與動機

科技推動技術創新，使人類生活變得更智慧便利，各項技術發展在科技的幫助 下進步飛速，進而促使世界連結更緊密。其中半導體產業為奠定科技進步的基礎，半導體製程產出的晶片元件應用範圍如雲端運算、物聯網、智慧家電等，近幾年熱度及技術越發穩定的自動駕駛汽車、人工智慧、生成式AI等多個重要領域都屬於其應用範圍，由此可見半導體技術可謂數位化世界不可或缺的重要先驅。

2022年，美國政府通過《晶片與科學法案》（CHIPS and Science Act），2024年2月，美國白宮宣布撥款鉅額執行《晶片研發計畫》（HIPS Research and development (R&D) programs），透過對半導體產業投資以激勵發展，進而加強美國整體經濟和供應鏈彈性，此法案間接促使全球半導體企業開始湧入大量的投資基金，也給產業帶來發展潛力可觀的前景。

科技發展也一併帶動了各國的經濟發展，近年來台灣國內商業、經濟也熬過了疫情的衝擊逐漸趨於穩定，特別是台灣股市交易活動也逐年熱絡起來，根據臺灣證券交易所統計之歷年股票市場概況表顯示，國內股市成交總金額從2019年的新台幣26兆元逐年上升至2023年的新台幣63兆元，股市的活躍程度也反映出股票這項投資工具越發受到投資人青睞。說到台股，不得不提及台灣股市中有「護國神山」美稱之台積電股票，作為台灣最重要的半導體代工企業，台積電不僅在技術創新方面領先，對國內金融經濟帶來的影響也十分顯著。

在全球市場對於半導體產業極度重視的背景下，推動產業發展之力量龐大，台積電在全球半導體供應鏈中的重要性也將進一步提升，而台積電股票對台灣股市帶來的影響也隨著產業發展產生更顯著的變化。

現代人們的日常交流逐漸社群化，使用者在在社群媒體中討論的內容也包含多 種主題，除了日常生活基本的食、衣、住、行、育、樂外，也在金融經濟、政治外交、科技人文等領域有諸多的討論。將社群媒體中的討論內容加以分析後，也在諸多領域實際應用，例如透過產品點擊率與搜尋內容讓企業更了解使用者偏好，對特定顧客進行精準的廣告投放，達到更好的行銷效果；在政治方面，也有利用網路投票預估選情，判斷不同地區選民意向的案例。

作為人們創作、分享、交流意見和觀點及經驗之平台，社群媒體能快速反應人們對事物的看法，其中討論台灣股市的社群平台如PTT、CMoney、Histock、Dcard股市版、鉅亨網......，使用者常在理財相關的社群平台分享對股市預測或交易結果。根據美網MagnifyMoney，在2021年對1,536名18至40歲受訪者的調查結果顯示，40歲以下的投資者中，有六成的人是金融理財論壇的會員，說明投資人會在理財社群平台活動、參考平台中的投資建議或大眾評論。且有兩成的投資人會同時在多個社群平台瀏覽貼文、留言作為個人投資參考依據。可見除新聞媒體及報章雜誌等傳統媒體外，現代投資人也在金融網站或理財社群平台獲取股市新資訊。

影響投資人選股及評價的股票分析方法有基本面、技術面、籌碼面和消息面等面向分析，由於社群網路已成為現代社會中重要的訊息交流平台，「消息面分析」所帶來的影響也在通訊網路普及的時代背景下對股市的影響更深遠，因此，分析社群網路中的情緒對於理解市場趨勢和做出投資決策具有重要意義，投資人會在平台上瀏覽產業資訊和社群輿論，並結合自身的金融知識進行股票交易，這些投資決策結果最終將反映到股市上。

情緒分析是自然語言處理（NLP）中的一個重要應用，傳統上使用基於詞典的方法或簡單的機器學習技術來分析文本情緒。隨著生成式AI和大型語言模型的出現，情緒分析的準確性和適用性得到了顯著提升。這些新技術利用深度學習和大量預訓練資料，能夠更準確地捕捉文本中的情感語境的細微差別，為研究提供更可靠的情緒分析結果。

本研究將結合焦點小組和機器學習技術，對台股中影響程度較大的台積電及台股加權指數的討論內容進行情緒分析，利用焦點小組的研究結果產生金融社群情緒字典的相關影響分析，優化模型中的情緒字詞之權重以提升情緒分數計算的準確性，並探索社群情緒與整體股價走勢之間的關聯。

新技術的應用不僅能提升情緒分析的準確性，還能為投資者提供更有價值的見解，有助於投資人綜觀全局、快速分析社群媒體中所含的情緒狀態，以做出更明智的投資決策。通過研究社群媒體中的社群情緒，為投資分析提供新的工具和方法，並推動相關領域的研究和應用，探究網路輿論對股市的真實影響，協助投資人進行投資決策。

1.2 研究目的

本研究旨在利用最新的機器學習技術，對社群媒體中有關台積電和台股加權指數討論版上使用者活動內容製成文本，對文本進行情緒分析產生社群情緒分數，並分析社群情緒對股市波動和投資人行為產生的影響。

同時在各大平台中建立獨立的討論區，結合焦點小組（focus group）的研究方式，蒐集投資人對社群媒體中大眾討論的股票消息的影響，透過不同的討論主題，實際蒐集出對投資人及大眾而言較容易受影響的金融社群情緒字詞，為情緒分析模型建立字詞權重的部分提供有效的基礎，加速模型產生的速度和提升情緒分數的準確度。

本研究的主要目的：

1. 金融社群情緒字典與情緒詞權重

以基礎的社群情緒字作為基礎，再結合焦點小組的研究方式對情緒字詞提供更有用的權重，包含常見的社群情緒詞彙和網路短語等，準確的字詞權重有助於提高情緒分析的準確性。

1. 社群情緒分析模型

利用生成式AI和大型語言模型來建立一個高效的情緒分析模型。此模型將被訓練來識別和分類社群媒體中有關台積電與台灣加權指數討論的情緒分數，包括正面、負面和中立情緒。

1. 社群情緒與股市活動之關聯

將情緒分析結果與實際股市進行比對，探討社群情緒變化與股市之間的關聯性，分析情緒波動是否能預測股市的短期或長期走勢、交易量的波動變化等股市交易活動。

1. 社群情緒對投資行為的影響

評估社群情緒對投資人投資行為的潛在影響。

1.3 研究架構

圖 1.1. 研究架構流程圖

第二章 文獻探討

2.1 台積電與台股加權股價指數

股票指數（stock index），又稱股價指數、股市指數，用於衡量整體股市表現的指標，股票指數的計算方式主要分為價格加權或市值加權，而各國股市皆有屬於自己的股價指數，如美國的道瓊工業指數、NASDAQ指數、S&P500等股價指數，還有歐洲的DAX指數、CAC-40和亞洲的富時中國A50指數、日經225指數等。

台灣股市也有台灣加權股價指數（Taiwan Capitalization Weighted Stock Index, TAIEX），常被國人簡稱為「大盤」，是台灣證券交易所（TWSE）編製的綜合性股票市場指數，用來衡量台灣股市整體的表現。

台灣加權指數是由台灣股市中具有代表性的上市公司股票所組成，每家公司的影響力取決於其市值，市值較大的公司在指數中的所佔權重就會較大，在台觀加權指數中市值佔比最重為台積電(2330)、第二為鴻海，緊接著是排名第三的聯發科，而大盤走勢反映了組成股票的整體波動情況，也代表著台灣股市整體漲跌的縮影。

表 1.1. 台灣加權指數市值佔比排行前10股票

|  |  |  |  |
| --- | --- | --- | --- |
| **排行** | **證券代碼** | **證券名稱** | **市值占大盤比重** |
| 1 | 2330 | 台積電 | 31.7881% |
| 2 | 2317 | 鴻海 | 3.3553% |
| 3 | 2454 | 聯發科 | 2.462% |
| 4 | 2382 | 廣達 | 1.5581% |
| 5 | 2412 | 中華電 | 1.4924% |
| 6 | 2881 | 富邦金 | 1.3953% |
| 7 | 2308 | 台達電 | 1.2916% |
| 8 | 2882 | 國泰金 | 1.1493% |
| 9 | 6505 | 台塑化 | 1.0671% |
| 10 | 2891 | 中信金 | 1.0382% |

台灣加權股價指數反映了整體股市表現，是投資人判斷股市情況的重要指標，因此大盤漲跌對投資人的心理預期會產生一定的影響，可能在大盤指數上漲時投資人會容易對股市呈現樂觀的態度，就更願意進行金融投資行為或對台灣股市更有信心；反之當大盤指數下跌時，投資者心態就容易變得保守，對有興趣的股票採取觀望或出售的投資策略。

我們也發現了針對加權指數進行分析的研究文獻，如劉照群(2008)對美國與台灣股市關聯性進行分析，在研究的共整合檢定中，發現費城半導體指數、台灣加權指數與台積電股價三者之間存在長期均衡關係。

Liao, Kao和Yuan的研究中探利用深度學習技術中的卷積神經網路（CNN），來對臺灣加權股價指數(TAIEX)進行股價預測。透過將台灣加權股價指數（TAIEX）的日金融數據轉換為圖像數據，並使用Xception模型進行訓練，結合Gramian技術將時間序列數據轉化為極座標表示，也提升了模型對股市趨勢的預測準確性。

對美國的S&P500加權股價指數也有相關研究，CHEN,KUAN-YU(2022)利用Glassdoor的員工評論作為文本，並結合Yahoo Finance和Macro Trends的數據，使用(Random Forest)、極限梯度提升(EGB)、循環神經網路(RNN)、長短期記憶(LSTM)等多個模型技術來預測股價走勢，並試圖找影響股價的因子。

Zhang等人研究了Twitter對美國大盤指數的預測（道瓊指數、納茲達克指數、標普500）。結果發現Twitter中的情感變化大盤指數呈現負相關。另外，Bollen等人利用OpinionFinder分析社群網站Twitter中的正負面情緒，再利用Google的向量工具把正負面情緒分為Calm、Alert、Sure、Vital、Kind、Happy六個向度來分析Twitter中的發文內容，最後透過回歸的模糊神經網路預測道瓊指數的收盤價。

2.2 情緒分析

2.2.1 傳統媒體情緒分析

社群媒體是自由、碎片化、具時效性的討論平台，與傳統媒體如新聞、報章雜誌等不同，社群媒體的交流特點之一是用字較不嚴謹、充斥著各種流行用語，使用者能隨時在社群平台上表達當下的想法及最新動態。

過去的文獻也有對傳統媒體做情緒分析的研究，CHEN, CHIAN-YU(2022)的研究中以公開資訊觀測站的重大訊息為文本，結合三個字典法:財經領域情緒詞典(FESD)、中文財務情緒詞典(CFSL)、中文金融情感詞典(CFSD)等對文本分析，但在研究中單純是利用累積關鍵詞的字數去判斷情緒值，且發現這類公告消息中較缺乏情緒詞，所以在判斷情緒效果較為不佳，且在研究中發現相較於PTT、facebook、Cmoney三平台中，Cmoney是投資者情緒指數最大的。

Juan,Zong-Yong(2022)的研究中利用bert對中大型股票的新聞資料進行分析，發現了對情緒詞做加權的處理方式能將準確率提高，但也要考慮文本的內容特點，發生重大事件、產業環境改變、新聞寫作風格等都會影響分析結果，另外發現「詞頻低影響大的詞」對情緒分數的影響容易被「詞頻高但較無意義的詞」所稀釋，故文本、加權、詞頻調整等對於預測結果有顯著影響。

2.2.2 社群媒體情緒分析

隨著資訊流通方式的變化及通訊電子在全球的普及，近年的研究中逐漸重視社群媒體在各方面的影響力。

隨著社群媒體的迅速發展和網路流行術語的頻繁變化，社群更新對投資者的情緒產生了直接影響。研究發現，社群媒體中的情緒波動與股票市場的波動具有高度關聯性，特別是在投資者決策過程中，社群輿論成為重要的參考依據。相比較於新聞中的情緒分析，由於新聞文字較為固定，社群媒體提供了更即時且多樣的數據來源，為投資行為研究提供了豐富的資料支持。

2.2.3 生成式AI文字探勘的應用與舊技術的差別

一、人工判讀相關文獻 Bryan (1997) 透過人工閱讀 250 篇 MD&A。Cole & Jones (2004) 閱讀 568 篇 MD&A。陳予得 (2021) 透過人工閱讀致股東報告書前十 行，探討其語調對未來公司績效之影響。人工判讀的方式較耗時費 力，較不適合處理大量資料。

近年來，社群情緒分析的相關研究不斷增加。例如，Kumar等人（2019）研究了Twitter上金融相關的推文，發現情緒變化可以顯著影響股價波動。Zhang和Huang（2021）的研究則強調了社群媒體更新和網路流行術語變化對投資者情緒和決策的影響。這些研究顯示，社群情緒分析能夠提供即時的市場情緒指標，對於理解市場趨勢和投資行為具有重要意義。

不同於英文的語言結構，中文因其語言的複雜性，例如多音字、多義詞及不規則的字句，使得中文的情緒分析應用困難度較高、技術相對英文不成熟。

[Fan, Y., Zhou, F., An, Y., Yang, J.](https://www.scopus.com/record/display.uri?eid=2-s2.0-85109266427&origin=inward&txGid=574a51709285a126ec7181bba80e0ee4) (2021)等人的研究中發現情緒對於股市確實會帶來影響，甚至有可能帶來股市崩盤的極端可能性

L. Zou, J. X. Zhang和W. Yang於2020年發表的論文《利用社交媒體的情感分析來預測股票市場變動》，探討如何利用社交媒體數據進行情感分析以預測股票市場趨勢。該研究使用自然語言處理技術來分析Twitter數據，並檢查公眾情緒與股票價格波動之間的關係。透過引入機器學習模型，作者展示了情感分析可以提供有價值的見解，並提高股票市場預測的準確性。一張含有 文字, 螢幕擷取畫面, 數字, 字型 的圖片

自動產生的描述

[利用投資人情緒預測股指的整合學習方法](https://www-sciencedirect-com.ntust.idm.oclc.org/science/article/pii/S0957417423022121" \l "b0140)

裡面有提到整合學習

整合學習因其在研究和應用中的出色表現而成為[機器學習領域的熱點（](https://www-sciencedirect-com.ntust.idm.oclc.org/topics/computer-science/machine-learning" \o "從 ScienceDirect 的 AI 生成的主題頁面了解有關機器學習的更多信息)[Dietterich，2000a](https://www-sciencedirect-com.ntust.idm.oclc.org/science/article/pii/S0957417423022121" \l "b0110)，[Dietterich，2000](https://www-sciencedirect-com.ntust.idm.oclc.org/science/article/pii/S0957417423022121" \l "b0115)）。它透過某種策略將多個[單一機器](https://www-sciencedirect-com.ntust.idm.oclc.org/topics/computer-science/single-machine" \o "從 ScienceDirect 的 AI 生成的主題頁面了解有關單機的更多信息)學習模型組合起來產生穩健的學習模型，可以提高機器學習模型的[泛化能力](https://www-sciencedirect-com.ntust.idm.oclc.org/topics/computer-science/generalization-ability)和穩定性。

[分析社群媒體中投資人情緒對股票報酬的影響：生存分析方法](https://ieeexplore-ieee-org.ntust.idm.oclc.org/document/8289945/references#references)

[利用社會情緒和互動數據進行股票走勢預測：整合 NLP 和貝葉斯框架](https://ieeexplore-ieee-org.ntust.idm.oclc.org/document/9885960/references#references)

社群情緒分析的相關論文發現與影響

社群更新與網路流行術語的變化

網路社群之影響與投資人的關聯

應用於新聞中情緒分析的很多因為文字比較固定

2.3 文字探勘與生成式AI的幫助大型語言模型

技術的應用、英文與中文差距、中文斷詞的困境與新興技術

第三章 [研究方法](https://daotw.com/%E8%B3%AA%E6%80%A7-%E9%87%8F%E5%8C%96%E7%A0%94%E7%A9%B6%E6%96%B9%E6%B3%95/)

3.1 文本資料蒐集與處理

3.1.1 文本蒐集

根據前提及「金融社群網站CMoney」是先前研究中實證結果情緒指數最大平台，故我們選定此Cmoney作為資料蒐集的平台。

本研究利用Python的套件中將「台積電個版」和「台股大盤」中有關台積電(2330)和台灣加權指數(TWA00)的發文內容、留言等資料透過爬蟲技術進行蒐集。

為了讓模型建立足夠的訓練資料集，且考慮到疫情期間對股市衝擊帶來的特殊影響，我們將爬蟲資料樣本區間設定為10年，為20年3月27日至2022年3月27日，總計366天，總文章篇數67941篇。研究中部分分析將排除股市未開市的天數，排除後的資料天數為244天，總文章篇數57798篇。

接下來本研究將以「交易日」作為排除未開市後總計共244天的資料代稱，以「全年」作為全年資料總計共366天的資料代稱。

資料代稱 資料天數 文章篇數

全年 366 67941

交易日 244 上漲 105 57798

下跌 120

平盤 19

表2、資料樣本說明

焦點小組的設計與研究計畫

1. 研究設計

1.1 焦點小組的目的

深入了解投資者對社群媒體內容的情緒反應及其對投資決策的影響。

蒐集投資者認為重要的情緒詞彙和短語，作為情緒分析模型調整的依據。

1.2 焦點小組的組織

參與者選擇：邀請10-15名活躍的投資者參與，每位參與者需在社群媒體上有至少一年的活躍發言歷史。

組織形式：分為2-3組，每組5-7人，確保討論深入且多樣化。

討論時長：每次討論約1.5至2小時，共進行3次討論。

2. 討論議題與流程

2.1 討論議題

投資者如何看待社群媒體中的討論內容？

問題引導：社群媒體上的討論對你的投資決策有何影響？

預期結果：了解投資者對社群媒體內容的關注點和影響程度。

社群討論內容對其投資決策的影響程度？

問題引導：你如何根據社群媒體上的討論調整你的投資策略？

預期結果：分析社群情緒如何影響投資者的決策過程。

常見的情緒詞彙和短語有哪些？

問題引導：在社群媒體上，你認為哪些詞彙或短語最能代表正面、負面和中立的情緒？

預期結果：蒐集並分類投資者認為重要的情緒詞彙和短語。

2.2 討論流程

引言與目標說明（10分鐘）

介紹焦點小組的目的和流程，說明參與者的角色和貢獻。

個人經驗分享（30分鐘）

讓每位參與者分享其在社群媒體上獲得的信息及其對投資決策的影響。

討論議題一：社群媒體的影響（30分鐘）

引導討論投資者對社群媒體內容的看法及其對投資決策的具體影響。

討論議題二：情緒詞彙與短語（30分鐘）

引導參與者分享和討論常見的情緒詞彙和短語。

總結與反饋（20分鐘）

總結討論內容，邀請參與者給出建議和反饋，並說明後續研究計畫。

3. 研究假設與限制

3.1 研究假設

投資者的情緒受社群媒體討論內容的顯著影響。

焦點小組能夠提供準確的情緒詞彙和短語，改進情緒分析模型的準確性。

3.2 研究限制

參與者樣本限制：參與者數量有限，可能無法完全代表所有投資者的觀點。

主觀偏差：參與者的個人經驗和觀點可能帶有主觀偏差，影響結果的普遍性。

4. 焦點小組的設立與實施

4.1 預備工作

招募參與者：通過社群媒體、金融論壇等途徑招募合適的投資者。

設計問卷：準備問題清單和討論議題，引導焦點小組討論。

4.2 討論實施

主持人引導：由經驗豐富的主持人引導討論，確保討論有序且深入。

記錄與分析：對討論過程進行錄音和筆記，提取關鍵資訊和情緒詞彙。

4.3 結果應用

模型調整：根據焦點小組提取的情緒詞彙和短語，調整情緒分析模型中的詞權重。

情緒詞典建立：將提取的情緒詞彙和短語整理成金融社群情緒詞典，提升情緒分析的準確性。

通過結合生成式AI和焦點小組，本研究將在情緒分析模型的構建和優化過程中，提供更準確和有價值的數據支持，進而提升對社群媒體中情緒的分析能力，幫助投資者做出更明智的投資決策。

### 研究方法計畫：結合生成式AI與焦點小組的情緒分析

#### 1. 研究設計與架構

\*\*1.1 研究步驟概述\*\*

- \*\*數據收集\*\*：從社群媒體平台（如PTT、Dcard、CMoney等）收集有關台積電和台灣加權指數的討論文本。

- \*\*數據預處理\*\*：去除噪音、停用詞和標點符號，進行分詞處理。

- \*\*焦點小組研究\*\*：組織焦點小組討論，以深入了解投資者對社群媒體中討論內容的情緒反應。

- \*\*生成式AI模型訓練\*\*：利用生成式AI模型對文本進行情緒分析和標註。

- \*\*模型優化\*\*：結合焦點小組研究結果，調整模型中的情緒詞權重，提升情緒分析的準確性。

- \*\*結果分析\*\*：將情緒分析結果與股市走勢進行比對，探討社群情緒變化與股市之間的關聯性。

#### 2. 具體方法

\*\*2.1 數據收集\*\*

- \*\*平台選擇\*\*：選擇台灣主要的金融理財社群平台（如PTT、Dcard、CMoney）作為數據來源。

- \*\*數據類型\*\*：包括討論帖、評論、回應等文本數據。

- \*\*時間範圍\*\*：設定特定的時間範圍（如最近一年的數據）進行收集。

\*\*2.2 數據預處理\*\*

- \*\*清洗數據\*\*：去除無關信息（如廣告、垃圾郵件）。

- \*\*分詞處理\*\*：使用自然語言處理（NLP）工具進行中文分詞。

- \*\*去除停用詞\*\*：去除常見的停用詞和標點符號。

\*\*2.3 焦點小組研究\*\*

- \*\*組織焦點小組\*\*：邀請10-15名活躍的投資者參與討論。

- \*\*討論議題\*\*：

- 投資者如何看待社群媒體中的討論內容？

- 社群討論內容對其投資決策的影響程度？

- 常見的情緒詞彙和短語有哪些？

- \*\*結果記錄\*\*：對焦點小組的討論進行錄音和筆記，提取關鍵情緒詞和短語。

\*\*2.4 生成式AI模型訓練\*\*

- \*\*選擇模型\*\*：選擇GPT-4等生成式AI模型進行情緒分析。

- \*\*訓練數據\*\*：使用收集的文本數據和焦點小組提取的情緒詞進行模型訓練。

- \*\*情緒標註\*\*：將文本標註為正面、負面或中立情緒。

- \*\*模型調整\*\*：根據焦點小組研究結果調整模型中的情緒詞權重。

\*\*2.5 模型優化\*\*

- \*\*調整情緒詞權重\*\*：根據焦點小組提供的意見，調整模型中情緒詞的權重。

- \*\*驗證模型\*\*：使用獨立的測試數據集驗證模型的準確性，確保模型能夠準確捕捉文本中的情緒。

\*\*2.6 結果分析\*\*

- \*\*情緒分數計算\*\*：計算每條文本的情緒分數，統計整體的情緒走勢。

- \*\*股市走勢比對\*\*：將情緒分析結果與台積電和台灣加權指數的股價走勢進行比對，分析情緒波動與股市走勢的相關性。

- \*\*投資行為影響\*\*：評估社群情緒對投資者行為的潛在影響，探索情緒變化是否能預測股市的短期或長期走勢。

#### 3. 新技術與舊技術的區別

\*\*舊技術（基於詞典和傳統機器學習）\*\*

- \*\*特點\*\*：

- 基於固定的情緒詞典進行情緒分析。

- 使用簡單的機器學習模型（如SVM、隨機森林）進行情緒分類。

- 需要大量人工標註數據。

- 難以捕捉文本語境和細微情緒差別。

- \*\*優缺點\*\*：

- \*\*優點\*\*：簡單易行，計算速度快。

- \*\*缺點\*\*：準確性低，適用範圍有限。

\*\*新技術（生成式AI和大型語言模型）\*\*

- \*\*特點\*\*：

- 利用深度學習和大量預訓練數據進行情緒分析。

- 能夠捕捉文本中的細微情緒差別和語境。

- 自我監督式學習，提高模型的準確性和適用性。

- \*\*優缺點\*\*：

- \*\*優點\*\*：準確性高，能處理複雜的情緒分析。

- \*\*缺點\*\*：需要大量計算資源和數據。

通過結合生成式AI和焦點小組，本研究可以提高情緒分析的準確性，並更好地理解社群情緒對股市的影響，為投資者提供更有價值的見解。

### 結合新技術進行情緒分析的建議

#### 研究方法架構

1. \*\*數據收集\*\*

- 從社群媒體平台如PTT、Dcard、CMoney等收集台積電和台灣加權指數相關的評論和貼文。

- 使用網路爬蟲技術自動化收集資料。

2. \*\*數據處理與預處理\*\*

- 進行文本清理，包括去除噪音、停用詞和標點符號。

- 使用分詞技術進行文本分詞。

3. \*\*情緒分析模型構建\*\*

- \*\*傳統方法\*\*：基於詞典的方法，使用現有的情緒詞典如NTUSD或SentiWordNet，進行簡單的詞頻統計和情緒計算。

- \*\*新技術\*\*：利用生成式AI（如GPT-3或GPT-4）和大型語言模型（LLM），通過深度學習技術進行情緒分類和預測。

4. \*\*模型訓練與優化\*\*

- 使用歷史數據進行模型訓練，並通過交叉驗證等技術優化模型參數。

- 將焦點小組提供的情緒字詞和權重加入模型，進行模型微調。

5. \*\*情緒分數計算與分析\*\*

- 根據模型輸出計算文本的情緒分數，分為正面、負面和中立。

- 分析情緒分數與股價走勢之間的關聯性。

#### 新技術與舊技術的比較

- \*\*準確性\*\*：新技術利用深度學習和大規模預訓練數據，能更準確地捕捉文本中的情緒細微差別；而傳統方法基於詞典，易受到詞典覆蓋範圍和更新頻率的限制。

- \*\*適應性\*\*：生成式AI能夠自動學習和更新，適應新詞彙和語境；傳統方法則需要手動更新詞典，較為僵化。

- \*\*計算效率\*\*：深度學習模型在硬件支持下計算效率高，但需要較大的計算資源；傳統方法計算量小，但精度和靈活性不足。

### 焦點小組的研究設計

#### 研究假設

1. 社群媒體上的情緒能顯著影響台積電和台灣加權指數的股價走勢。

2. 使用生成式AI進行的情緒分析比傳統詞典方法更準確有效。

3. 投資者對社群媒體中情緒詞的敏感度不同，影響其投資行為。

#### 研究限制

1. \*\*數據來源的偏差\*\*：僅選取特定平台的數據可能不全面，無法代表所有投資者的觀點。

2. \*\*情緒詞彙的多樣性\*\*：社群媒體中的情緒表達多樣，難以完全覆蓋。

3. \*\*技術局限性\*\*：生成式AI和大型語言模型的計算資源需求高，且可能存在過擬合問題。

#### 焦點小組的設立

1. \*\*小組組織\*\*：

- 選取不同背景的投資者，包括專業投資人、普通投資者和金融分析師。

- 組建多個焦點小組，每組6-10人，進行多次討論。

2. \*\*討論主題\*\*：

- 討論社群媒體上的股票信息對其投資決策的影響。

- 確定對其投資行為影響最大的情緒詞彙和表達方式。

- 評估不同情緒分析方法的準確性和實用性。

3. \*\*數據收集與分析\*\*：

- 逐字記錄討論內容，提取關鍵情緒詞彙。

- 將提取的情緒詞彙納入情緒分析模型，進行詞彙權重調整。

通過結合焦點小組的實地調研和生成式AI的技術優勢，本研究希望能夠提供一個更準確和實用的社群情緒分析工具，幫助投資者做出更明智的投資決策。

3.1 實驗設計與架構

專家法 焦點小組 建立平台或開啟

3.2 研究假設限制

3.3 內容分析法

3.4 金融社群情緒字典與檢核

3.5 情緒分析模型建置

筆記

~~金融重大訊息情緒分析之研究~~[~~https://ndltd.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=IgsQ9a/record?r1=1&h1=1~~](https://ndltd.ncl.edu.tw/cgi-bin/gs32/gsweb.cgi/ccd=IgsQ9a/record?r1=1&h1=1)

~~中大型股票新聞資料探勘那篇https://hdl.handle.net/11296/pnc6pw~~

The effect of stock index futures to stock market volatility

<https://ieeexplore-ieee-org.ntust.idm.oclc.org/document/6081227>

An integrated approach of ensemble learning methods for stock index prediction using investor sentiments

<https://www-sciencedirect-com.ntust.idm.oclc.org/science/article/pii/S0957417423022121>

The Rising Power of the Individual Investor: How Social Media Sentiments and User Activity Impact Stock Price Volatility and Trading Volume

<https://scholarship.claremont.edu/cmc_theses/2873/>

Does investor sentiment really matter?☆

<https://www.sciencedirect.com/science/article/pii/S1057521916301569>

An integrated approach of ensemble learning methods for stock index prediction using investor sentiments

<https://www-sciencedirect-com.ntust.idm.oclc.org/science/article/pii/S0957417423022121#b0130>

Tim Loughran 和 Bill McDonald 將財務領域的用詞分為這六大類情緒，奠定王釧茹團隊的軟資訊分析基礎。圖│研之有物(資料來源│Tim Loughran and Bill McDonald, 2011, When is a Liability not a Liability? Textual Analysis, Dictionaries, and 10-Ks, Journal of Finance, 66:1, 35-65.)

<https://research.sinica.edu.tw/computational-finance-wang-chuan-ju/computational-finance-cjwang-03/>

### 第二章 文獻探討

#### 2.1 台灣股市之台積電與台股加權指數

- \*\*2.1.1 台灣研究台積電與台股加權指數的文獻回顧\*\*

- 過去研究的概述

- 研究的不足之處

- \*\*2.1.2 台積電與台股加權指數之間的關聯性\*\*

- 影響因素分析

- 實證研究的發現

#### 2.2 社群平台之情感分析

- \*\*2.2.1 社群情緒分析的相關研究\*\*

- 情緒分析的基本概念

- 主要發現與影響

- \*\*2.2.2 社群更新與網路流行術語的變化\*\*

- 新興術語的出現及其影響

- \*\*2.2.3 網路社群之影響與投資人的關聯\*\*

- 社群媒體對投資決策的影響

- 投資人行為分析

- \*\*2.2.4 應用於新聞中的情緒分析\*\*

- 新聞情緒分析的應用

- 固定文本情緒分析的挑戰和解決方案

#### 2.3 文字探勘與生成式AI的應用

- \*\*2.3.1 技術應用概述\*\*

- 文字探勘技術的發展

- 生成式AI在情緒分析中的應用

- \*\*2.3.2 英文與中文文本處理的差異\*\*

- 英文文本處理技術

- 中文文本處理的特殊挑戰

- \*\*2.3.3 中文斷詞的困境與新興技術\*\*

- 中文斷詞技術的發展

- 相關新興技術的介紹

### 第三章 研究方法

#### 3.1 實驗設計與架構

- \*\*3.1.1 專家法\*\*

- 專家訪談的設計與實施

- \*\*3.1.2 焦點小組\*\*

- 焦點小組的組織與運作

- \*\*3.1.3 建立平台或開啟\*\*

- 平台建設的步驟與方法

#### 3.2 研究假設與限制

- \*\*3.2.1 研究假設\*\*

- 主要研究假設的提出

- \*\*3.2.2 研究限制\*\*

- 研究範圍與局限性

#### 3.3 內容分析法

- \*\*3.3.1 內容分析法的介紹\*\*

- 方法論的基礎

- \*\*3.3.2 應用於本研究的內容分析法\*\*

- 具體應用步驟

#### 3.4 金融社群情緒字典與檢核

- \*\*3.4.1 情緒字典的構建\*\*

- 字典構建的過程

- \*\*3.4.2 情緒字典的檢核\*\*

- 字典驗證的方法

### 字典權重設定

在建立情緒詞典時，可以考慮以下修改和設計：

1. **詞彙加權**：對於不同情緒詞彙，可以設計不同的加權值，以反映其在股市情緒分析中的重要性和影響力。例如，將正面詞和負面詞設置較高的權重，以便更有效地捕捉市場情緒的變化。
2. **專有名詞處理**：對於專有名詞（例如股票代碼、金融術語），可以單獨考慮其情緒價值或將其納入到情緒詞典中，以增強對於特定股票或行業的情緒分析能力。
3. **語境考慮**：考慮詞彙在不同語境中的情緒表達，例如同一詞彙在不同文章中可能具有不同的情緒含義，可以根據上下文進行調整和加權。

### 其他修改建議

除了字典權重設定外，還可以針對研究方法進行以下修改和調整：

1. **爬蟲技術優化**：尋找更有效的爬蟲技術，確保能夠全面且準確地收集社群媒體中的相關資料，考慮利用API或其他開放資料來源。
2. **斷詞工具比較**：除了使用Jieba和CkipTagger外，可以考慮其他中文斷詞工具或自然語言處理庫，以提高文本資料的處理效率和準確性。
3. **情緒分析模型**：尋找更適合金融市場情緒分析的模型和工具，例如LSTM、BERT等，以提高情緒分析的準確性和鮮度。
4. **結果應用優化**：除了提供市場情緒指標外，可以探索更多的應用場景，例如與其他技術分析方法結合，提供更全面的投資建議和決策支援。

這些修改建議可以根據具體研究需求和方法設計進行適當的調整和優化，以提高研究的科學性和可靠性。

#### 3.5 情緒分析模型建置

- \*\*3.5.1 模型建置的步驟\*\*

- 模型選擇與建置

- \*\*3.5.2 模型驗證與調整\*\*

- 模型的測試與優化

參考文獻

1. Evans, Julie Ryan. (2021). *Nearly 60% of Young Investors Are Collaborating Thanks to Technology, Often Turning to Social Media for Advice*. Magnifymoney. <https://www.magnifymoney.com/news/young-investors/>.
2. Liu, Chao-Chun. (2008). An Empirical study of the Relationship Among Philadelphia Semiconductor Index in American, Taiwan Stock Index and TSMC Stock Price (Unpublished master’s thesis). National Taipei University , Taiwan, ROC.
3. Liao, C. H., Kao, T. L., & Yuan, S. M. (2020). Forecasting Taiwan Capitalization Weighted Stock Index by Using Convolutional Neural Network. IEEE Xplore Logo. <https://ieeexplore-ieee-org.ntust.idm.oclc.org/document/9301956>
4. Chen,Kuan-Yu. (2020). Predicting Stock Price Trends using Social Reviews: Evidence from the S&P 500 Firms, (Unpublished master’s thesis). National Chung Cheng University, Taiwan, ROC.
5. Chen, Chian-Yu. (2022). A Study of Sentiment Analysis on Financial Material Information, (Unpublished master’s thesis). Feng Chia University , Taiwan, ROC.
6. Juan,Zong-Yong. (2022). 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, (Unpublished master’s thesis). National Yunlin University of Science and Technology , Taiwan, ROC.
7. [各種參考文獻的範例格式](chrome-extension://bocbaocobfecmglnmeaeppambideimao/pdf/viewer.html?file=https%3A%2F%2Fwww.archives.gov.tw%2FUserFiles%2Ffile%2F%25E5%258F%2583%25E8%2580%2583%25E6%2596%2587%25E7%258D%25BB%25E6%25A0%25BC%25E5%25BC%258F(105%25E9%2581%25A9%25E7%2594%25A82).pdf)
8. [網路資料的引用規則與產生器](https://wordvice.com.tw/apa-citation-generator)