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

### 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。

#### 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 金融社群情緒詞典與檢核

- \*\*3.4.1 情緒詞典的構建\*\*

- 詞典構建的過程

- \*\*3.4.2 情緒詞典的檢核\*\*

- 詞典驗證的方法

### 詞典權重設定

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

除了詞典權重設定外，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

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。