二、研究計畫內容（以 10 頁為限）：

（一）摘要

數位時代已經來臨，在科技飛躍性的進步和通訊電子的高度普及下，網際網路讓世界各地的人們能在社群媒體上進行無距離、無時差的直接交流，人們也經常在社群媒體上討論有關人文、社會、經濟發展等議題。

其中，使用者也在金融理財的社群媒體上分享對股票的操作、進行討論，除了基本分析和技術分析外，從新聞、網路社群各方得知的股票消息也會影響投資人對股票的預期心理和操作態度，本研究計畫透過文字探勘（Text Mining）和自然語言處理（NLP）等技術，針對社群中有關台積電股票的評論及貼文進行情緒分析（Sentiment Analysis），並應用資料庫統整成金融市場情緒詞典，探討社群中對股市相對影響的正面、中立和負面情緒詞彙，以利投資人交易時掌握最新的股票資訊，協助投資人進行投資決策。

關鍵字：股市、社群媒體、內容分析法、文字探勘（Text Mining）、情緒分析（Sentiment Analysis）、VADER Sentiment

（二）研究動機與研究問題

近年來台灣股市交易活動逐年熱絡，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 2016 年的新台幣 16 兆元逐年上升至 2020 年的新台幣 45 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，台積電隸屬晶圓代工產業，2021 年台積電代工生產份額更是佔全球的 56%，是全球半導體產業中重要的生產公司，其產品隨著技術多元化而精進。台積電影響著各行業的經濟命脈，可謂股市中不容忽視的一股力量。[1]

隨著科技日新月異，網路的普及讓人們隨時可以傳遞消息，不同類型的社群媒體也相繼推出。在 TWNIC 台灣網路報告中提及，累積至 2020 年，台灣 12 至 24 歲的 Z 世代網路使用率達到 100%、25 至 55 歲的 X、Y 世代的網路使用率也高達 95.3%，且網路服務使用項目中，12 至 39 歲的用戶使用社群論壇的比例高達 95.6%，40 至 55 歲之使用比率也達到 79.5%，可見各年齡層對網路的依賴程度與日俱增。現代人的日常交流逐漸社群化，使用者在社群媒體上的活動產生了大量資料流量，而這些大數據資料目前也應用在經濟、行銷、政治、人文等領域。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。[2]

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，已成為現代人日常生活密不可分的一部份。常見的社群媒體例如 Facebook、Podcast、Instagram、Twitter.....，都是近年來火紅的社群平台。其中也有討論股市的社群平台如 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 23%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，其中「消息面分析」卻常被視為輔助角色而忽視了其對整體經濟和個股評價的價值，鑒於多種調查及相關論文實證現代人交流趨向社群化，不只有社群活動的大數據資訊產生的附加價值，人們也開始關注社群對現實生活所帶來的影響。[3][4]

社群上股票評論已成為投資決策中重要的影響因素，投資人在平台上瀏覽產業資訊和社群輿論，結合自身的金融知識並進行股票交易，最後將決策反映到股市上。本研究認為，若將社群平台中關於股市的貼文、留言、論壇內容透過自然語言處理（NLP）之技術初步處理後，再對其進行語意分析（Semantic Analysis），得到之資料價值將有助於投資人綜觀全局、有效分析社群評論對股市的情緒狀態。

本研究以台積電為例，蒐集為期一年有關台積電股票之網路留言或貼文資料。不同於財經節目或新聞媒體制式化的報導內容，社群網路的評論內容較亂無章法，因此本研究須先以 Jieba（結巴）、word2vec、GloVe、VADER Sentiment 等相關技術，對文字進行斷詞、斷句等初步處理，後將文字資料轉換成規則的結構化資料，再透過自然語言處理（NLP）之技術進行情緒分析（Sentiment Analysis），將資料分類成正面、中立和負面等情緒標籤，並利用資料庫統整成相關的情緒詞典，驗證股價漲跌走勢與社群評論之關係，探究社群網路輿論對股市的真實影響，協助投資人進行投資決策。

（三）文獻回顧與探討

1. 情緒分析應用

社群媒體是一個零碎、具時效性的討論平台，與傳統媒體不同，社群媒體之評論用字較不嚴謹，使用者能隨時在社群平台上表達當下的想法及最新動態。

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。

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。[5]

2. 文字探勘應用

文字探勘和資料探勘區別可以由資料類型、資料明確及資料分析簡單區分。本研究主要使用技術是文字探勘。

	文字探勘（Text Mining）	資料探勘（Data Mining）
資料類型	非結構式資料。例如：文字	結構式資料。例如：數字
資料明確	文字意義因上下文不同而語意不同、文字意義模糊	數字較精確
資料分析	情緒探勘、情緒分析（Sentiment Analysis）、意見探勘	統計分析（Statistical Analysis）、預測模型（Predictive Model）等

表 1、文字探勘與資料探勘之比較表[19]

由於網路已成為傳播訊息的主要管道，其產生的訊息數量遠高於人工能負荷的範圍，因此，文字探勘（Text Mining）技術將取代人工分析，該技術因能處理大量文字資料、擷取訊息中涵義而極具商業價值。現今的文字探勘（Text Mining）也已運用在許多產業上，例如醫療、金融、房地產等都有相關研究分析及預測。

以金融市場的文字探勘（Text Mining）為例，Arman Khadjeh 等人彙整許多的文字探勘文獻並歸納出流程圖。文字探勘流程可以簡單分為三個步驟，首先為文本資料擷取與輸入，接著為資料處理（Data Processing），最後以機器學習（Machhine Learning）的方式來進行金融市場的漲跌預測[6]。

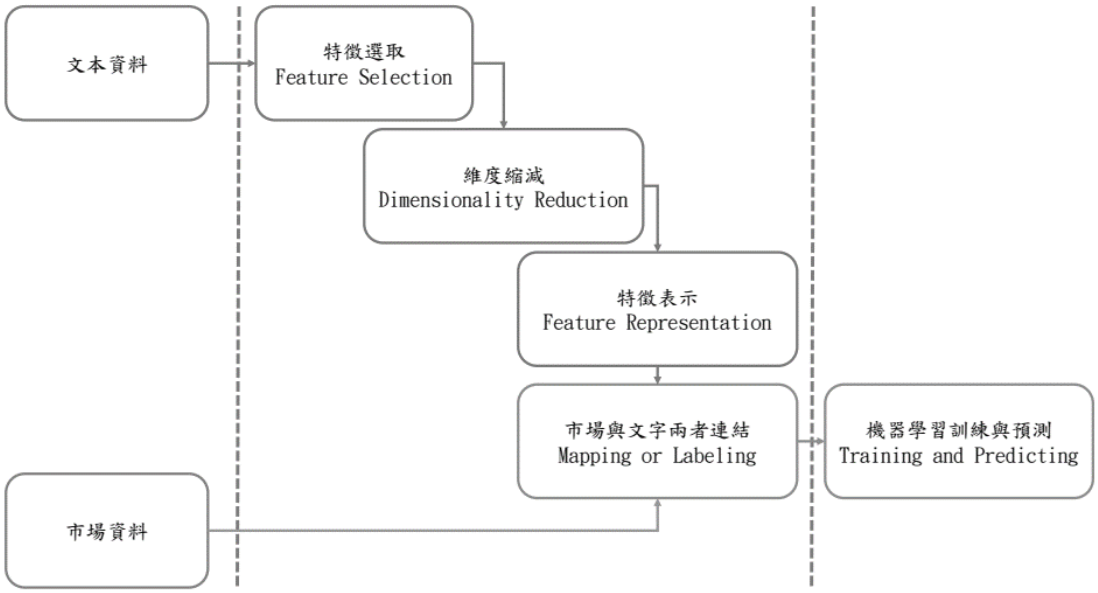

圖 1、文字探勘三階段

文字探勘步驟：

1. 使用爬蟲技術抓取文本資料和市場資料。
2. 對文本資料進行斷詞（特徵選取）、去重複等處理，再製作成詞典（特徵標示）。
3. 將資料交由機器學習，最後進行市場預測。

3. 股市輿論風向的影響

隨著科技發展，社群討論逐漸成為人民對股市預期心理的指標。為了哄抬股價，有些公司會選擇買輿論（帶風向）的方式增加股票的討論度，利用社群輿論的方式營造公司良好的形象。而投資人若缺乏專業分析能力（如基本、技術面分析等），容易因為輿論風向而混淆判斷，例如某家公司在各社群中以假人頭的方式散布未來將大量出口增加訂單的議題，讓非理性投資人看好公司未來效益而跟進。根據信心股價理論，投資人若對於股市情況樂觀，信心越強，就必然以買入股票來表現其心態，股價因而上升。事實上公司並沒有能力運用投資人的資金產出更大效益，而這樣的情勢逐漸導致投資人失去投資信心，股價下跌，原本看好的投資人不僅無法得到理想預期報酬，甚至出現賠本的情況發生。大部分投資人跟隨著這樣的風向判斷跟買/賣，同時也反映股價將上漲或下跌，所以根據情緒分析的結果可即時反應投資市場傾向程度，這樣的傾向與股價的走勢之間有可循的關聯，即可用自然語言處理（NLP）中的情緒分析來預測股價。[7][8]

4. 內容分析法

內容分析法又稱文本分析，源自十八世紀中的瑞典。Bowers（1970）定義出的內容分析，不只針對內容分析的方法是否客觀且有系統與量化，更著重在內容分析的價值，將內容利用系統客觀和量化方式加以歸類、統計，並且根據這些類別的統計模型做推論。相較傳統分析方式較符合經濟效益。[9][10][11]

（四）研究方法及步驟

本研究目標為探討社群網路平台關於股市的正、負和中立情緒詞彙，有別於傳統資料探勘所處理「結構性」的資料，文字探勘（Text Mining）所處理的資料是沒有特定結構的純文字，考慮到本研究探討社群領域會有許多非結構式資料，故採用文字探勘的技術。

以往相關文獻多以新聞領域為研究焦點，與社群平台相較之下，新聞的專業用詞多、資料範圍較小。本研究認為，隨著科技世代蓬勃發展，人與人的交流與互動逐漸社群化，有些精簡化的詞也賦予豐富情緒，因此社群網路的討論聲量也成為投資判斷重要的一環。藉由此研究動機建立出以下研究流程圖（如圖 2），並將研究流程圖區分為三大部分，分別是：

第一部分：確立研究目標，針對社群平台有關台積電一年內資料為研究範圍。
第二部分：執行研究，運用相關技術分析自然語言處理。
第三部分：將研究結果運用在系統中，提供股市投資者判斷的依據。

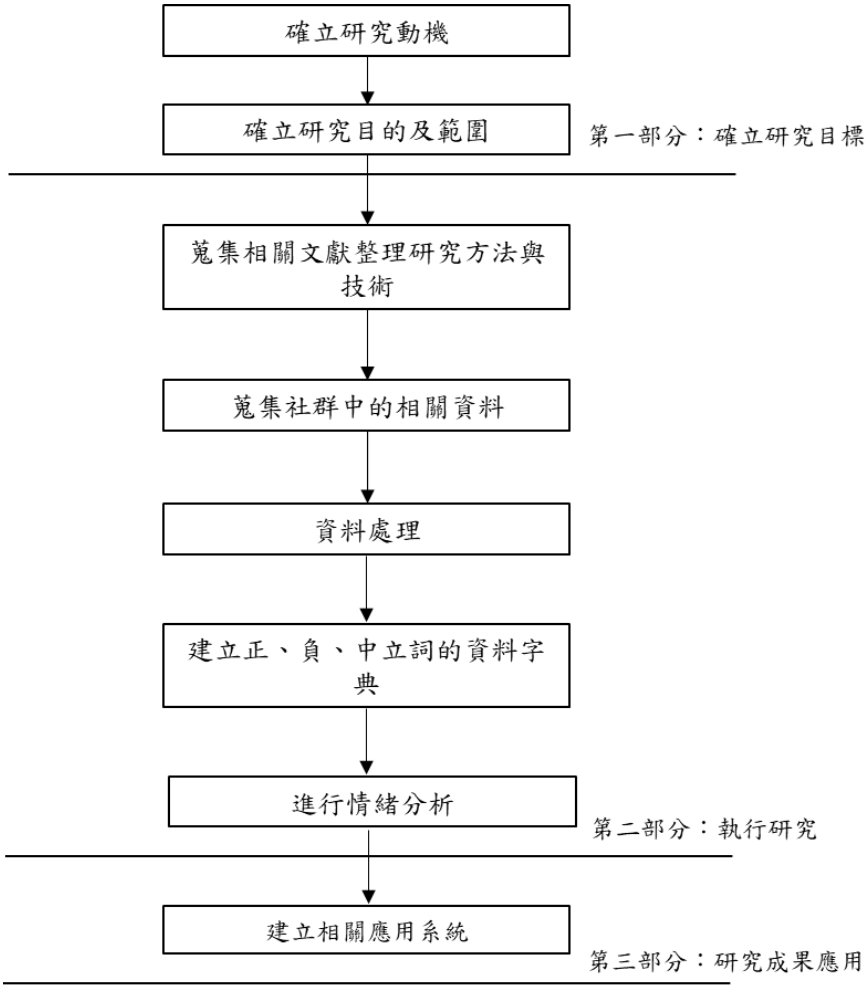

圖 2、研究流程圖

第一部分：確立研究目標

研究目標是建立社群媒體中金融市場情緒詞典。根據美國金融研究公司 InvestorPlace 2021 年文章指出，台積電掌握極紫外線（EUV）技術，擁有全球一半安裝基礎和 60%產量，已成為全球重要的半導體公司，因此台積電同時也成為台灣股市影響因子之一。本研究認為針對台積電分析情緒並建立詞典，可以讓投資人更有效判斷股價走勢，故以台積電於社群媒體一年內的相關文章為研究範圍。[12]

第二部分：執行研究

研究相關文獻中的技術，首先以爬蟲技術蒐集股市的相關資料，在此步驟中，將抓取資料範圍分為兩部分，第一部分為市場資料，以股市文章做為資料範圍進行內容分析法，給定情緒分析套件依據的金融市場情緒詞典，第二部分為文本資料，以社群媒體中股市論壇有關台積電一年內的文章為範圍，用來執行情緒分析，實現本研究目標。抓取資料的過程使用 Selenium 及 PhantomJS 的分散式爬蟲框架來完成網站中資料的蒐集。[13][14]

1. Selenium 是一個瀏覽器自動化的套件，可以利用 Python 撰寫自動化的腳本訪問瀏覽器，包含開啟瀏覽器、點擊按鈕、填寫表單及取得網站內容等，用以簡化繁瑣及耗時的網站讀取工作，是 Python 自動化應用非常重要的套件。

2. PhantomJS 是一種基於 webkit 的 JavaScript API。作為核心瀏覽器的功能，使用 webkit 來編譯解釋執行 JavaScript 程式碼、CSS 選擇器、JSON、HTML5、Canvas 和用來縮放向量圖形的 SVG 等，最主要應用於模擬人類操作系統，如網路監測、網頁截圖、Web 測試、頁面訪問自動化等。

文章範例	
範例一	沒意外會再倒一波，636 賣單加堆，635 買單沒人掛，連假單都不掛，就是要倒了的意思
範例二	期指 17747--17824 高要過才會繼續攻，2330 不夠強,637 要過，今天期指 17702 不破也算還好，關鍵需要 5-8 天時間整理，上面還有缺口要回補也不要看太壞
CkipTagger 斷詞範例	
範例一	[['意外','倒','賣單','加堆','買單','人','掛','假單','掛','倒','了','意思'],
範例二	['期指','高','才','繼續','攻，','夠','強,','今天','期指','破','還好','關鍵','需要','時間','整理','上面','缺口','回補','看','太','壞']]
詞典範例	
正面詞	買單、高、過、攻、強、整理、回補
中立詞	假單、加堆、掛
負面詞	倒、賣單、缺口、壞
專有名詞	期指

表 2、文章標註範例表

將兩部分已抓取完的資料使用 Jieba（結巴）及中研院的 CkipTagger（又稱 Ckip）來進行斷詞、去重複等資料處理，處理完的資料稱為詞袋（Bag of Words Model），將第一部分的詞袋進行內容分析法，由研究者訂立出分類邏輯確保一致性，再製作成詞典，類別有正面詞、中立詞、負面詞、專有名詞，範例（如表 2），目的是提供情緒分析套件辨別情緒的依據，而第二部分的詞袋將匯入後續的情緒分析套件中進行分析。

接著此過程將使用機器學習的其中一種技術－自然語言處理（NLP），此技術解決了口語及書面語言在計算機輔助分析語言之間的困難。本研究探討自然語言處理的一個的特定領域－情緒分析（Sentiment Analysis），由於研究需求是將文本資料分類為情感類，因此需要將已建立好的詞典匯入到套件中，再由套件自動將詞袋中的文本資料進行分類，以計算方式將詞囊轉換為數字數據或是嵌入（Word Embedding）到維度之中，使用到的套件有 word2vec、GloVe、VADER Sentiment。

1. word2vec 為 Google 公司推出的開源工具之一，能根據輸入的詞袋計算出詞與詞之間的距離。word2vec 將「字詞」轉換成「向量」形式，把詞袋中的詞簡化為向量空間中的運算，以計算出向量空間上相似度的方式，來表示語義之相似度，例如「空單」的詞向量預測到高機率出現在該向量附近的「看跌」。[16]

2. GloVe（Global Vectors for Word Representation），它是一個基於全局詞頻統計（Count-based & Overall Statistics）的詞表徵（Word Representation）工具，可以將單詞由實數組成的向量來表示。這些向量能夠捕捉單詞之間的語義特性，如相似性（Similarity）、類比性（Analogy）等。[17]

   該技術實現模型過程分為三大步驟：

   第一步驟：根據前述已建立的字典構建一個共現矩陣（Co-occurrence Matrix）χ，矩陣中的每一個元素χ_{ij}代表單詞 i 和單詞 j 在字典中（Context Window）內共同出現的次數，根據兩個單詞在字典中的距離d，提出了一個衰減函數（Decreasing Weighting）：decay=1/d，用於計算權重，距離越遠的兩個單詞所佔總計數（Total Count）的權重越小。

   第二步驟：構建詞向量（Word Vector）和共現矩陣（Co-occurrence Matrix）之間的近似關係，並計算以下公式表達兩者關連：

   $$w_i^T \overline{w}_j + b_i + \overline{b}_i = \log\ (\chi_{ij})$$

   其中，w_i^T和$\overline{w}_j$是我們最終要求的詞向量；b_i和$\overline{b}_i$分別是兩個詞向量的偏誤。

   第三步驟：構造它的損失函數，依照最基本的均方誤差建立權重函數$f(\chi_{ij})$，而期望權重須滿足大於較少出現在一起的詞（Rare Co-occurrences），所以函數必須是遞減函數，推導出以下公式：

   $$\mathrm{f(x)} = \begin{cases} (\chi/\chi_{max})^{\alpha} \\ 1 \end{cases} \quad \begin{matrix} if\ x < x_{max} \\ otherwise \end{matrix}$$

3. VADER Sentiment 提供了正面、中立、和負面情緒的衡量標準。此模型是專門為社群媒體文本數據所開發和調整，VADER 接受了一組完整的人類標記過的數據的訓練，包括常見的表情符號、UTF-8 編碼之表情符號、口語術語和縮寫（例如 meh、lol、sux）對於給定的輸入文本數據，轉換後的文本數值為[-1,1] 範圍內的數值，其中數值大於 0.001 的情緒為正面詞，數值小於-0.001 則被認為是負面詞，否則為中立。例如「The weather is not good」經轉換後可得到文本數值為-0.3412，可判斷為負面詞。「The weather is great」文本數值為 0.6588，可判斷為正面詞。[18]

第三部分：研究成果應用

接續第二部分的分析結果建立資料庫。本研究目標建立一套方便投資人參考的系統，能有效幫助投資人在關鍵時刻做出正確的判斷，假設台積電股票目前正面反應 32%、負面反應 45%、中立反應 23%（本研究以台積電 2330 為期一年的社群資料為範例），提供整體的市場情緒指標，利用機器學習技術降低時間成本，並且根據金融情緒詞典預測股市漲跌，驗證輿論對股價的真實影響，並且達到更客觀的股市情勢分析。

（五）預期結果

使用者經常在社群軟體上討論股市預測及結果，因此除了基本面分析和技術分析外，從社群軟體上得知的股票消息經常也會影響投資人對股票的預期心理和操作態度。本研究計畫期望建立出社群媒體中影響股市字詞的情緒詞典。透過文字探勘，將使用者在各大社群媒體上所發文及留言的字詞中，對於將影響股市的字詞分類為正面、中立及負面，於系統中提供市場情緒指標給投資人參考，以利投資人在交易股票時能夠掌握當下的市場情緒，協助投資人進行投資決策。

（六）參考文獻

[1] 台灣證券交易所【歷年股票市場概況表】年報
https://www.twse.com.tw/zh/statistics/statisticsList?type=07&subType=232
[2] 台灣網路報告 TWNIC
https://report.twnic.tw/2020/
[3] 股票分析四大面向：基本面、技術面、籌碼面、消息面
https://enjoyfreedomlife.com/four-aspects-of-the-stock-market/
[4] Nearly 60% of Young Investors Are Collaborating Thanks to Technology, Often Turning to Social Media for Advice
https://www.magnifymoney.com/blog/news/young-investors-survey/
[5] 朱夢珺，蔣洪汎，許偉.（2016）. 基於金融微博情感與傳播效果的股票價格預測.
[6] 文字探勘與機器學習於股票市場的應用與三大步驟
https://bigdatafinance.tw/index.php/data-visualization/862-2019-05-26-14-56-40
[7] 信心理論
https://wiki.mbalib.com/zh-tw/%E4%BF%A1%E5%BF%83%E7%90%86%E8%AE%BA
[8] 顏士杰.（2021）. 透過傾向分析進行股價趨勢預測的實務經驗分享.
[9] 內容分析法
https://nccur.lib.nccu.edu.tw/bitstream/140.119/32479/7/75201407.pdf
[10] Yu-Teng Su.（2019）. Research on the Analysis of Marketing Strategy of Fitness Club by Content Analysis.
[11] Sheng-Cheng Chung.（2019）. A Study on the Development of Hydropower Industry by Content Analysis.
[12] 美國金融研究公司台積電全球地位文章
https://investorplace.com/2020/09/tsm-stock-the-most-important-company-in-the-world/?mod=mw_quote_news&fbclid=IwAR1ZTRAsBQHGWQmo1yB-8E5dHgtnqQdz4kpuGAf04IgAj2tCX4Xg-7B5bFM
[13] 爬蟲技術
https://www.itread01.com/content/1546794011.html

[14] 张晔 孙光光 徐洪云 庞婷 曲潇洋,（2020）. 国外科技网站反爬虫研究及数据获取对策研
[15] 從 word2vec 到情感分析
https://studentcodebank.wordpress.com/2019/02/22/%E7%B9%81%E9%AB%94%E4%B8%AD%E6%96%87-nlp-%E5%BE%9Eword2vec%E5%88%B0-%E6%83%85%E6%84%9F%E5%88%86%E6%9E%90/
[16] word2vec 簡介
http://ilms.ouk.edu.tw/d9534524/doc/43713
[17] Es, S.（2020）. Sentiment Analysis in Python: TextBlob vs. Vader Sentiment vs. Flair vs. Building It From Scratch.
[18] Shihab Elbagir and Jing Yang.（2019）. Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment
[19] 資料探勘與文字探勘之比較
https://medium.com/marketingdatascience/%E8%B3%87%E6%96%99%E6%8E%A2%E5%8B%98%E8%88%87%E6%96%87%E5%AD%97%E6%8E%A2%E5%8B%98%E4%B9%8B%E6%AF%94%E8%BC%83-4410964ded2e
[20] Sung-Shun Weng.（2008）. Using Support Vector Machine and Text Mining For Stock Price Trends Prediction.
[21] Nielsen, A.（2020）. Practical Time Series Analysis: Prediction with Statistics & Machine Learning, O'Reilly
[22] Uhr, P., J. Zenkert, and M. Fathi.（2014）. Sentiment Analysis in Financial Markets,'' 2014 IEEE International Conference on Systems, Man and Cybernetics（SMC）: San Diego, CA, USA.
[23] Bhati, R. G.（2020）. Sentiment Analysis: a Deep Survey on Methods and Approaches. Int'l Journal of Disaster Recovery and Business Continuity. Vol. 11, No. 1, pp. 503-51.
[24] Bohmian.（2020）. Sentiment Analysis of Stocks from Financial News Using Python.
[25] Briggs, J.（2020）. Sentiment Analysis for Stock Price Prediction: How we can predict stock price movement using Twitter.
[26] Lin Yu.（2020）. Princing Anomaly from the Text Sentiment in Social Community Forum.
[27] 莊凱翔.（2018）. The prediction of trend toward stock price by text mining and sentiment analysis on social media: Using SVM and LDA Algorithm

（七）需要指導教授指導內容

本研究主要利用文字探勘（Text Mining）技術，以網路爬蟲、Selenium 及 PhantomJS 等技術蒐集研究之資料，透過自然語言處理（NLP）和 word2vec 等技術對文字進行處理、分析，實證社群平台中有關台積電股票之評論是否對現實台灣股市造成影響，以利用本研究產生之金融情緒詞典輔佐投資人進行投資決策。

為正確使用技術以達到準確的情緒分析（Sentiment Analysis）、並對蒐集之資料內容做出適當文字分類與文字處理，將向指導教授請教計畫欲使用之網路爬蟲、文章斷詞、去重複工具如 Jieba、CkipTagger 技術的使用方式，並和指導教授討論情緒分析（Sentiment Analysis）之理論與實際應用，以及如何對已分類的情緒字詞篩選、分級等，進而正確解釋社群網路評論與真實股市走勢之關係。

表 C802