國立臺北科技大學

資訊與財金管理系

學生畢業專題

研究社群媒體資料之金融市場情緒詞典—以台積電為例

學生：108AB0712 林姍如
108AB0727 黃雅婄
108AB0744 廖劭其
108AB0752 張倢菱

指導教授：劉亮志老師

中華民國一百一十二年一月

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），並統整成金融市場情緒字典，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、社群媒體文字探勘（Text Mining）、情緒分析（Sentiment Analysis）

誌謝

本專題能夠順利的完成，首先要感謝是我們的專題老師，劉亮志教授和技術指導鄭麗珍教授，劉老師和鄭老師在本專題的各個研究階段中都充分參與，無時無刻給我們寶貴的建議與指導，由於這不辭辛勞的的付出，本專題才能以較完整的風貌展現。還有學校所提供的場所，供我們做專題的場地在這段期間中竭誠的協助與辛勞的投入本專題的研究與討論，才使得本專題能夠如期的完成。

最後還要謝謝各位同學及組員們的配合，才能順利的完成專題製作，雖然還有很多要學習的地方，但還是非常感謝專題老師們還有各位組員，也讓我們知道「專題」是大家一起齊心努力的，沒有努力是沒有成果的。

目錄

表目錄

圖目錄

第一章 前言

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）、CkipTagger 斷詞等相關技術，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），將資料分類成正面和負面等情緒標籤，並統整成相關的情緒字典，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。

第二章 相關文獻

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、意見探勘、情緒分析（Sentiment Analysis）	統計分析（Statistical Analysis）預測模型（Predictive Model）

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]。

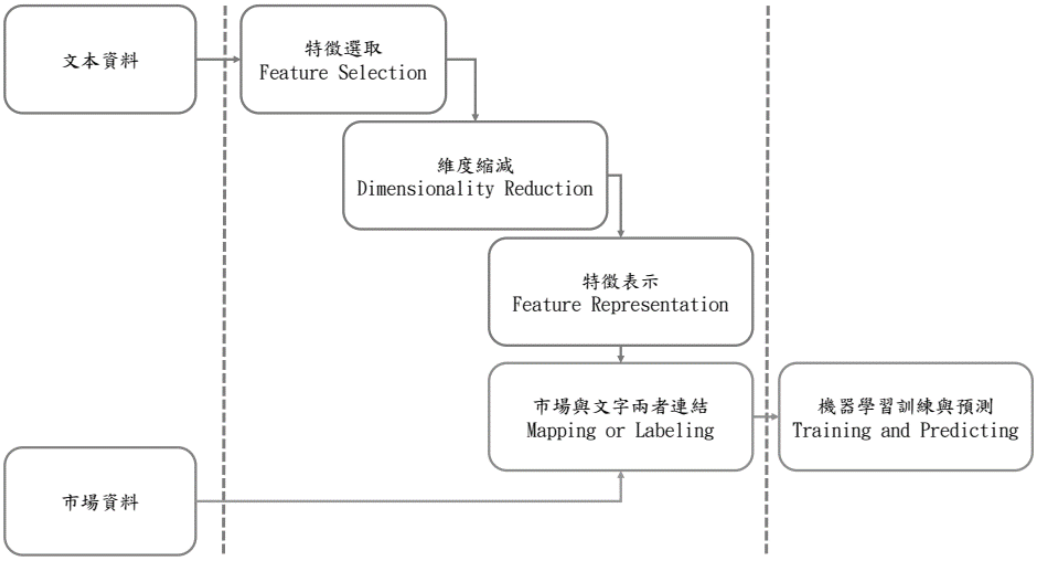

圖 1、文字探勘三階段

3. 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。

4. 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]

5. 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]

第三章 資料蒐集與處理

本研究利用 Python 的套件將「金融社群網站 CMoney」中台積電個版的資料透過爬蟲技術進行蒐集，**資料樣本區間為 2021 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

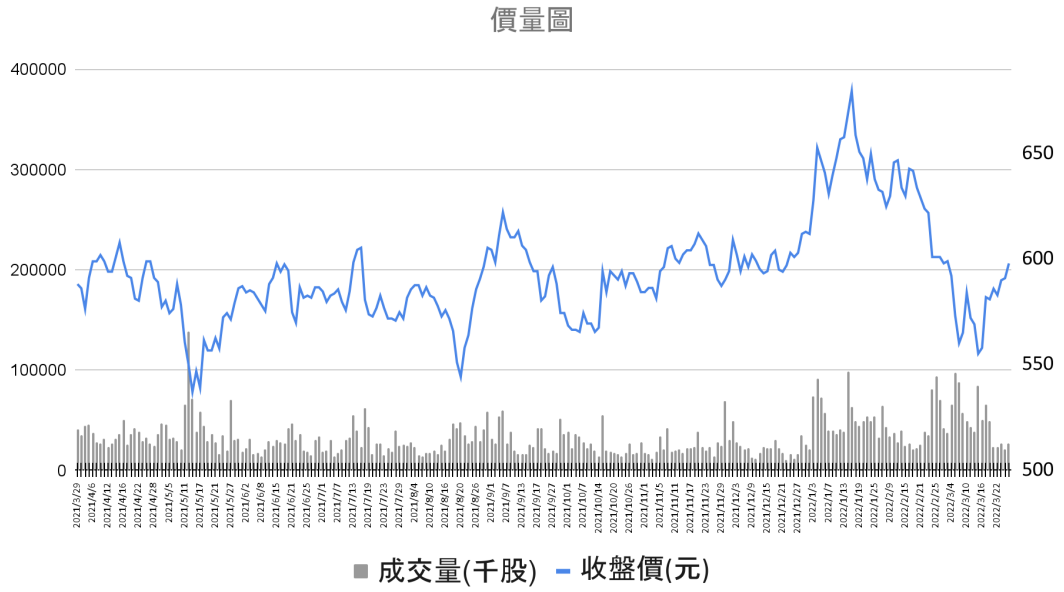

圖 2、樣本區間之交易日價量圖

1. 斷詞處理

本研究使用中央研究院中文詞知識庫小組的套件 CkipTagger（又稱 Ckip）來進行斷詞、詞性標註等前置處理。由全組組員訂立出分類邏輯確保判斷一致性，再製作成情緒字典，類別有正向詞及負向詞，目的是讓系統計算正負向字詞的數量，作為辨別文章情緒的依據。

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']

表 3、斷詞範例

2. 人工撿詞

（1）社群文章

本研究利用社群文章資料製作人工撿詞字典，使用者在社群中的發文內容不似較專業的報章雜誌，社群中的用詞大多都較口語化，同時充斥著大量的網路用語、流行用語，故需要以人工的方式將被斷開的字詞重新撿回，並給定 CkipTagger 新的人工撿詞字典。撿詞標準為全體組員（四人）及指導教授共同檢核，達成共識後再新增撿詞，其中撿詞也考慮了上下文影響，如拉/不拉、漲/不會漲、站穩/不會站穩等類似的上下文字詞，若被斷開則有不同含意，本研究也將這類的字詞重新撿回，社群文章的總撿詞為 135 筆。

社群文章人工撿詞字典							
135 筆							
不拉	業外	漲跌-	台積電+	百德集團	開低走低	資本利得	任重而道遠
小跌	護盤	滿水位	台積電-	借券賣出	開低走平	機會財股	股價淨值比

多單	續抱	隔日沖	加密貨幣	烏俄戰爭	開平走高	融券賣出	百元俱樂部
走跌	避險	會不會	月線往上	租賃三雄	開平走低	TSM+	景氣循環股
抗俄	不看漲	噴起來	月線向上	航運三雄	開平走平	TSM-	殖利率倒掛
站回	不看跌	潛力股	中興保全	逢低布局	跌不下去	IC 設計	費城半導體
被嘎	不看好	競爭者	中華電信	通貨膨脹	超額利潤	護國神山	TSM +
崩盤	不是賣	不要買進	再度站上	通用電器	買賣超-	櫃買市場	TSM -
做東	3 奈米	不要再追	回補空單	貨櫃三雄	統一投顧	獲利了結	聯電 ADR
追新	不會漲	大盤指數	波段交易	貨幣緊縮	程式交易	千元俱樂部	台積電 ADR
創惟	不會賣	大幅下跌	季線往上	被動元件	無法站穩	不會再創高	道瓊工業指數
買賣	同欣電	人道走廊	季線向上	現金水位	開低走高	未實現損益	S&P500
稅後	要不要	人工智慧	供需失調	清算價值	集中市場	外資提款機	富時 100 指數
新高	往上賣	28 奈米	亞果生醫	高速傳輸	輔助交易	台積電 +	費城半導體指數
億豐	開紅盤	不會上漲	見好就收	骨牌效應	漲跌 +	台積電 -	賠了夫人又折兵
齊漲	新藥股	不會升息	均線往上	追高殺低	漲跌 -	台股 ADR	NSDAQ 綜合指數
嘎盤	漲跌+	民意調查	均線向上	逢低回補	道瓊指數	元大高股息	

表 4、社群文章字典

（2）專業文章

除網路用語外，社群討論內容也會充斥著各種專業術語。本研究參考許多機構公開的產業報告書、半導體產業調查報告及多篇經濟期刊等專業文章，將台灣半導體產業中常見的專業字詞進行人工撿詞，將被斷開的字詞重新撿回。專業字詞的撿詞標準為全體組員（四人）共同檢核及指導教授共同檢核，達成共識後再新增撿詞，專業文章的總撿詞為 126 筆。

專業文章人工撿詞							
126 筆							
封城	醫療設備	每股盈餘	數位科技	世界先進	邏輯 IC	美中貿易戰	先進駕駛補助系統
新冠	新興技術	現金股利	總體經濟	戰略地位	車用晶片	美中科技戰	3D Fabric
遷廠	財務預測	領導地位	集成電路	供需失衡	量子電腦	中美科技戰	美國半導體工業協會
5G	高階主管	邏輯密度	地緣戰略	景氣循環	量子通訊	智慧型手機	美國半導體行業協會
缺貨	全球經濟	異質整合	新冠肺炎	晶圓代工	車用市場	筆記型電腦	美國半導體產業協會
遠端	遠距學習	矽中介層	先進封裝	成熟製程	垂直整合	報復性消費	雞蛋放在同一個籃子
2 奈米	地緣政治	利益平衡	去中心化	銷售折讓	商業模式	市場滲透率	世界半導體貿易統計組織
供應鏈	基礎架構	中國大陸	分散風險	銷貨折讓	半導體產業	車用半導體	FTSE4 Good
7 奈米	潛在市場	公衛機構	外銷導向	平板電腦	技術差異化	三維積體電路	Mitsubishi Electric
5 奈米	車用電子	再生能源	市場供需	單點失效	高效能運算	開放創新平台	Open Innovation Platform
4 奈米	系統整合	永續發展	消費電子	單點故障	全世代製程	道瓊永續指數	Semiconductor Industry Association
1 奈米	運算密度	公司治理	新興科技	基礎設施	營業利益率	極紫外光微影	World Semiconductor Trade Statistics
價值鏈	先進製程	經營管理	生產進程	產能過剩	稅後純益率	積體電路設計	
矽晶圓	合併營收	世界指數	網路攻擊	潛在風險	利害關係人	全球半導體聯盟	S&P Global
數位轉型	稅後淨利	封裝測試	出口管制	進口替代	台灣證交所	整合元件製造商	Robeco SAM
無線技術	晶片設計	比較利益	營運成本	出口替代	滾動式調整	COVID-19	

表 5、專業文章字典

3. 正負向字詞標記

本研究需對社群文章中的正負向字詞進行統計，以計算出社群輿論討論漲跌的情況。人工撿回被斷開的詞後，重新處理全年資料，產出正確的詞頻及斷詞，並以人工標記正負向字詞的方式製作出金融情緒字典。若此字詞對股市有正向敘述則判斷為正向詞、負向敘述則為負向詞。判斷標準為全體組員（四人）及指導教授共同檢核，達成共識後給定正負向情緒分類，情緒字典的正負向字詞總計為 157 筆。

正向詞字典				負向詞字典			
87 筆				70 筆			
攻	拉高	漲勢	買回來	弱	扼殺	跌幅	不要買進
抱	長多	漲價	漲上來	挫	走弱	落後	不會上漲
拉	看好	領先	漲停板	割	走跌	緊縮	台積電-
強	看漲	領漲	漲跌+	跌	軋空	賠錢	回補空單
買	突破	增長	噴起來	賣	看跌	賣出	忐忑不安
漲	留倉	樂觀	不會升息	下挫	重挫	賣掉	貨幣緊縮
上去	站上	獨強	月線向上	下殺	降溫	賣超	通貨膨脹
上揚	站回	獲利	月線往上	下跌	倒掛	賣壓	無法站穩
上漲	站穩	賺到	台積電+	下彎	除息	壓回	買賣超-
大單	做多	賺錢	再度站上	大跌	做空	縮表	漲跌 -
大買	強勁	擴產	均線向上	小跌	停利	虧損	獲利了結
大漲	強勢	翻紅	均線往上	不利	停損	轉弱	TSM-
止跌	創高	轉強	供不應求	不拉	通膨	不看好	不會再創高
牛市	復甦	續抱	季線向上	升息	減資	不看漲	台積電 -
可望	買超	續強	季線往上	出清	超賣	不會漲	殖利率倒掛
回補	買進	不是賣	跌不下去	多空	跌到	漲跌-	TSM -
多頭	新高	不看跌	開高走高	收黑	跌破	大幅下跌	

收紅	補回	不會賣	漲跌＋	利空	跌停	不要再追	
有望	補漲	邪惡波	TSM+				
利多	解封	往上賣	台積電＋				
抓住	漲停	拉上去	TSM＋				
抱緊	漲幅	起漲股					

表 6、正負向文章字典

4. 正向詞字典文字雲

用以上的正向詞情緒字典計算全年所有文章的詞頻。在文字雲中的字體越大表示出現在文章中的頻率次數較高。如**買、漲、上漲、新高、獲利**，在全年的所有文章中各出現超過 10000 次。

圖 3、正向詞字典文字雲

5. 負向詞字典文字雲

用以上的負向詞情緒字典計算全年所有文章的詞頻。在文字雲中的字體越大表示出現在文章中的頻率次數較高。如**下跌、跌、賣、賣超、升息**，在全年的所有文章中各出現超過 5000 次。

圖 4、負向詞字典文字雲

6. 名詞說明

名詞	說明
交易日	排除股市未開市天數，以了解股市開市時的社群型態。排除後的**資料天數為 244 天，總文章篇數 57798 篇**。
系統情緒	以「情緒字典」將**交易日的社群文章**進行正負向字詞統計，當日正向詞較多則系統情緒為正向，負向詞多過於正向詞則系統情緒判斷為負向，若統計結果正向詞和負向詞數量相等則為中立。
股價漲跌	股價漲跌幅情況。當日收盤價較昨日收盤價高則股價漲跌為正向，反之為負向，若該日收盤價和昨日收盤價相同，則股價漲跌為中立。

表 7、名詞說明

● **驗證結果**

將系統情緒及股價漲跌做比對，判斷社群文章與股價漲跌的相關性。
若系統情緒與股價漲跌相等，則驗證結果為O，共有以下3種情況：

1. 系統情緒為正向，股價漲跌為正向。

2. 系統情緒為負向，股價漲跌為負向。

3. 系統情緒為中立，股價漲跌為中立。

若非以上三種情況，則驗證結果皆為X。

本研究將**交易日的社群文章**進行資料處理，說明以下3種不同實驗方式，以調整因子及參數的方式，驗證社群文章與股價漲跌是否存在關聯。

（1）系統情緒_對照組

將**交易日的社群文章**進行系統情緒統計，**不做任何變數調整**。

（2）系統情緒_實驗組（篩選台積電）

對**交易日的社群文章**進行篩選，本研究將社群文章內有出現「**台積電、臺積電、護國神山、神G、台g、台G、TSMC、tsmc**」字詞之社群文章視作「**與台積電有關之社群文章**」，**以下以「篩選關聯性」為代稱**。

（3）系統情緒_實驗組（篩選台積電&等級57以上）

將交易日的社群文章篩選出與台積電有關之社群文章，再進行發文者等級篩選。本研究將發文者等級進行四分位距統計，**第三四分位數（Q3）為等級57等**。本研究欲分析**發文者等級較高者是否具較大影響力、同時驗證其發文內容與股價漲跌之關聯**，故排除發文者等級低於57等之社群文章，**以下以「篩選發文者等級」為代稱**。

第四章 發文內容分析

以下資料分析之統計資料以交易日的社群文章為樣本，資料樣本區間為 2021 年 3 月 27 日至 2022 年 3 月 27 日，**以股市交易日 244 天、總文章篇數 57798 篇**為內容進行深度分析。由於平台於交易日平均每小時發文量不到十篇，文章樣本數量不多，本研究按照目前平台使用者情況進行分析。

本研究針對社群文章內容、社群發文時段進一步分析，欲了解使用者在社群平台活動的時間、發文習慣等，並推論出此平台使用者的習性及文章內容與股價的關聯。

1. 系統情緒統計單位

本研究將交易日中所有文章資料進行分析，進行兩種不同的計算方式：

（1）系統情緒（各篇加總）

以一篇社群文章為一個情緒單位，對當日的社群文章進行個別社群文章的正負向字詞統計以計算個別社群文章情緒。統計當日社群文章的情緒值，若當日正向情緒的社群文章較多，則判斷當日系統情緒為正向；負向社群文章較多則判斷當日系統情緒為負向。

例如：當日有 200 篇正向社群文章及 100 篇負向社群文章，即當日系統情緒為正向。

（2）系統情緒（全天加總）

以一天為一個情緒單位，彙總當日社群文章之正負向字詞進行統計，計算單日的系統情緒。

例如：當日總計有 500 個正向字詞及 300 個負向字詞，即當日系統情緒為正向。

對照組			實驗組 （篩選台積電）			實驗組 （篩選台積電&等級 57 以上）		
系統情緒-各篇加總								
	當日	隔日		當日	隔日		當日	隔日
符合	125	105	符合	118	104	符合	122	109
不符合	119	139	不符合	126	140	不符合	122	135
總天數	244	244	總天數	244	244	總天數	244	244
符合率	51.23%	43.03%	符合率	48.36%	42.62%	符合率	50.00%	44.67%
系統情緒-全天加總								
	當日	隔日		當日	隔日		當日	隔日
符合	129	115	符合	130	114	符合	140	113
不符合	115	129	不符合	114	130	不符合	104	131
總天數	244	244	總天數	244	244	總天數	244	244
符合率	52.87%	47.13%	符合率	53.28%	46.72%	符合率	57.38%	46.31%

表 8、系統準確度分析表全天加總與各篇加總

根據上表可得知系統情緒在**以天為單位（全天加總）的準確度較高**，故以下分析將以**每日社群文章正負向字詞加總，以計算當日的系統情緒**。

2. 正負向天數/正負向字詞平均

將交易日的社群文章進行系統情緒的正負向天數統計，分別是：

實驗		正向	負向
對照組	天數	213	31
	佔比	87.30%	12.70%
篩選台積電	天數	212	31
	佔比	87.24%	12.76%
篩選台積電&等級 57 以上	天數	198	46
	佔比	81.15%	18.85%

表 9、正負向天數分析

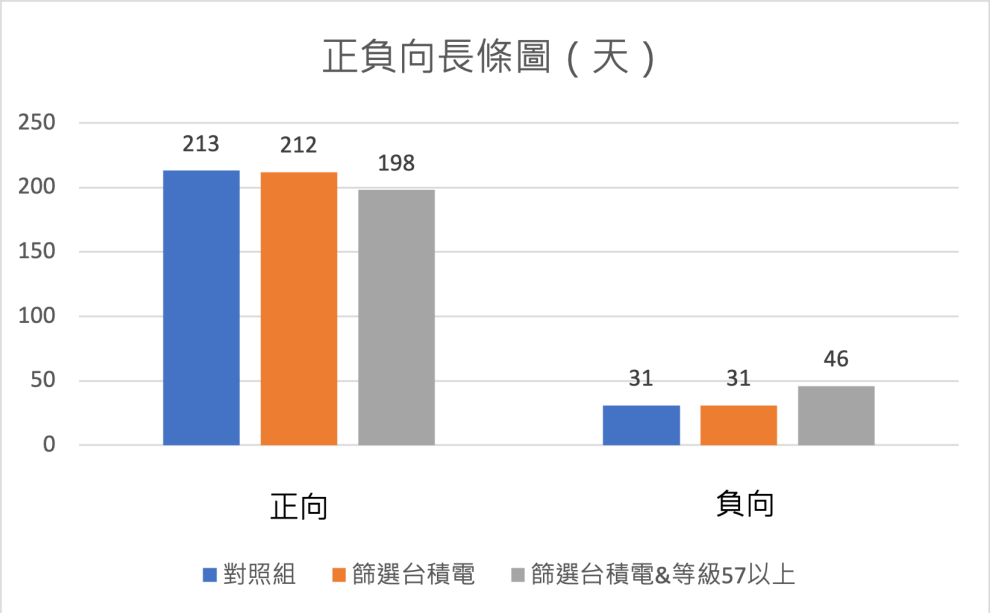

圖 5、正負向天數分析

由上述圖表可知，三種實驗結果的正向天數都明顯多過於負向天數，此社群平台的發文內容都相對正面，其中篩選台積電&等級 57 以上系統情緒判斷為負向的天數最多。

組別	發文量平均	正向詞平均	負向詞平均
對照組	237	642	344
篩選台積電	111	520	262
篩選台積電&等級 57 以上	28	197	103

表 10、發文量、正負向字詞平均

本研究統計了交易日（244 天）中所有文章的正負向字詞，計算出每日正向詞及負向詞平均數量。發現在三種實驗結果中，平均每日正向詞的數量會高於負向詞近兩倍或以上，也更加應證了平台中較多正面發文，使用者普遍看好股市的現象。

3. 各時段發文趨勢比較

本研究對資料進行發文時段的詳細分析，從股市開盤前一小時的上午 08：00 至收盤的下午 13：30 進行統計，每 30 分鐘為一區間，對**各時段發文量、正負向輿論比例**進行分析，分析包含**全年資料（366 天）、交易日資料（244 天）**。

（1）發文時段趨勢

本研究藉由觀察各發文時段中發文量的統計，推論出使用者活動最頻繁的時段，並分析發文量與股價漲跌之關聯。

（2）發文時段正負向比例趨勢

針對各時段社群輿論內容的正負向文章統計，觀察各時段中使用者發文情緒的變化。

發文時段趨勢	
對照組	實驗組（篩選台積電&等級 57 以上）
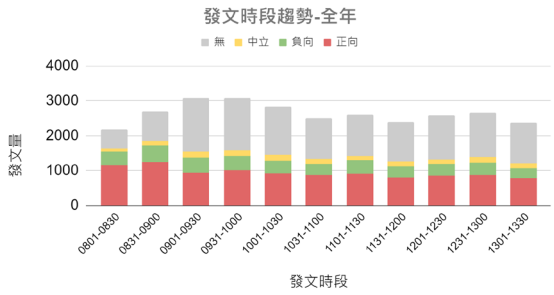	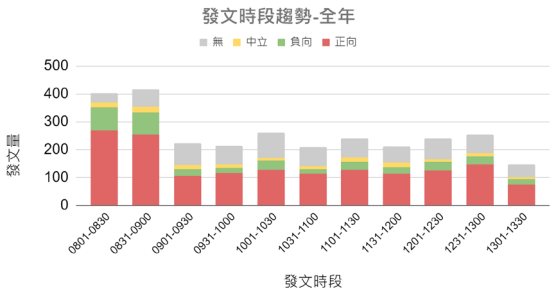
圖 6、對照組發文時段趨勢-全年	圖 7、實驗組發文時段趨勢-全年
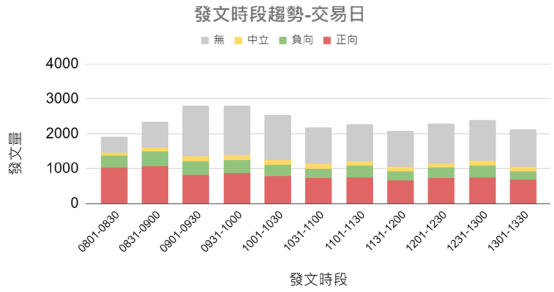	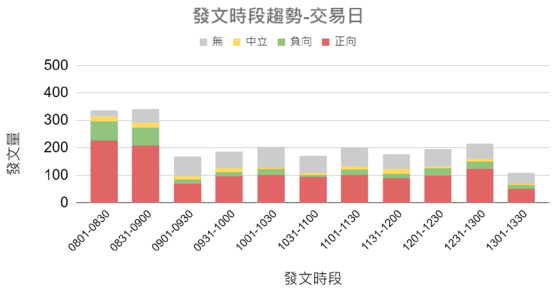
圖 8、對照組發文時段趨勢-交易日	圖 9、實驗組發文時段趨勢-交易日

表 11、發文時段趨勢

發文時段正負向比例趨勢	
對照組	實驗組（篩選台積電&等級 57 以上）
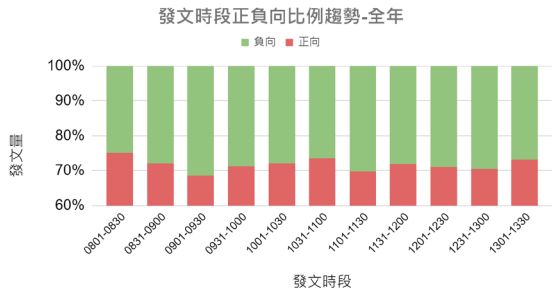	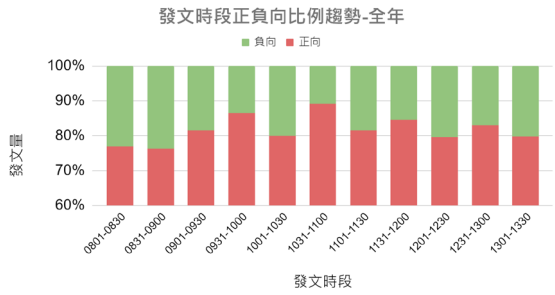
圖 10、對照組發文時段正負向比例趨勢-	圖 11、實驗組發文時段正負向比例趨勢-全

全年	年
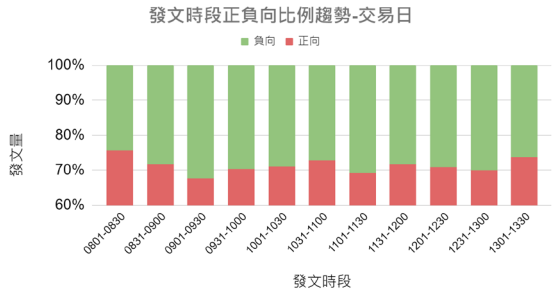	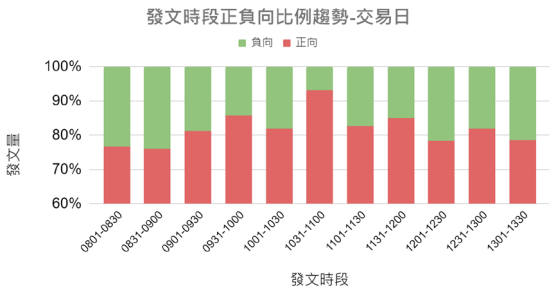
圖 12、對照組發文時段正負向比例趨勢-交易日	圖 13、實驗組發文時段正負向比例趨勢-交易日

表 12、發文時段正負向比例趨勢

藉由圖表可得知：

1. 藉由圖 6、圖 7 可知，篩選條件增加，無情緒的文章數量越少。
2. 圖 7、圖 9 顯示，等級 57 等以上的使用者發文時段較集中在開盤前一小時，也就是上午 08：00 至 09：00 之間。
   而圖 6、圖 8，無篩選的對照組，其發文趨勢就相對平穩，但同樣在上午的發文量較多，使用者在 09：00 至 10：00 之間有比較密集的發文量。
3. 觀察圖 10、圖 11，與無篩選相比，等級 57 等以上使用者的發文內容中，整體正向文章的比例高出不少，特別是實驗組在全年的資料中，上午 10：30 至 11：00 這半小時的正向文章比例更接近 90%。
4. 對比圖 10 到圖 13，可發現全年與交易日的發文型態無論在發文時段趨勢或是正負向文章比例都極相似。

第五章 系統準確度分析

本研究透過三個階段的確認，觀察在**股價驗證時間、社群文章關聯性、發文者等級的因子**下系統的準確度變化，以找出系統準確度最高的情況：

針對**交易日資料（244 天）**進行系統情緒判斷，並與**當日及隔日股價**情緒做比對。

1. 第一階段：股價驗證的時間

分別以驗證正確、驗證錯誤的結果呈現，如下圖：

（1）當日股價

當日系統情緒對照當日股價，以驗證當日社群言論是否與當日股價較有關聯。

舉例：12/27 系統情緒為正向，股價漲跌也為正向，則 12/27 驗證正確。

（2）隔日股價

當日系統情緒對照昨日股價，以驗證昨日股價與今日的社群言論是否較有關聯。

舉例：12/27 系統情緒為正向，12/26 股價漲跌也為正向，則 12/27 驗證正確。

股價驗證時間	正確率	錯誤率
系統情緒與**當日股價**比對	52.87%	47.13%
系統情緒與**隔日股價**比對	47.13%	52.87%

表 13、股價驗證時間

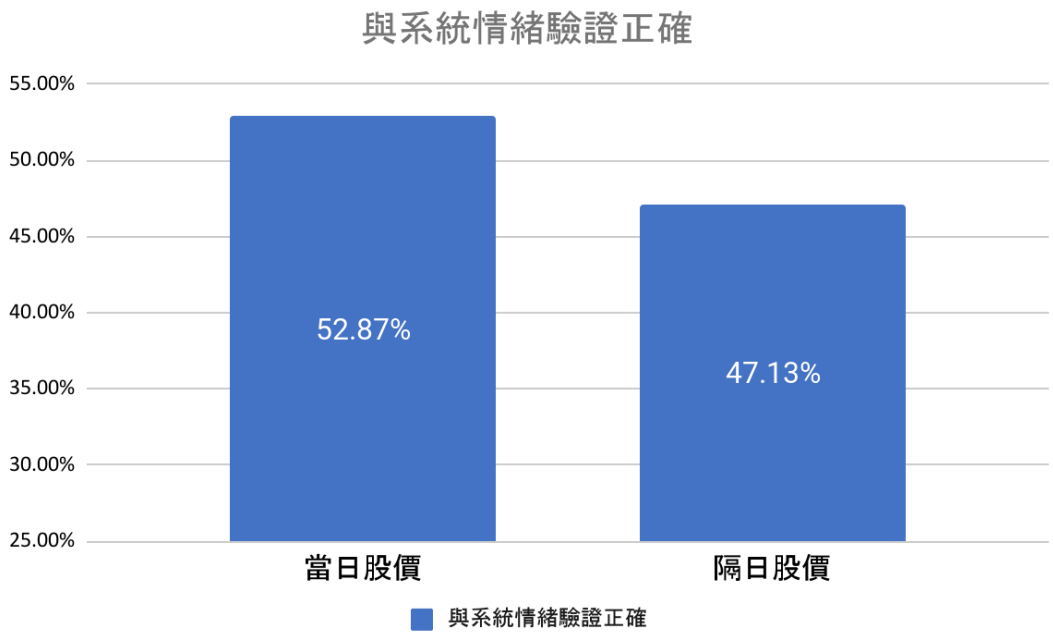

圖 14、系統情緒驗證正確比例

將當日及隔日股價情緒與系統情緒做比較，由統計結果可得知**以當日股價做比對的精準度較高（52.87% > 47.13%）**，故作為實驗的基準點。

（3）驗證正確比例分析

第一階段 對照組與當日股價情緒分析				
組別	驗證	總佔比	正負向	佔比
對照組	正確	52.87%	正向	41.80%
			負向	11.07%
	錯誤	47.13%	正向	45.49%
			負向	1.64%

表 14、第一階段驗證分析

正向天數比例	負向天數比例
87.30%	12.70%
系統情緒正向且與股價情緒比對正確	系統情緒負向且與股價情緒比對正確
47.89%	87.10%

表 15、第一階段正負向天數及驗證分析

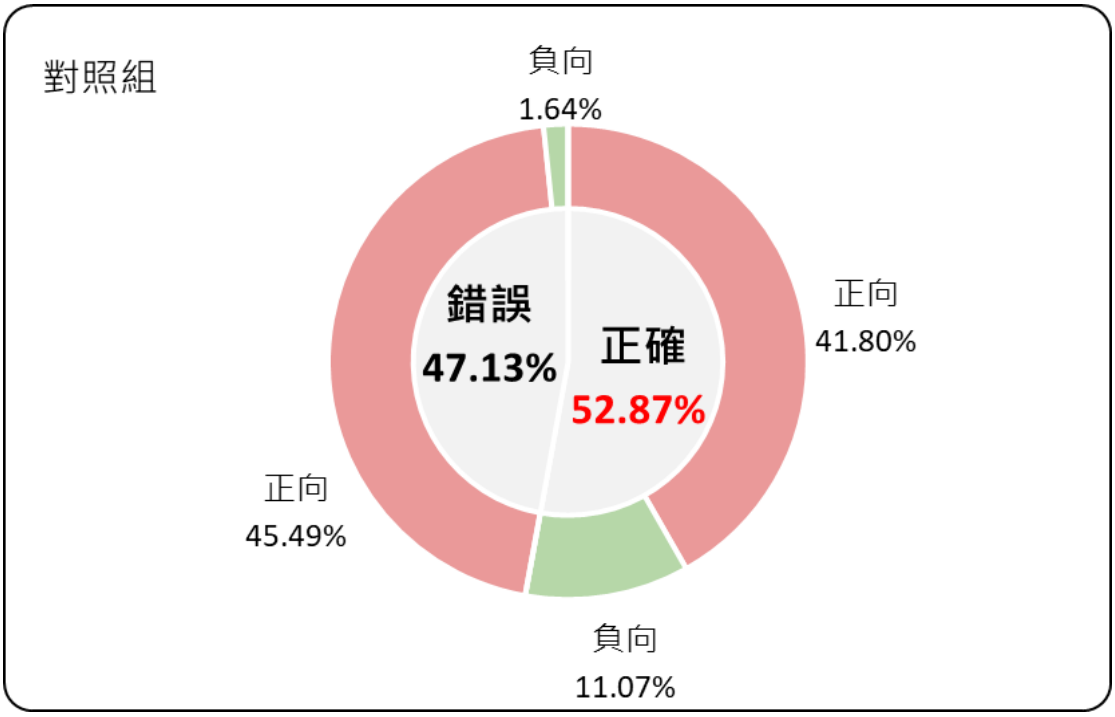

圖 15、第一階段驗證分析

第一階段以當日股價為驗證時間基準，在對照組中：

在系統驗證正確的數據中，情緒正向比例為 41.80%、負向比例為 11.07%，整體系統情緒為正向的比例為 87.3%，負向比例為 12.7%。並針對驗證正確的社群文章進行詳細分析，可發現：

所有正向社群文章中，將有 41.8% / 87.3%= **47.89%** 與股價驗證正確

所有負向社群文章中，將有 11.07% /12.7%= **87.1%** 與股價驗證正確

可得知正向文章的數量雖較多，但經過個別分析後，當日社群情緒為**負向**時，**系統驗證的準確度較高**。

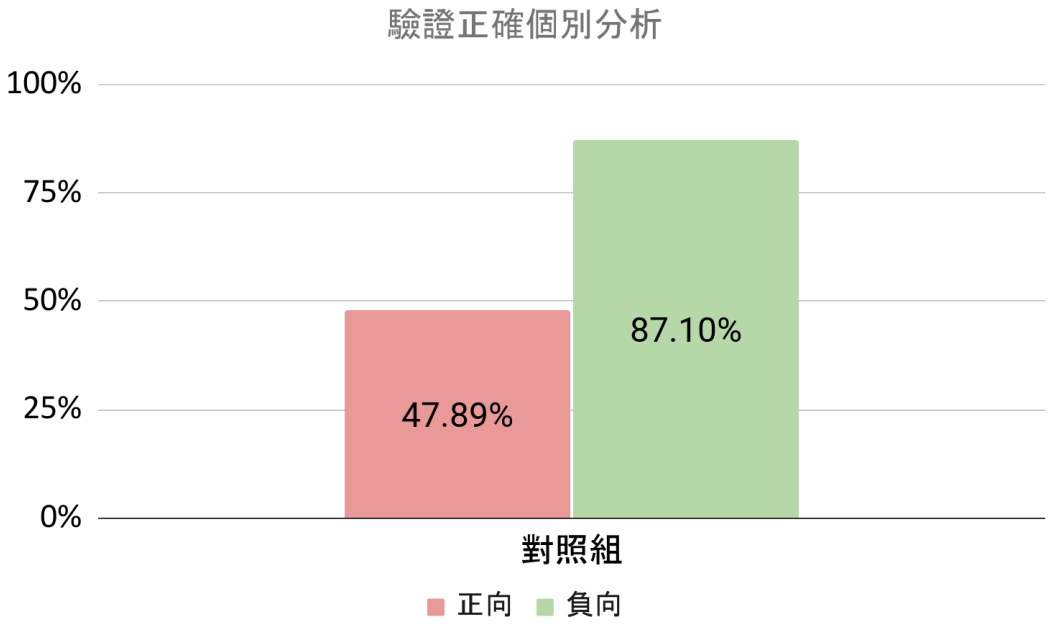

圖 16、第一階段驗證正確個別分析

2. 第二階段：社群文章關聯性

確定股價驗證時間後，進行第二階段的調整，本研究將文本**篩選有關台積電的文章**進行系統情緒判斷，再與當日股價比對結果區分為正確跟錯誤：

（1）驗證正確比例分析

第二階段				
組別	驗證	總佔比	正負向	佔比
實驗組（篩選台積電）	正確	53.28%	正向	42.21%
			中立	0%
			負向	11.07%
	錯誤	46.72%	正向	44.67%
			中立	0.41%
			負向	1.64%

表 16、第二階段驗證分析

正向天數比例	負向天數比例
86.88%	12.71%
系統情緒正向且與股價情緒比對正確	系統情緒負向且與股價情緒比對正確
48.58%	87.10%

表 17、第二階段正負向天數及驗證分析

將對照組與篩選台積電的分析進行比對，可得知**篩選文章關聯性的整體準確度從 52.87%提升至 53.28%**，表示**篩選台積電有助於判斷股市的情緒**，並以篩選台積電的資料往下進行實驗分析。

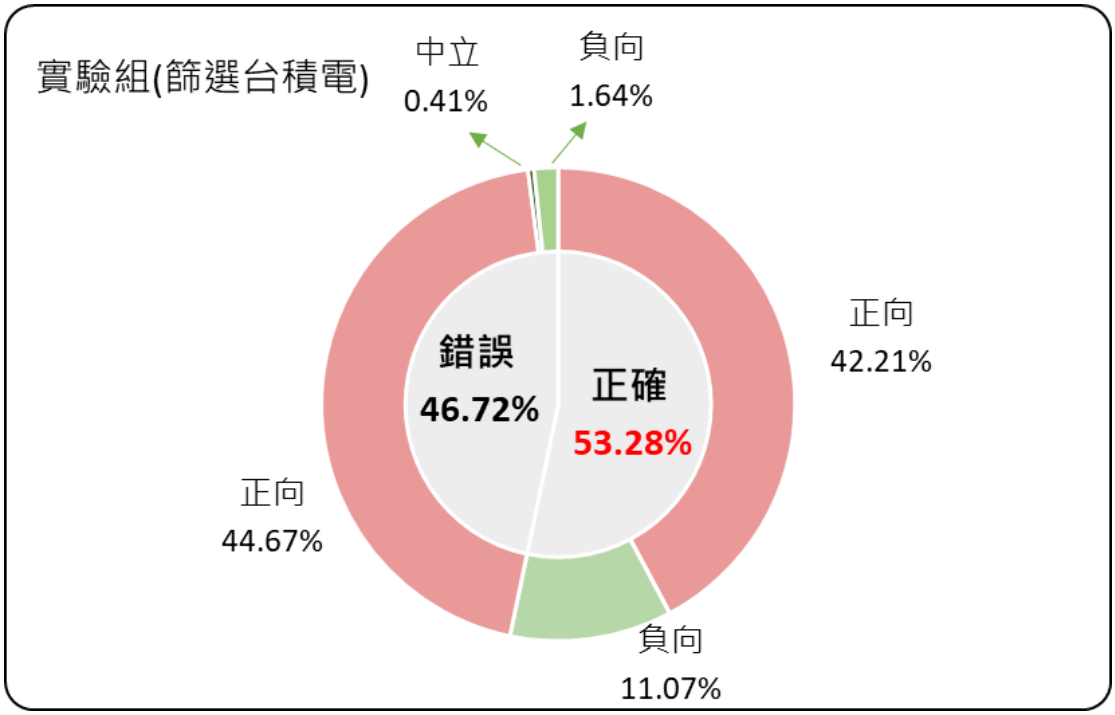

圖 17、第二階段驗證分析

第二階段以當日股價為驗證時間基準，篩選文章關聯性中：

在系統驗證正確的數據中，情緒正向比例為 42.21%、負向比例為 11.07%，整體系統情緒為正向的比例為 86.88%，負向比例為 12.71%。並針對驗證正確的社群文章進行詳細分析，可發現：

所有正向社群文章中，將有 42.21% / 86.88% = **48.58%** 與股價驗證正確

所有負向社群文章中，將有 11.07% /12.7%= **87.1%** 與股價驗證正確

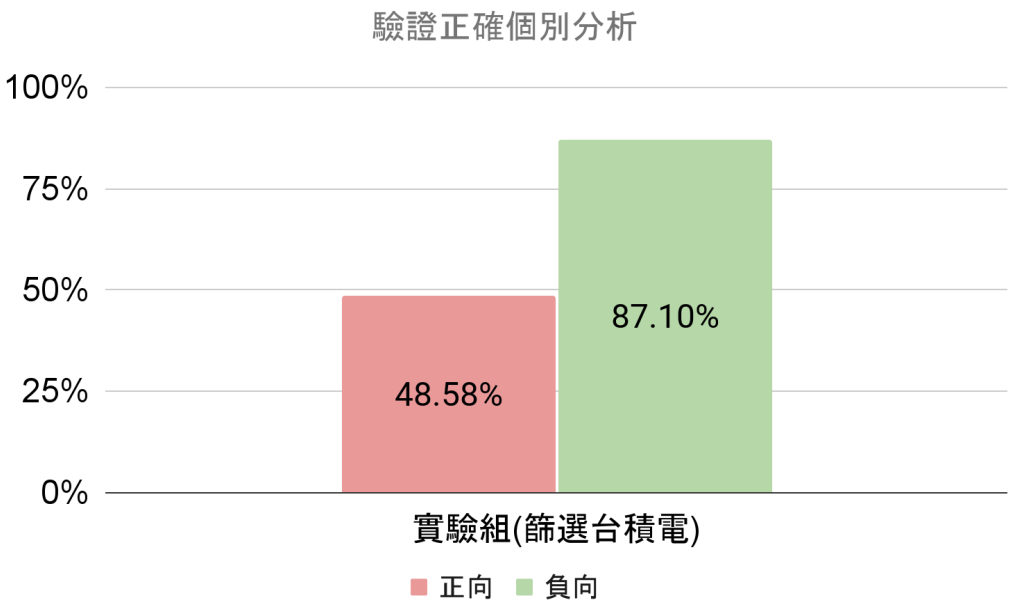

圖 18、第二階段驗證正確個別分析

3. 第三階段：使用者發文等級影響

第二階段與第三階段**差異因子在於篩選發文等級 57 等以上使用者**，實驗結果如分析表，**系統情緒的準確度從 53.28%提升至 57.37%**，表示**篩選台積電與發文者等級有利於本情緒字典的判斷**。

（1）驗證正確比例分析

第三階段				
組別	驗證率	驗證正確率	正負向	**佔比**
實驗組（篩選台積電&等級 57 以上）	正確	57.37%	正向	41.80%
			負向	15.57%
	錯誤	42.63%	正向	39.35%
			負向	3.28%

表 18、第三階段驗證分析

正向天數比例	負向天數比例
81.15%	18.85%
系統情緒正向且與股價情緒比對正確	系統情緒負向且與股價情緒比對正確
51.51%	82.6%

表 19、第三階段正負向天數及驗證分析

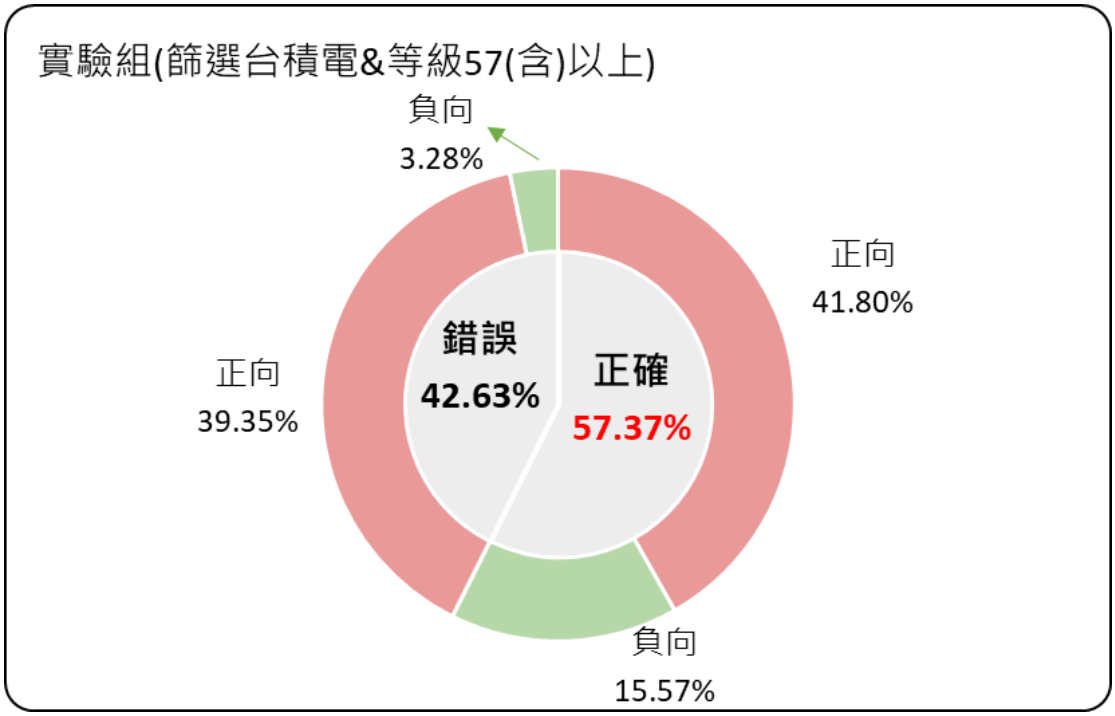

圖 19、第三階段驗證分析

第三階段以當日股價為驗證時間基準，篩選文章關聯性、發文者等級中：

在系統驗證正確的數據中，情緒正向比例為 41.80%、負向比例為 15.57%，整體系統情緒為正向的比例為 81.15%，負向比例為 18.85%。並針對驗證正確的社群文章進行詳細分析，可發現：

所有正向社群文章中，將有 41.80% / 81.15% = **51.51%** 與股價驗證正確

所有負向社群文章中，將有 15.57% / 18.85% = **82.60%** 與股價驗證正確

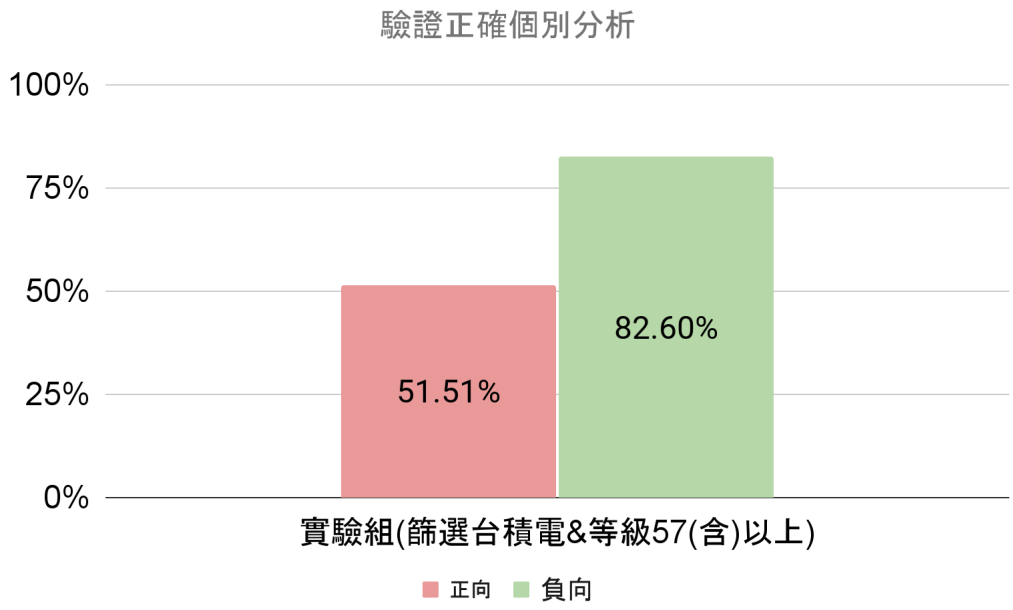

圖 20、第三階段驗證正確個別分析

（2）整體正負向驗證正確比例分析

	第一階段		第二階段		第三階段	
正確率	正向	負向	正向	負向	正向	負向
	47.89%	87.10%	48.58%	87.10%	51.51%	82.6%

表 20、各階段驗證分析

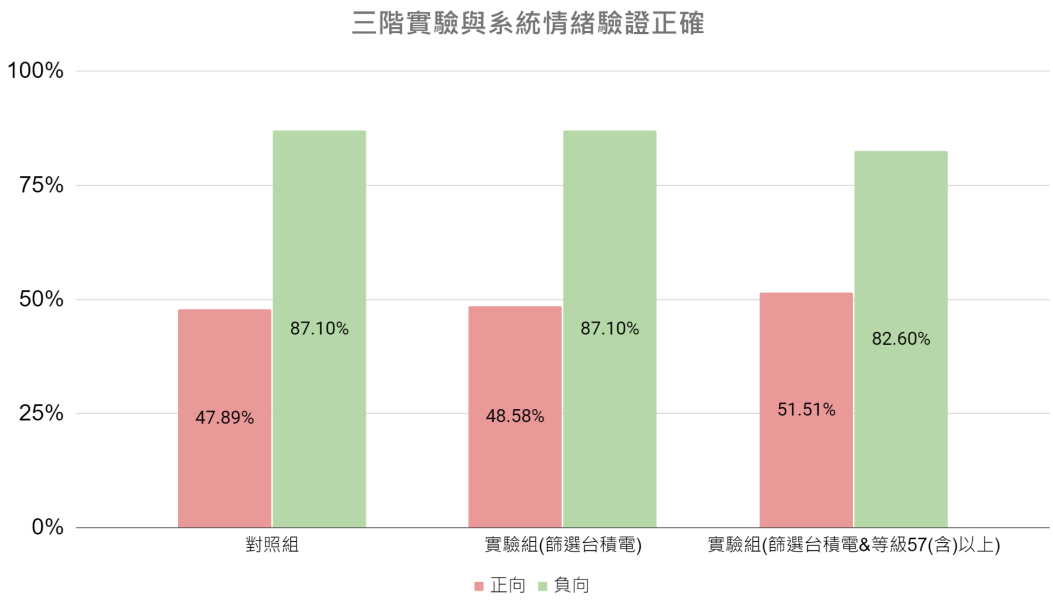

圖 21、各階段驗證分析

將三個階段的正負向情緒準確度放在一起比較，可以發現三個階段中系統情緒被判斷為負向的準確度都明顯高於正向，**可見我們系統及字典驗證出的負向情緒對於股市判斷較有參考價值。**

第六章 計畫成果自評

本計畫執行的進度與結果符合原定之計畫規劃，製作完成金融市場情緒詞典，由於在驗證準確度中發現下列可進行的優化，故列點提出以下的建議及未來展望：

1. 斷詞斷句與製作字典

中文因其文法複雜，伴隨著標點符號的交錯影響，如何讓系統精準的斷詞仍是比較大的挑戰，若能深入鑽研斷詞斷句的模式，讓系統在處理文章時提昇斷詞精準度，並且在製作字典時加上每個正負向字詞的權重，則能提升系統情緒的準確度。

2. 給定正負向字詞動態權重

除了上述提到可在字典上替正負向字詞人工加上權重外，也可利用程式自動判斷正負向字詞間的相似性、向量距離等。動態給定該正負向字詞在該句中應有的權重，使得各正負向字詞在各句中所佔的權重能夠獲得正確的情緒比重。

3. 結合機器學習

研究中資料處理如正負向字詞判斷、撿詞等較多為人工製作。本研究認為目前市面上尚未有中文情緒分析的股市系統，未來可以藉由目前製作的社群情緒字典，運用機器學習技術發展系統自動化的社群情緒分析，且加入機器學習能更有效率地進行撿詞、正負向字詞標示。

4. 尋找平均發文量較多的平台

因本研究發現 CMoney 平台的發文量偏低（交易日平均每小時不到十篇），導致出現單一數據影響整體分析結果的情況，若是尋找平均發文量較多的社群股市平台，較不會因單一用戶使用習慣，而導致研究結果較為偏差。

5. 建議擴大 CMoney 資料樣本區間

若資料樣本增加或是擴大研究區間，即可針對目前研究樣本擴大，執行更詳細驗證，提升本研究的公信力。

第七章 結論

1. 發文內容分析

（1）論壇使用者的發文情緒偏正面

無論當天股價是漲是跌，社群中的情緒大部分都是看好股市。

經統計發現，社群中情緒為正向的天數高於負向天數，以實驗組（篩選台積電&等級 57 以上）為例，正向天數約占全部天數的 81.15%。且在正負向數量的部分，各項實驗每日出現在社群中的正向詞都會高於負向詞平均，正向詞的數量會高於負向詞近兩倍或以上。

（2）平台使用者在上午的活動較活躍

分析所有等級的使用者習慣，每日上午的 08：00 至 10：00 之間的發文量最多，其中等級 57 等以上發文者則集中在上午 08：00 至 09：00 發文較活躍，過了這個時段後等級 57 等以上發文者在平台發文量就會急速減少，討論度就接近穩定。

（3）等級 57 等以上發文者文章情緒較明顯

藉由篩選條件增加，無情緒的文章數量越少。

與對照組相比，等級 57 等以上發文者的發文內容中，整體正向文章的比例高出不少，特別是實驗組在全年的資料中，上午 10：30 至 11：00 這半小時正向文章比例更接近 90%。

（4）非交易日的發文量較少

因非交易日的發文量較少，對於系統情緒較無太大的影響，故全年與交易日的發文型態無論在發文時段趨勢或是正負向文章比例都極相似。

2. 系統準確度分析

（1）社群使用者的發文內容情緒，和當日股價較有關聯。

相較於隔日情緒，社群情緒與當日股價相比，準確率較高。

（2）情緒判斷為負向的驗證正確之比率皆較高、錯誤率較低。

故本字典作為系統情緒的判斷，驗證出的負向情緒對於股市判斷較有參考價值。這個結果也符合大部份財務文獻的既有結果。

（3）同時篩選關聯性及發文者等級的驗證結果準確度較高。

故等級 57 等以上發文者並提及台積電的文章，對於本研究較有參考價值。

參考文獻

[1]顏志杰, "遺傳演算法在股票投資組合風險值模型建構之應用," 天主教輔仁大學資訊管理學系在職專班碩士論文,2005。

[2]吳聲昌, "以資料探勘技術於台灣股票市場尋找低風險投資組合之研究," 世新大學管理學院資訊管理學系碩士學位論文,2006。

[3]陳育季, "重大新聞訊息對台灣股票市場影響之研究以經濟日報頭版新聞為例," 南華大學管理經濟學系經濟學碩士班碩士論文,2009。

[4]吳真慧, "專業性報紙頭版新聞對股票價量的影響," 逢甲大學企業管理學系碩士論文,2000。

[5]吳昀錚, "利用文字探勘技術預測台股加權指數," 國立中央大學資訊管理研究所碩士論文,2008。

[6]陳志龍, "運用類神經網路與技術指標預測股票型基金漲跌及交易時機之研究," 朝陽科技大學資訊管理系碩士論文,2006。

[7]陳俊翰, "人工智慧方法應用於臺灣股票指數期貨隔日漲跌預測之研究," 國立高雄第一科技大學資訊管理系碩士論文,2009。

[8]林姿吟, "財經新聞語料探勘在摩根台指成份股漲跌差異訊息之研究," 銘傳大學財務金融學系碩士論文,2007。

[9]L. Cao, "Domain Driven Data Mining （D3M）," IEEE International Conference on Data Mining Workshops（ICDMW'08）, pp.74-76, 2008.

[10] L. Cao, "Domain-Driven Data Mining： Challenges and Prospects," IEEE Transactions on Knowledge and Data Engineering, Vol. 22 , No. 6, pp.755-769, 2010.

[11] L. Cao, Y. Zhao, H. Zhang, D. Luo, C. Zhang, and E.K. Park, "Flexible Frameworks for Actionable Knowledge Discovery," IEEE Transactions on Knowledge and Data Engineering, Vol. 22 , No. 9, pp.1299-1312, 2010.