

計畫內容

1.[選題依據] 國內外相關研究的學術史梳理及研究動態；本課題相對於已有研究的獨到學術價值和應用價值等。

1.1 國內外相關研究的學術史梳理及研究動態

過去許多文獻已實證發現投資人的情緒會顯著影響股票市場報酬率。Baker and Wurgler (2006)檢測投資人情緒的橫斷面影響，結果顯示當情緒較低(高)時，對小股、年輕股、高波動股及貧困股的後續報酬會相當高(低)。Dash and Maitra (2018)使用回歸及小波模型檢測印度股市投資人內含情緒指數及價值加權股市指數之間的關係，結果呈現投資人情緒對股票報酬存在強烈正向的影響。這些文獻結果多顯示投資人情緒和當期股票報酬有正向關係，而對預期股票報酬有負向關係。進而 Sun, Najand and Shen (2016)探討由新聞媒體計算的高頻投資人情緒和股票報酬之間的預期關係，結果發現甚至落後報酬存在之下，落後半小時的投資人情緒會最優勢的預測日內 S&P500 指數報酬。Renault (2017)使用一個大眾於股市的微博平臺，建構由線上大眾於分享股市多頭或空頭看法時使用的字詞並形成情緒權重，發現前半小時投資人情緒變動可預測後半小時 S&P500 指數 ETF 報酬。

過去文獻上使用投資人情緒來預測股票報酬率多使用三種方法來衡量投資人情緒。第一種為使用“調查基礎”的投資人情緒指數衡量(Brown and Cliff, 2005)，如 AA II 投資人情緒調查。第二種為“市場基礎”的變數。Baker and Wurgler (2006; 2007)使用如 IPO 第一天報酬、IPO 量及交易量以衡量投資人情緒。Brown and Cliff (2004), Smales (2017) and Dash and Maitra (2018) 使用賣權相對買權數目的比率、市場周轉率相對股市總市值比率、隱含波動度指數 (VIX) 以衡量投資人情緒。第三種為使用“傳統媒體新聞基礎”的投資人情緒衡量(Tetlock, 2007; Sun et al., 2016)。Da et al. (2015)提及當調查無反應率高時，調查基礎衡量變成不可靠；市場基礎衡量可能夾雜經濟推動的結果。傳統媒體新聞基礎通常不是來自單一來源(Tetlock, 2007)，就是混雜股市之外來源的財務情緒資料形成的指數(Sun et al., 2016)；前者獲得的情緒衡量不夠客觀，後者則可能捕捉投資人於股市以外的情緒。

近年來行為財務學者已構建來自網路資料的投資人情緒指標。Renault (2017)提及來自網路的投資人情緒資訊應可克服 1. 來自於調查基礎衡量可能夾雜和答案相關的偏差，2. 來自於市場基礎衡量可能有特質及非情緒相關的成分，與 3. 來自於傳統媒體新聞基礎可能混雜因果關係。因為網路溝通的持續變動特質，新聞雜訊的問題可被頻繁地重新評估。近年一些文獻已使用來自網路的資訊以預測股票市場報酬，但他們的網路資訊來源多來自於如 StockTwits 的單一網路平臺(Da et al., 2015; Chen et al., 2014; Renault, 2017) 有可能搜集到特定客群的情緒反應，需要獲得來自不同投資客群的情緒反應才能結合社會聆聽。並且目前文獻未能直接捕捉投資人對某檔股票的情緒，僅能捕捉投資人對整體股市看多或看空的時間序列趨勢（如 Renault, 2017）。若能掌握投資人對個股樂觀及悲觀的情緒，可落實擇股及擇時的投資決策。同時，若能透過網路資訊進一步掌握投資人對個股樂觀及悲觀情緒的強度，更能應用投資人對個股樂觀及悲觀情緒的網路聲量以預測對股票報酬率的影響。

我國資本市場自 2014 年起陸續推出大資料指數基金，指數選股標準除了市值、盈餘等傳統因數外，還加入了在搜尋引擎的訪問次數及社群媒體的評論正負項得分等大數據因數。隨著相關技術的成熟，證券產業也開始出現應用大資料分析的案例，例如 2010 年美國印第安那大學的 Johan Bollen 與 Huina Mao 等人利用分析 Twitter 發文的情緒，可成功預測 87.6%的美國道瓊工業指數漲跌趨勢。2013 年英國華威商學院的 Tobias Preis 等人發現利用分析 Google 的關鍵字搜尋次數，可以預測股市走向。國外市場上亦已有業者推出即

時分析網路文章、社群媒體、新聞等訊息以做為投資參考的服務，例如美國 The Stock Sonar 及我國的情緒寶等。這些文獻證實結合金融科技及社群聆聽方式可影響並預測股票市場報酬率。但是應用金融科技捕捉對股票報酬影響的文章，尚缺乏從投資人對個股情緒的面向深入分析對股票價量的影響。相關文獻需進一步利用金融科技直接衡量投資人對個股正面及負面情緒，並將正負面情緒置入模型中分析其對股票報酬的影響。因此本研究進而從金融科技視角結合社群聆聽，來探討投資人樂觀及悲觀情緒是否及如何影響進而預測股票報酬率。

1.2 本課題相對於已有研究的獨到學術價值和應用價值

本課題通過網路爬蟲技術爬取各網路社群平臺，直接捕捉投資人對某檔股票漲跌看法的資訊，形成樂觀及悲觀情緒的網路聲量，並將正負面情緒的強度勾稽至金融大資料平臺，探討正負面情緒強度對股價報酬的影響進而預測報酬。

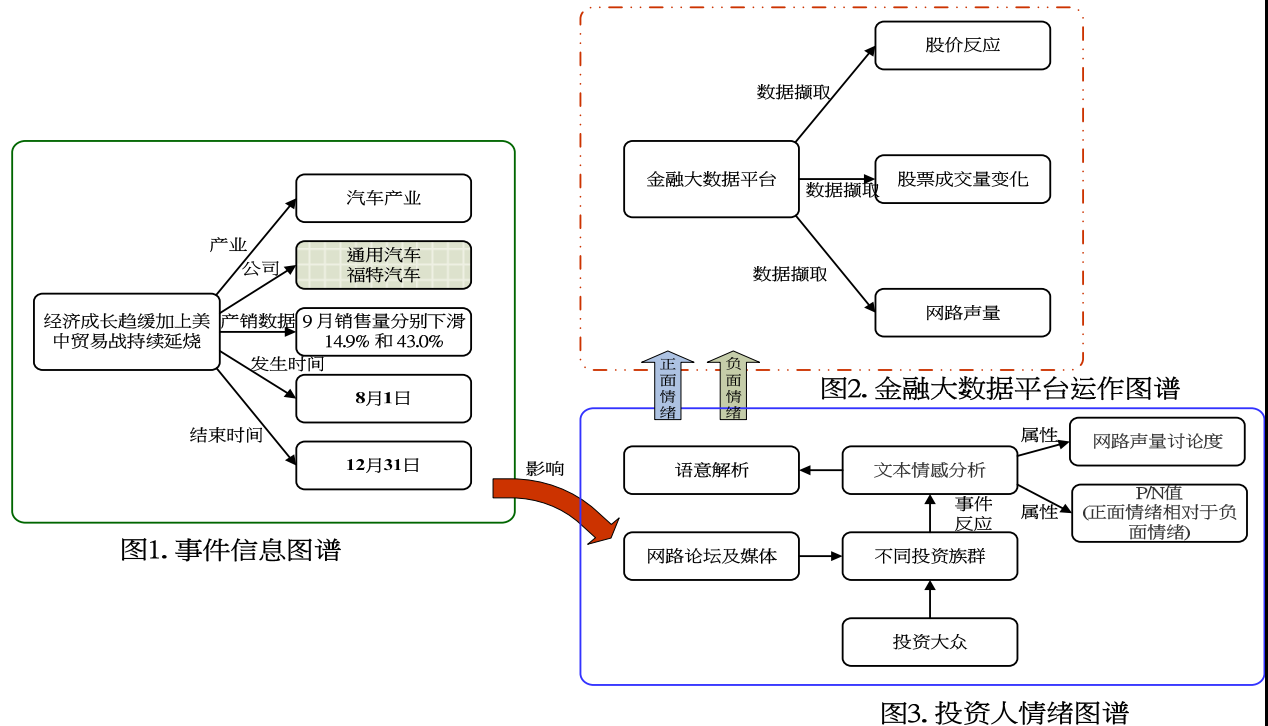
2.[研究內容] 本課題的研究物件、總體框架、重點難點、主要目標等。

2.1 本課題的研究物件

本計畫以上海及深圳股市投資人為研究物件，搜集近十年來這些投資人每日在這兩個股市所有相關社群網站對分別 ETF 個股發表的風險偏好、預期報酬、觀點及看法的充分資訊，紀錄日期、內容、個股、關鍵字，並由內容及關鍵字初步篩選為樂觀或悲觀情緒。

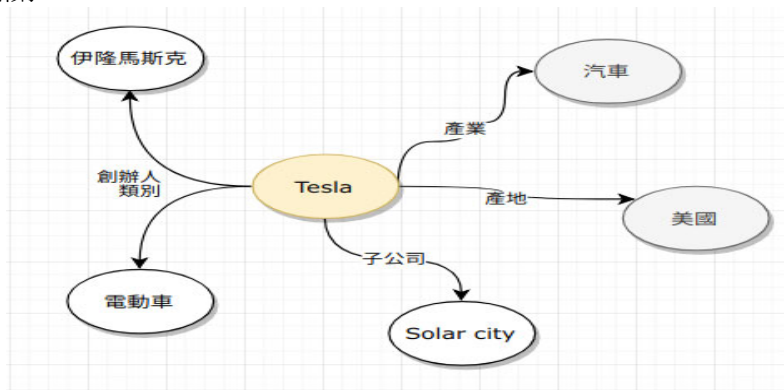
2.2 本課題的總體框架

當影響個股公司或股市大盤的事件發生時(圖 1)，網路各社群平臺討論資訊產生，本課題透過網路爬蟲程式密集採集投資人對個股漲跌看法，結合社群聆聽以語意解析來解讀投資大眾對公司發展及公司股價漲跌的評價。並採用「文本情感分析」來建構正面及負面情緒的網路聲量；設立投資人個股情緒的資料庫，並利用其來構建樂觀及悲觀情緒變數(圖 2)。進而將情緒變數勾稽金融大資料平臺(圖 3)以建置模型，分析樂觀及悲觀情緒對股票報酬影響及對後續股票報酬做預測。本計畫將事件訊息圖譜、投資人情緒圖譜、金融大資料平臺運作圖譜結合為知識圖譜。以下繪圖說明之：



(1) 事件資訊圖譜。主要提供機器人對於現實世界當中對於金融相關事件的理解。

新聞摘錄：中國將對美國 600 億美元商品課征 5~25%的關稅，商品包含太陽能電池、汽車零元件、家電商品等，去年美國出口到中國的商品總金額為 1,300 億美元。可發現對於 Tesla 將會有重大的影響，因其屬於汽車產業、美國公司、且子公司的業務亦為太陽能電池相關。



(2) 投資人情緒圖譜。以下為情感分析範例：

評論	正負
此事件將使得 Tesla 公司受到嚴重衝擊阿！	-0.9
不可能真的課這麼重的關稅，美國政府知道嚴重性，過一陣子又會變卦了	0.6
這事件對電動汽車產業，尤其是對 Tesla，或多或少都一些不利的影響	-0.1

(3) 金融大資料平臺運作圖譜。將投資人情緒圖譜所得到的正面及負面情緒匯入金融大資料平臺，並予對應的相關事件所牽連的產業及個股進行交叉分析，探討其對產業或個股股價及交易量的衝擊程度。



2.3 擬突破的重點和難點

(1) 透過網路爬蟲的程式來密集採集投資人對個股樂觀及悲觀情緒相關的資訊，需要一定人力、時間及經費投入來建置金融大資料平臺，且也需要運用知識圖譜來串接網路爬蟲程式。(2) 語意解析初步分辨正面及負面情緒的語意資料須查詢相關機構所建置的語意大資料資料庫，須投入相對的人力及時間。(3) 本計畫執行方式與過去的「量化投資」之主要差別在於運用機器學習技術，結合知識圖譜來搜集投資人對個股樂觀及悲觀情緒資

訊，並串聯事件發生對股價影響，為量化投資朝更精細化之技術。

2.4 本課題的主要目標

本計畫的研究目標想透過網路爬蟲的技術結合股市各投資客群的社群聆聽，勾稽投資人語意解析與財金訊息，建構投資人對個股正面及負面情緒的網路聲量，並建置模型分析樂觀及悲觀情緒對股票報酬率影響及對股票報酬率做預測。

3.[思路方法] 本課題研究的基本思路、具體研究方法、研究計畫及其可行性等。

3.1 研究的基本思路和具體研究方法

「第一階段」：資料搜集

語意解析引擎的建置需要大量的文本資料作為訓練準確度之基底，故前期需要藉由爬蟲的程式來進行網路理財投資的文本資料搜集的動作。本計畫以上海及深圳股市自 2009 年 1 月 1 日至 2018 年 12 月底之 10 年期間日資料為樣本期間；搜集資料包括每日投資人在這兩個股市所有相關社群網站對分別上市 ETF 個股發表的風險偏好、預期報酬、觀點及看法的充分資訊，建置完整的股票投資人個股情緒的資料庫，並定期更新。

「第二階段」：語意解析下結合網路社群聆聽

透過語意解析引擎立即進行語意的判斷並做出回饋，而語意解析引擎的建構須結合知識圖譜的建置，建構投資人語意解析與財金訊息勾稽之知識圖譜將會佈署在雲端等適合的環境中，最後再結合不同投資客群的社群聆聽來解讀投資大眾對未來公司發展及公司股價漲跌直接與真實的評價，來與實際股價走勢來相互印證。

「第三階段」：劃分樂觀及悲觀情緒的術語

本計畫分別針對投資人樂觀及悲觀情緒的 lth 類別來搜集所有潛在的特徵詞，並計算 lth 類別的 jth 特徵詞的詞頻。如 r^O 和 r^P 分別是樂觀及悲觀情緒的詞頻。在和樂觀及悲觀情緒相關的許多特徵詞中，代表性特徵詞相較於其他者更攸關。我使用卡方檢測以決定代表性特徵詞(Yang and Pedersen, 1997)。首先劃分屬於樂觀、悲觀及其他(和前兩類無關)類的特徵詞。檢定每個特徵詞是否屬於特定類情緒，每次鑒定兩類($l=2$)，如樂觀及非樂觀。虛無假設是 jth 特徵詞屬於非樂觀類，而對立假設是 jth 特徵詞屬於樂觀類。悲觀類可設計為相似的假設作法。卡方獨立性檢定的統計量為下列公式：

$$\chi^2(L, j) = \frac{(A+B+C+D) \times (A \times D - B \times C)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (1),$$

此處 A 和 B 是指 lth 類 jth 和 $non-jth$ 特徵詞的詞頻；C 和 D 是指 $non-lth$ 類 jth 和 $non-jth$ 特徵詞的詞頻；且 N 是指所有特徵詞的詞頻。假設 jth 特徵詞獨立於 lth 類。若卡方值高於臨界值，獨立於樂觀的虛無假設被拒絕；故判定 jth 特徵詞被視為樂觀(lth)類的代表性術語。

「第四階段」：估計情緒特徵詞的權重

本計畫接續採用焗權重法分別計算樂觀及悲觀情緒的特徵詞之權重。透過權重可指出特徵詞的強度。此方法已被廣泛使用在語意分析做為所有代表性特徵詞相對大小的衡量。以樂觀情緒為例，焗值公式如下述：

$$e_j^O = -k \sum_{i=1}^m r_{ij}^O \ln r_{ij}^O, k = 1 / \ln m; i = 1, 2, 3, \dots, m; j = 1, 2, 3, \dots, n. \quad (2)$$

此處 m 是廠商的數目； r_{ij}^O 是投資人對 ith 廠商樂觀情緒中 jth 特徵詞的詞頻；且 e_j^O 是樂觀情緒中 jth 特徵詞的焗值。本研究接續分別計算樂觀及悲觀情緒的 jth 特徵詞的權重。

$$w_j^O = (1 - e_j^O) / n - \sum_{j=1}^n e_j^O, j = 1, 2, 3, \dots, n. \quad (3)$$

$$w_j^P = (1 - e_j^P) / n - \sum_{j=1}^n e_j^P, j = 1, 2, 3, \dots, n. \quad (4)$$

此處 n 是投資人樂觀或悲觀情緒特徵詞的數目。權重愈大表示樂觀或悲觀的特徵詞愈重要。

「第五階段」：量化產生樂觀及悲觀情緒變數

此階段計算投資人在某月 t 對 ith 廠商的“樂觀情緒密度”(IOS_{it})及“悲觀情緒密度”(IPS_{it})。藉由對樂觀情緒計算特徵詞，量化投資人對特定廠商的樂觀情緒之程度。本計畫首先將投資人每日對 ith 廠商樂觀情緒 jth 特徵詞的詞頻 r_{idkj}^O 乘以相對應的權重 w_j^O ，並累積對樂觀情緒的所有代表性特徵詞，計算為每日 IOS (IOS_{idk})。接續，我加總每日的 IOS 計算為每月的 IOS，並加總每日的 IPS 計算為每月的 IPS。

$$IOS_{it} = \sum_{d=1}^{30} \sum_{k=1}^K IOS_{idk} = \sum_{d=1}^{30} \sum_{k=1}^K \sum_{j=1}^n r_{idkj}^O w_j^O \quad (5-1)$$

$$IPS_{it} = \sum_{d=1}^{30} \sum_{k=1}^K IPS_{idk} = \sum_{d=1}^{30} \sum_{k=1}^K \sum_{j=1}^n r_{idkj}^P w_j^P \quad (5-2)$$

此處下標 i , d , 和 k 是指 ith 廠商在 dth 天 kth 則網路資訊； r_{idkj}^O 是指投資人對 ith 廠商在 dth 天 kth 則網路資訊為樂觀情緒的 jth 特徵詞的詞頻； r_{idkj}^P 則是指相對應悲觀情緒的詞頻；且 w_j^O 和 w_j^P 是投資人樂觀及悲觀情緒的 jth 特徵詞的權重。被計算的 IOS_{it} 和 IPS_{it} 可隨時追蹤及更新投資人對 ith 廠商樂觀及悲觀情緒的程度。因為 K (網路資訊的數目) 和 n (特徵詞的數目) 是無限的，這兩個密度變數的最大值是無限的，本計畫分別對 IOS_{it} 和 IPS_{it} 將非 0 的密度值，使用均等分配的公式 $[99 * (x - x_{min}) / (x_{max} - x_{min}) + 1]$ (x 是密度值)，將之標準化介於最小 1% 和 最大 100% 之間；然而 0 密度值仍然為 0。較高的密度值意謂投資人有較強的樂觀或悲觀情緒。

「第六階段」：設立回歸模型—檢測情緒對股票報酬的影響及預測

最後本計畫使用下列的 OLS 回歸，以檢測投資人對廠商的樂觀及悲觀情緒對此月及後續月該股票累積異常報酬的影響：

$$CAR_{it,t+j} = \beta_0 + \beta_1 IOS_{it} + \beta_2 IPS_{it} + \lambda Z_{it}^{Ret} + \epsilon_{it}, \quad (6)$$

$$j = 0, 1, 2, 3$$

此處 i 是在本樣本 ith 廠商，且 t 是 tth 月。 CAR 是股票 i 於 t 至 $t+j$ 月 ($j=0, 1, 2, 3$) 的累積異常報酬，此處異常報酬 AR 是從市場模型所估計。估計期涵蓋在檢定期之前 250 個交易日。本計畫認為投資人情緒影響股票報酬應該不會超過一季。 AR 樣本期間是從 2009 年 1 月第一個交易日至 2018 年 12 月底最後一個交易日。 IOS 是指投資人對廠商樂觀情緒的程度，而 IPS 是指投資人對廠商悲觀情緒的程度。 Z^{Ret} 是橫斷面股票報酬的控制變數，包括 $Beta$, $Size$, BM 和 $Momentum$ (Fama and French, 1992; Jegadeesh and Titman, 1993)。每股“ $Beta$ ”是由估計市場模式得到，和 CAR 作法相似。 $Size$ 是每個廠商流通在外股票的去年底市值的對數值。 BM 是指淨值市價比。 $Momentum$ 是指每股報酬減檢定期前一年的報酬。最後採用真實報酬資料進行回溯測試，檢測以此模型預測報酬的準確程度。

3.2 計畫進度

(1) 2019 年為知識圖譜建置及網路爬蟲設計的階段，包含

2019.01~2019.06: 模型建構、搜集股票投資網路社群平臺和追蹤國內外文獻。

2019.07~2019.12: 知識圖譜建置、系統建置，及網路爬蟲程式設計與測試。

(2) 2020 年上半年進行投資人個股情緒資料庫建置的階段，包括

2020.01~2020.06: 結合各社群大資料得到不同類型投資人對個股發表的風險偏好、

預期報酬、觀點及看法的網資訊，透過網路爬蟲技術建構出投資人個股情緒的資料庫。

(3) 2020 年下半年完成情緒對股票報酬影響模型建置的階段，包括

2020.07~2020.12：將建置完成的投資人個股情緒的資料庫來構建樂觀及悲觀情緒變數，進而建置模型分析樂觀及悲觀情緒對股票報酬的影響及預測。

(4) 2021 年完成金融大資料平臺建置及成果展現的階段，包括

2021.01~2021.03：將情緒資料庫與情緒影響股票報酬模型自動串接，並依據產業分類完成金融大資料平臺建置；並採用真實資料進行回溯測試，檢測其預測報酬的準確程度。

2021.04~2021.12：將研究結果撰寫報告、申請結題；並撰寫成文章投稿至學術研討會或國際期刊。

4.[創新之處] 在學術思想、學術觀點、研究方法等方面的特色和創新。

(1) 本計畫是**第一篇**融合金融科技的最新技術，建構事件與投資人情緒的知識圖譜、網路爬蟲程式及金融大資料平臺，來捕捉投資人對個股樂觀及悲觀情緒的強度，以分析這兩種情緒強度對股票報酬的影響進而預測報酬率。

(2) 過去文獻衡量投資人情緒來預測股票報酬多使用“調查基礎”，“市場基礎”，“傳統媒體新聞基礎”三種方法來衡量投資人情緒。此三種方式各有其缺失，如當調查無反應率高時，調查基礎衡量變成不可靠；市場基礎衡量可能夾雜經濟推動的結果；傳統媒體新聞基礎通常不是來自單一來源，就是混雜股市之外來源的財務情緒資料形成的指數。本計畫建構金融大資料網路平臺以衡量投資人情緒模式，可解決上述三種方式的衡量缺失。

(3) 網路溝通具持續變動特質，近年已有少數文獻使用網路的資訊以預測股票報酬，但他們的網路資訊來源多來自單一網路資訊平臺，可能搜集到特定客群的情緒反應。本計畫建構金融大資料平臺模式爬取更多不同網路平臺的資訊，獲得來自不同投資客群的情緒反應才能結合社會聆聽，分析不同投資族群對於事件產生的情緒反應。

(4) 過去文獻處理網路資訊的做法，僅能捕捉投資人對整體股市看多或看空的時間序列趨勢，本計畫則可捕捉投資人在網路平臺上對某檔股票看多或看空的橫斷面及時間序列面的資訊，能掌握投資族群對個股樂觀及悲觀的情緒，可落實擇股及擇時的投資決策。

(5) 過去文獻處理網路資訊的做法，僅能透過網路資訊捕捉投資人對股市情緒資訊的頻率，本計畫能將情緒關鍵字的權重(重要性)結合其頻率轉化為情緒強度，同時創造投資人對個股樂觀及悲觀情緒的網路聲量，並可進一步預測對股票報酬的影響。

5.[預期成果] 成果形式、使用去向及預期社會效益等。

5.1 成果形式：

本課題預期成果的考核主要是：**(1) 投資人個股情緒資料庫的建置。**培訓相關人員利用網路爬蟲程式爬完上海及深圳股市上市股票所有網路社群平臺投資人討論個股看好或看壞等資訊，並定期抓取最新的資料產生投資人個股情緒資料庫。**(2) 研究報告。**從建置知識圖譜、使用網路爬蟲技術、產生投資人對個股正負面情緒的網路聲量、如何將正負面情緒強度勾稽金融大資料平臺、建置模型並依據產業分類，檢測樂觀及悲觀情緒對股票報酬影響等，完整說明於研究報告中。

5.2 使用去向及預期社會效益：

本課題研究成果的主要去向為：**(1) 服務投資人股票投資決策。**結合投資人對個股樂觀及悲觀情緒的資訊，提供投資人選股、擇時的股票投資建議。**(2) 服務金融主管機關股票投資政策。**服務金融主管機關股票投資的政策建議及諮詢。**(3) 成果發表。**將理論學術成果立足發表，發展中國特色的股票投資決策。

6 · [參考文獻] 開展本課題研究的主要中外參考文獻。

- Baker, M., and J. Wurgler, 2006. Investor sentiment and the cross-section of stock returns. *Journal of Finance*, 61, 1645–1680.
- Baker, M., and J. Wurgler, 2007. Investor sentiment in the stock market. *Journal of Economic Perspectives*, 21, 129–157.
- Brown, G.W., and M.T. Cliff., 2005. Investor sentiment and asset valuation. *Journal of Business*. 78 (2), 405–440.
- Brown, W., and M. Cliff, 2004. Investor sentiment and the Near-term stock market, *Journal of Empirical Finance*, 11(1), 1-27.
- Chen, H., P. De., Y.J. Hu and B.H. Hwang, 2014. Wisdom of crowds: the value of stock opinions transmitted through social media. *Review of Financial Studies*, 27 (5), 1367–1403.
- Cooper, M., R.C. Gutierrez and A. Hameed, 2004. Market states and momentum. *Journal of Finance*, 59, 1345–1365.
- Da, Z., J. Engelberg and P. Gao, 2015. The sum of all FEARS: investor sentiment and asset prices. *Review of Finance Studies*, 28 (1), 1–32.
- Dash, S.R., and D. Maitra, 2018. Does sentiment matter for stock returns? Evidence from Indian stock market using wavelet approach. *Finance Research Letters*, 26 (2018), 32–39.
- Fama, E.F., and K.R. French, 1992. The across-section of expected stock returns. *Journal of Finance*, 47, 427-465.
- Jegadeesh, N., and S. Titman, 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *Journal of Finance*, 48, 65-91.
- Renault, T., 2017. Intraday online investor sentiment and return patterns in the U.S. stock market. *Journal of Banking and Finance*, 84, 25–40.
- Smales, A.L., 2017. The importance of fear: investor sentiment and stock market returns. *Applied Economics*, 49 (34), 3395–3421.
- Sun, L., M. Najand and J. Shen, 2016. Stock return predictability and investor sentiment: a high-frequency perspective. *Journal of Banking and Finance*, 73, 147–164 .
- Tetlock, P.C., 2007. Giving content to investor sentiment: the role of media in the stock market. *Journal of Finance*, 62 (3), 1139–1168 .
- Yang, Y., and J.O. Pedersen, 1997, A Comparative study on feature selection in text categorization. *The International Conference On Machine Learning (ICML)*, 412–420.