

新聞情緒如何帶動股價走勢

－ 蘋概股三雄：台積電/鴻海/大立光

組別：伍陸柒捌

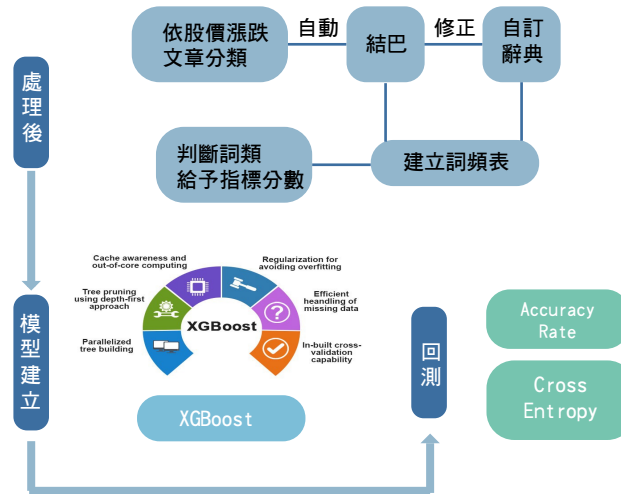
組員：工管四 廖守三
會計三 吳佳安

圖資一 張以柔
地理二 紀德鑫

壹、研究動機

近年來，金融商品數呈現爆炸性成長，投資人有更多選擇的背後，你有想過自己喜歡哪種分析策略嗎？
基本面 技術面 籌碼面？
技術分析沒有考慮到外部環境的影響，財報分析也很需要分析經驗的積累。有沒有指標能讓我們兼具兩者的同時，同時快速對多個標的進行分析呢？
我們都說新聞反應了大盤趨勢，某方面來說是已經經過消化的基本面資訊，然而隨著數據與資訊量的爆炸，投資者難以效率地吸收來自各方的質化資料。因此在這邊，本組希望發想一種利用新聞情緒分析股價波動性。
我們的目標並非最大化獲利，而是希望能在資訊爆炸的世界提供投資人精準但收斂的策略指標。

肆、分析流程架構



伍、資料清洗與特徵工程

關於當天股價究竟是漲還是跌，我們給此一個核心定義：

－ 漲：若當天股價的收盤價比前一日的最高價還要高出2%，那麼我們有足夠信心稱當天為股價漲日。

－ 跌：若當天股價的收盤價比前一日的最低價還要低2%，那麼我們有足夠信心稱當天為股價跌日。

－ 不漲不跌：由於本組對特殊事件要求嚴格，因此不符合上述條件的其餘交易日都被歸類為正常波動。

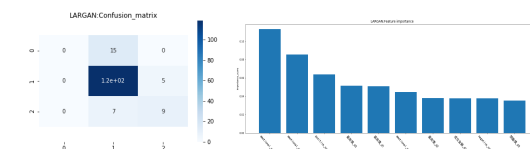
根據上述條件，在每日收盤後給予該日「漲」「跌」「正常」的訊號標籤。接著我們定義何謂「有影響力」的產經新聞。以下是我們的定義。

假設8/22收盤後，該日被認定是漲訊，則8/22往前推3日(8/21, 8/20, 8/19)的新聞被歸為漲訊文章；相反地，若8/22收盤後，該日被認定是跌訊，則8/22往前推3日(8/21, 8/20, 8/19)的新聞被歸為跌訊文章。不選擇8/22當天的新聞的，是因為我們發現投資人的反應並未那麼靈敏，多半會希望多抱幾天才脫手。

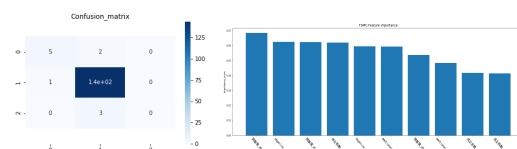
貳、主題簡介

在標的選擇上，考量標的不能討論度太低，否則訓練資料過少，同時該標的股價需易受國際局勢波動，漲跌的比例相當。結合上述條件，本組發現蘋概股非常適合作為研究對象。
蘋概股即為蘋果概念股，定義為上下游與蘋果公司密切合作的廠商。這類型公司非常容易受到政策、景氣、高層、技術的影響。台灣的蘋概股又以台積電、大立光與鴻海最具代表性。
該三家公司對台灣股市有顯著影響，本組以上述三家企業自中美貿易戰(2018年3月自2019年11月)至中美兩國達成協議前的時期，蒐集相關新聞與股價，以文字探勘進行分析對比，並驗證在事件日前若有相關新聞的發布，是否對於近日股價有所影響，探討新聞是否提前反應了股價。

陸、成果評估

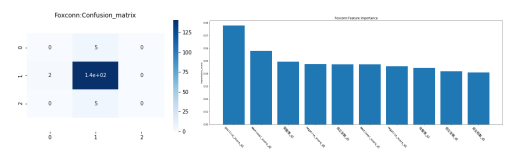


大立光
準確率：83%
重要特徵：3日前情緒分數/2日前正向分數/3日前最高價



台積電
準確率：97%
重要特徵：2日前開盤價/3日前負向分數/1日前開盤價

處理完每篇新聞的漲跌訊號之後，我們分別對正向新聞和負向新聞切詞(使用結巴內建字典及本組自定義字典，因此在對公司名稱、社會人物的切詞會較準確)並且使用多個指標篩出真正能代表正向負向文章的字詞。
在多项指標中，本組最終選擇 $t\text{-idf} * M * \text{df}$ 卡方值的前1000名字詞並使用CountVectorizer 將每一篇文章轉化為1x2000的稀疏矩陣向量。
根據文章有的正負向字詞豐富度，可以加權算出該篇文章的正負向情感分數和整理情感分數。



鴻海
準確率：92%
重要特徵：3日前正向分數/2日前情緒分數/2日前開盤價

資料取徑

1. 新聞資訊取得

利用爬蟲套件，蒐集中美兩國達成協議前中美貿易戰(2018年3月至2019年11月)時期網路(udn聯合新聞網、ptt等)標題提及台積電、大立光與鴻海的相關新聞。

2. 股價資訊取得

利用爬蟲套件，透過國家的股價資訊系統蒐集對應時期上述三家企業的日股價資料。