基於新聞評論之股價預測系統

專題編號：107-CSIE-S020
執行期限：106年第1學期至107年第1學期
指導教授：王正豪教授
專題參與人員： 102820030 袁士庭

一、摘要

股市的預測歷年來一直是投資人關切的議題，本專題將討論如何利用自動化的方式取得的投資目標相關的臺灣新聞媒體的財經新聞與評論，並運用機器學習的方法，分析該新聞的情緒性，並搭配其他股市重要的指標，進而預測其股市的漲跌。

為了能呈現期預測結果，本專題將以一個視覺化的方式作演示，讓使用者能有效的了解其預測的股市漲跌。

關鍵詞：股市預測、新聞情緒分析、監督式學習、類神經網路、情緒字典。

二、緣由與目的

關於投資目標的財經新聞與評論常為投資者所參考的重要資訊，而隨著資訊科技的發展，投資者能藉著網路的讀取各類的新聞媒體資料。但新聞媒體眾多，且新聞不斷地在產生，使得投資人在投資的過程中難以一一閱讀各家新聞媒體的新聞，取得更全面性的資訊來決定投資的方向。

在本專題中，希望藉著資訊檢索的方式自動獲得其投資目標相關的財經新聞與評論，並進行內容的分析得到該新聞的情緒性，進而預測其投資目標的漲跌，提供投資人一個參考。

三、技術與方法

本專題的架構主要為以下四個模組：「資料收集」、「語意分析」、與「股市預測」。

(一) 資料收集：

分為「新聞收集」與「股市收集」

1. 「新聞收集」：針對各家新聞媒體的網站，開發爬蟲程式，根據關鍵字取得相關新聞。
2. 「股市收集」：自股市網站上取得投資目標的股市資訊。

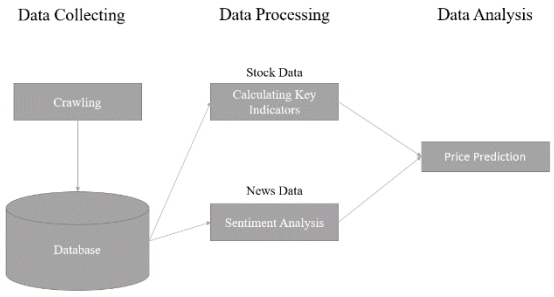
(二)語意分析：

利用如 SnowNLP 是針對中文語意的工具，TextBolb 針對英文語意的，將新聞內文進行語意分析，並計算新聞情緒正負值。

(三)股市預測：

將新聞評論情緒與其他股市重要指標做搭配，利用類神經網路的模型如 LSTM 進行股市預測。

四、方法架構



(圖1. 方法架構圖)

將收集的新聞資料與股市資料存進資料庫後進行後續的處理與分析，用製作的字典進行語義分析後的情緒值與重要指標結合，做訓練與預測。

五、實驗流程：

(一) 資料收集：

1. 新聞收集：台灣新聞媒體網站多有作新聞分類，依各網站的網頁結構製作爬蟲程式並蒐集新聞網站的財經分類下，投資目標關鍵字相關的新聞。另外針對英文也收集相關的新聞來分析

。

2. 股市收集：製作爬蟲程式並從台灣「公開資訊觀測站」收集投資目標的財務報表。也爬取美國知名指標道瓊工業平均指數來做分析。

(二)資料處理：

1. 新聞資料處理：雖然各新聞媒體網站有作關鍵字搜尋的功能，但其新聞帶有該關鍵字不一定是其相關的新聞，需依如標題，字頻作進一步分類。

2. 股市資料處理：依收集到的財務報表，計算如 KD 值、RSI、MA 等重要指標，並加上漲跌幅度的特徵，以利後續分析使用。

(三)情緒分析：

利用現有的工具搭配原本通用字典與上述方法製作的財經字典的結合，分析新聞內容的情緒正負值，作為該新聞的特徵值。

(四)預測模型：

使用類神經網路，做股市漲跌的預測，並以1. 股市重要指標、2.新聞情緒、3. 股市重要指標加上新聞情緒等三種組合作為模型的特徵，並進一步做比較來探討新聞情緒在股市預測佔有的重要性。

六、使用工具：

1. SnowNLP:

是一個 python 寫的類庫，可以方便的處理中文文本內容，中文分詞、詞性標註和情感分析等功能。

2. Keras:

Keras 是一個開放原始碼，高階深度學習程式庫，使用 Python 編寫，能夠運行在 TensorFlow。

3. BeautifulSoup:

Beautiful Soup 是一個 Python 的函式庫模組，以少量的程式碼，就可快速解析網頁 HTML 碼，從中取出想要的資料，加快程式撰寫速度

4. TextBlob:

TextBlob是一個用Python編寫的開源的文本處理庫。它可以用來執行很多自然語言處理的任務，比如，詞性標註，名詞性成分提取，情感分析，文本翻譯

七、成果與結論：

本專題嘗試將新聞評論與新聞股市的趨勢做連結，發現中文語意的研究和英文語意研究的差別，國外有許多針對英文字詞的研究和情緒字典。目前中文最有名的為台大的中文意見詞典 NTUSD，但僅做正負情緒的分析。若要加強中文語意的研究，勢必需要語文研究的基礎在更上一層。

另外新聞語意和股市的關聯性並無預期中的高，猜想影響股市的因素過多，資訊也不對稱，若想要進行股市預測勢必需要更多的特徵和因素加以考量。

參考文獻

[1] 王宏亘，「財經新聞情緒及投資市場走勢之相關性研究」，**國立台北科技大學資訊工程系**，2015，06月。

[2] 黃驤，「基於財經新聞詞彙分布之新聞極性分析」，**國立台北科技大學資訊工程系**，2014，08月。

[3] FinLab (2018)，「Python 財經」，https://www.finlab.tw/categories/%E8%B2%A1%E7%B6%93Python%E6%95%99%E5%AD%B8/。

[4]. Antonio Moreno-Ortiz “Lingmotif: Sentiment Analysis for the Digital Humanities” University of Malaga ´ Spain

[5] Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng and Christopher Potts “Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank”Stanford University