



# 台股新聞標題文字探勘及情緒預測

張庭瑄

國立清華大學計量財務金融研究所


# 目錄




 研究動機

 資料介紹

 探索性資料分析

 新聞資料預處理

 SVM預測結果

 結論

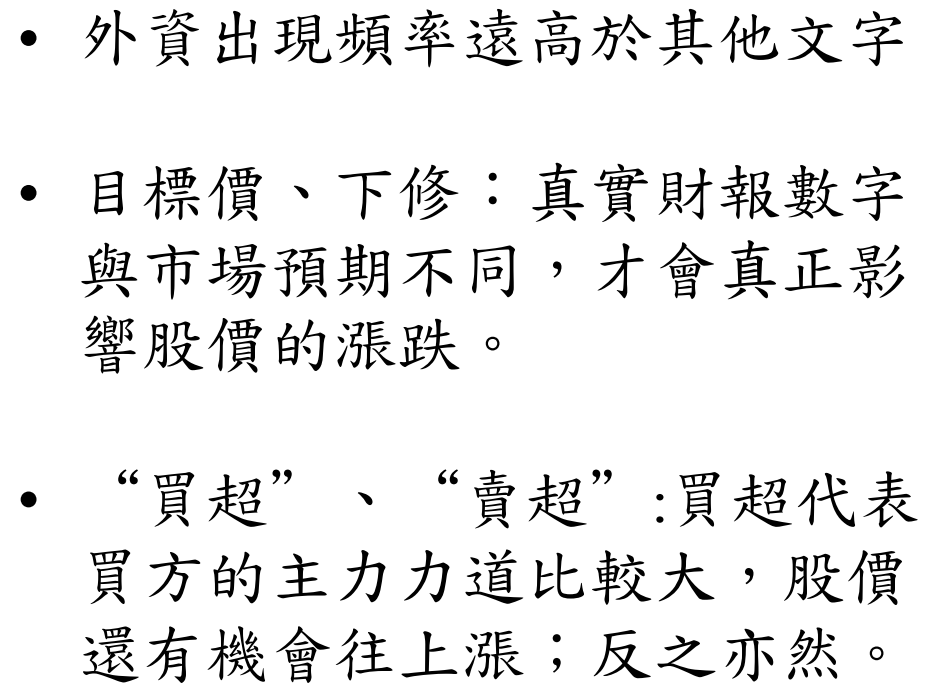
- 新聞消息面對投資人的投資決策十分重要
- 花費大量時間閱讀新聞沒效率也不可行
- 本文的研究目的希望透過文字探勘分析新聞標題的用詞，快速預測每篇新聞消息的好壞，幫助投資人能快速判斷市場行情的好壞。

# 資料介紹



- 資料樣本：台股新聞標題
- 資料樣本期間:2020/2/11~2020/4/30
- 樣本數：228筆
- 類別分成正、反兩類
- 新聞標題判斷標準：對台股影響

新聞消息	個數	比率
正面	137	47.57%
負面	151	52.43%



- 中文斷字處理 : 結巴( jieba )的詞庫 + 證交所公布台股上市證券名單
- TF-IDF的方法來對新聞資料做特徵選取，評估一字詞對於一個檔案集或一個語料庫中的其中一份檔案的重要程度

$$tfidf = \begin{matrix} & \begin{matrix} \text{詞1} & & \text{詞}t & & \text{詞}T \end{matrix} \\ \begin{matrix} \text{新聞1} \\ \vdots \\ \text{新聞}d \\ \vdots \\ \text{新聞}D \end{matrix} & \begin{bmatrix} tf_{1,1} * idf_1 & \cdots & tf_{1,t} * idf_t & \cdots & tf_{1,T} * idf_T \\ \vdots & & \vdots & & \vdots \\ tf_{d,1} * idf_1 & \cdots & tf_{d,t} * idf_t & \cdots & tf_{d,T} * idf_T \\ \vdots & & \vdots & & \vdots \\ tf_{D,1} * idf_1 & \cdots & tf_{D,t} * idf_t & \cdots & tf_{D,T} * idf_T \end{bmatrix} \end{matrix}$$



# SVM預測結果

Kernel : Linear

預測分類 / 真實分類	反面	正面
反面	15	29
正面	1	13

Kernel : Radial

預測分類 / 真實分類	反面	正面
反面	15	1
正面	36	6

Kernel : Sigmoid

預測分類 / 真實分類	反面	正面
反面	13	3
正面	24	18

- 準確度(Accuracy)=48.28%
- 特異度 (Specificity) =30.95%
- 靈敏度 (Sensitivity) =93.75%
- 準確度(Accuracy)為36.21%
- 特異度 (Specificity) 為85.71%
- 靈敏度 (Sensitivity) 為29.41%
- 準確度(Accuracy)為53.45%
- 特異度 (Specificity) 為85.71%
- 靈敏度 (Sensitivity) 為35.14%

# SVM預測結果



## 模型比較

模型	準確度 (Accuracy)	特異度 (Specificity)	靈敏度 (Sensitivity)
Linear	48.28%	30.95%	93.75%
Kernel: Radial	36.21%	85.71%	29.41%
Kernel: Sigmoid	53.45%	85.71%	35.14%

- Sigmoid 的準確度最高
- 線性模型的靈敏度最好
- Radial為核函數的模型不適合此樣本的分類
- 模型設定：陽性為負面消息，陰性為正面消息
- 對市場壞消息較敏感的投資人，線性的SVM模型比較適合；對好消息較敏感的投資人，則選擇Sigmoid模型較佳。



- 台灣財經新聞以外資、主力動向以及與市場預期落差為主要關注重點
- 針對本次研究，對市場壞消息較敏感的投資人，比較適合線性的SVM模型；對好消息較敏感的投資人，則選擇Sigmoid模型較佳。
- 模型預測結果不佳，因為新聞的資料搜集不易，加上新聞標題的正負面判斷費時費力，因此樣本數過少，會產生 $p \gg n$ 的問題



Thanks