# 統計計算期末書面報告

# 台股新聞標題文字探勘及情緒預測

清大計財所

學號：108071503

姓名：張庭瑄

目錄

# 一、 研究目的

新聞消息面對於投資人而言是極為重要的工具，特別是投資部門的基金經理人以及交易員，每天都必須大量閱讀大量財經新聞，來掌握金融市場動態，但每個人的時間有限，花費大量時間閱讀市場消息十分沒效率也不現實，因此本文的研究目的希望透過文字探勘分析新聞標題的用詞遣字，快速預測每篇新聞消息的好壞，以利投資人以少量時間就能快速判斷市場行情的好壞，將更多的時間留給產業研究或是策略撰寫上。

# 二、研究方法

本報告分成兩個步驟，第一步先利用文字探勘的模型及技術，將財經新聞的標題轉換成文字變數，本文採用Term Frequency-Inverse Document Frequency, TF-IDF來對文字資料做特徵選取與維度縮減，將新聞標題量化，取得文本的特徵，以利後續的步驟；第二步利用監督是機器學習中的支持向量機( Support Vector Machine, SVM )的模型來預測新聞消息的好壞，本文首先將樣本切分成訓練資料集以及測試資料集，利用訓練資料集的文字變數來建立SVM的模型，再將測試資料帶入模型預測該文章的類別，透過與真實類別比較，計算模型的正確度。

# 研究議題：台股新聞標題文字探勘及情緒預測

## 2.1 議題介紹

每天股市的新聞是投資人判斷市場的重要工具，但光是閱讀每日的新聞，就必須花費大量的時間，十分沒效率，因此若能透過文字探勘的方法將新聞標題資料轉換成文字變數，並利用監督是機器學習的模型來預測該篇新聞所隱含好壞消息，以利投資人快速消化每日的大量的新聞消息面，只要只要將每日的新聞資料帶入模型，就能快速判斷市場的動向，以利後續策略的調整以及產業、個股研究。

## 2.2 資料說明與探索性分析

### 2.2.1.資料說明：

資料樣本為2020年2月11至4月30的台灣股市新聞，共288筆資料，由人工搜集而來，並將新聞的消息面分成好壞兩個類別，並以工人智慧的方法判斷並註記，判斷原則是以該新聞對台股市場的影響，來決定該新聞標題為正面或是負面消息，樣本標記的結果如下，可以看出搜集的新聞消息為正面、反面的比率大概一半一半，因此本研究不考慮類別不平衡的問題。

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

### 2.2.2.探索性資料分析

圖 1 文字頻率圖

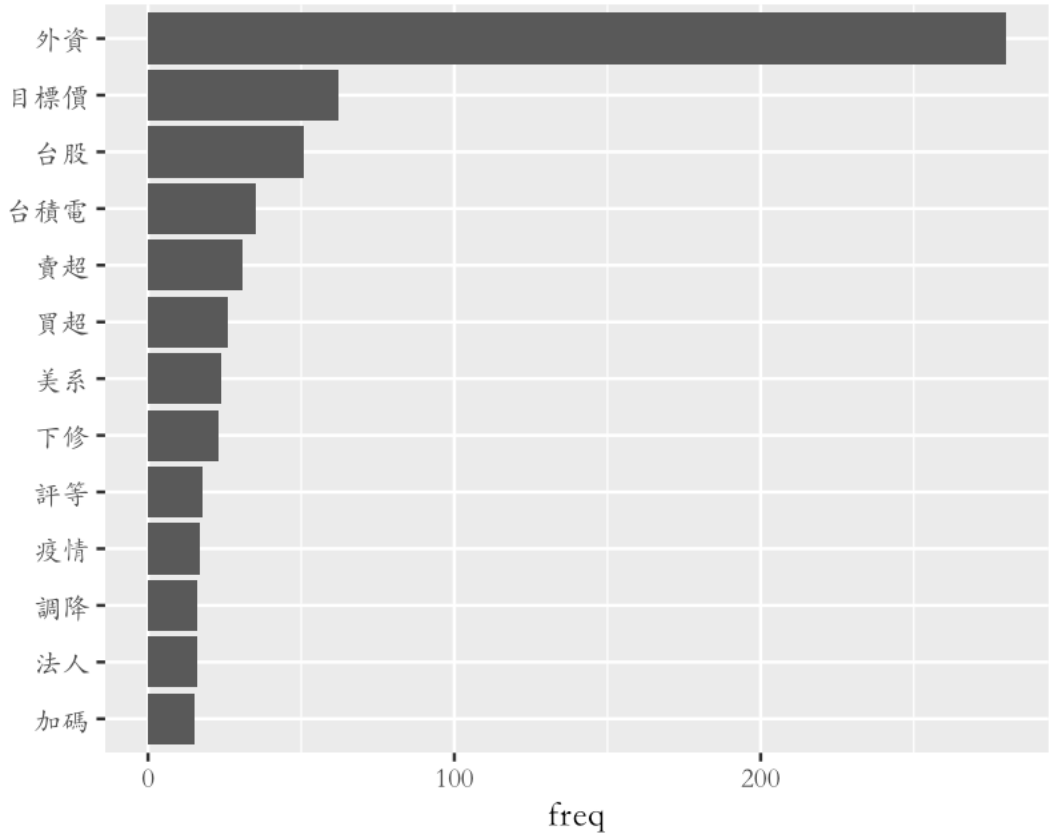


圖2 文字雲


從上面文字頻率圖以及文字雲兩張圖可以觀察出來，“外資”出現的頻率遠高於其他文字，這點十分符合市場的共識，主要因為台灣資本市場相對較小，在籌碼面容易受到法人的影響，而三大法人中，影響力和資本額最大的都是外資，因此台股市場的投資人喜歡以外資的動向為主要觀察的目標，提到外資的新聞自然而然越來越多，再補中一點，“美系”通常和“外資”一起出現，而所謂的美系外資是指美林、高盛、花旗、摩根大通以及台灣Morgan Stanley。

而目標價、下修這些字則代表了市場投資人在做決策時，都會與市場預期的價格比較，根據效率市場理論，股價已經充分反映過去市場的所有信息，因此只有當公布的財報或是即時消息，才會真正影響股價的漲跌，因此財經新聞也會著重在預期以及真實數值的公布上。

此外， “買超”及“賣超”是將主力券商的買賣超張數做加總，如果那天最後的結果是買超，代表買方的主力力道比較大，股價還有機會往上漲；反之，當最後結果是賣超時，代表賣方的力道更強，股票被放空的機率較高，因此市場當天主力的買賣情況也是投資人及財經媒體重點觀察的部分。

## 2.3 模型說明與資料預處理

此小節模型說明分成兩部分，第一部分首先介紹本文對新聞資料的預處理方法，以及TF-IDF的模型的詳細說明，而第二部分則為支持向量機模型的介紹。

### 2.3.1 新聞資料預處理

#### 2.3.1.1 中文斷字處理

本文在中文斷字處理的部分採用結巴( jieba )的詞庫來處理中文斷字的問題，但在財經相關專有名詞的部分，因為網路開放平台中繁體財金詞庫的資源並不多，因此透過筆者的領域相關知識自行建立，將台灣股市新聞常用的專有名詞及動詞加入詞庫，而在台灣股票名稱的部分，本研究採用台灣證券交易所公布「本國上市證券國際證券辨識號碼一覽表」中的股票名稱為標準，將台股名稱加入詞庫之中。此外，台灣股票新聞經常公布股票漲跌的詳細數字，但數字的大小對於本文的研究關係不大，因此本文將數字以及標點符號刪除，以利後續文本特徵選取的進行。

#### 2.3.1.2 TF-IDF ( Term Frequency – Inverse Document Frequency )

本文採用TF-IDF的方法來對新聞資料做特徵選取，該統計方法是一種用於資訊檢索與文字挖掘的常用加權技術，用來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，常作為檔案與使用者查詢之間相關程度的度量或評級，被應用在搜索引擎上。

直覺來說，字詞的重要性會隨著它在檔案中出現的次數成正比增加，但同時會隨著它在語料庫中出現的頻率成反比下降。所謂詞頻( Term Frequency，TF）指的是某一個給定的詞語在該檔案中出現的頻率。對於在某一特定檔案裡的詞語$t_i$來說，它的重要性可表示為：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

其中$n_{i,j}$是該詞在檔案$d_j$中的出現次數，而分母則是在檔案$d_j$中所有字詞的出現次數之和。

而逆向檔案頻率（Inverse Document Frequency，IDF）是一個詞語普遍重要性的度量，而某一特定詞語的idf，可以由總檔案數目除以包含該詞語之檔案的數目，再將得到的商取以10為底的對數得到：

$$idf_i = \frac{|D|}{|\{j: t_i \in d_j|}$$

其中

|D|：語料庫中的檔案總數

$|\{j: t_i \in d_j|$：包含詞語$t_i$的檔案數目（即$n_{i,j} \neq 0$的檔案數目），如果詞語不在資料中，導致分母為零，一般情況下使用1+$|\{j: t_i \in d_j|$。

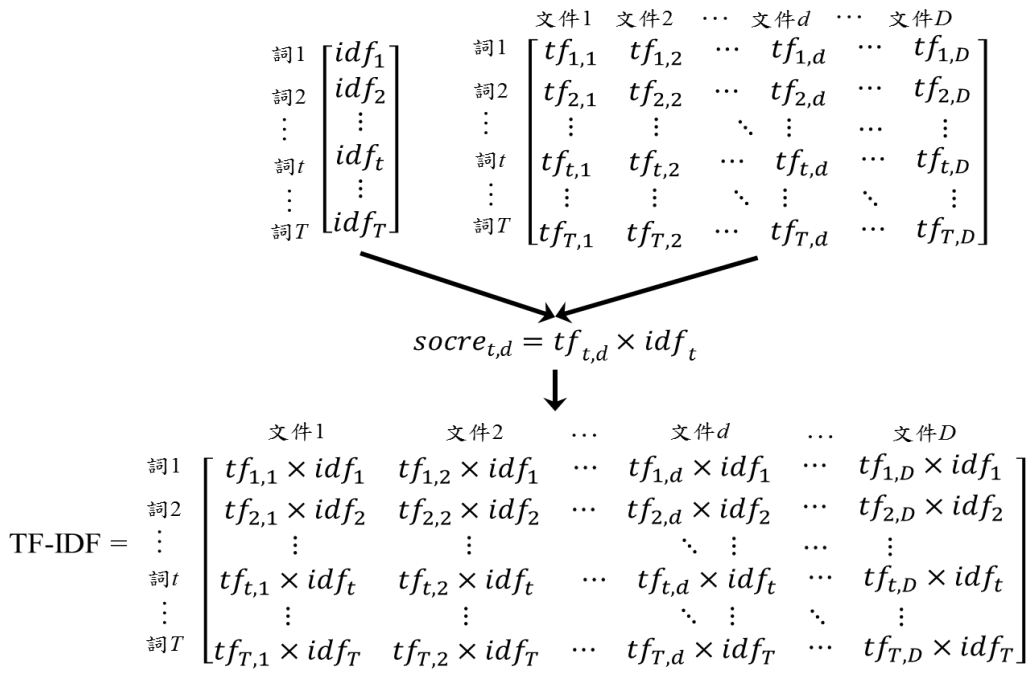
因此我們可以藉由上面兩項計算每一個「詞」對每一篇「文件」的分數 (score)，

定義為：

$$score_{t,d} = tfidf_{t,d} = tf_{t,d} * idf_t$$

由於某一特定檔案內的高詞語頻率，以及該詞語在整個檔案集合中的低檔案頻率，可以產生出高權重的$tfidf$。因此，$tfidf$傾向於過濾掉常見的詞語，保留重要的詞語。

TF-IDF的矩陣如下圖：




從上面的圖可以看出，TF-IDF將文件資料轉換成文件與詞彙的變數矩陣，而在此研究之中，column的文件即為新聞標題的樣本，row的詞則為此樣本所產生出的所有詞彙，也可以視為一特徵向量，之後再加入已經分類好的Y，便可以帶入SVM模型進行建模及預測。

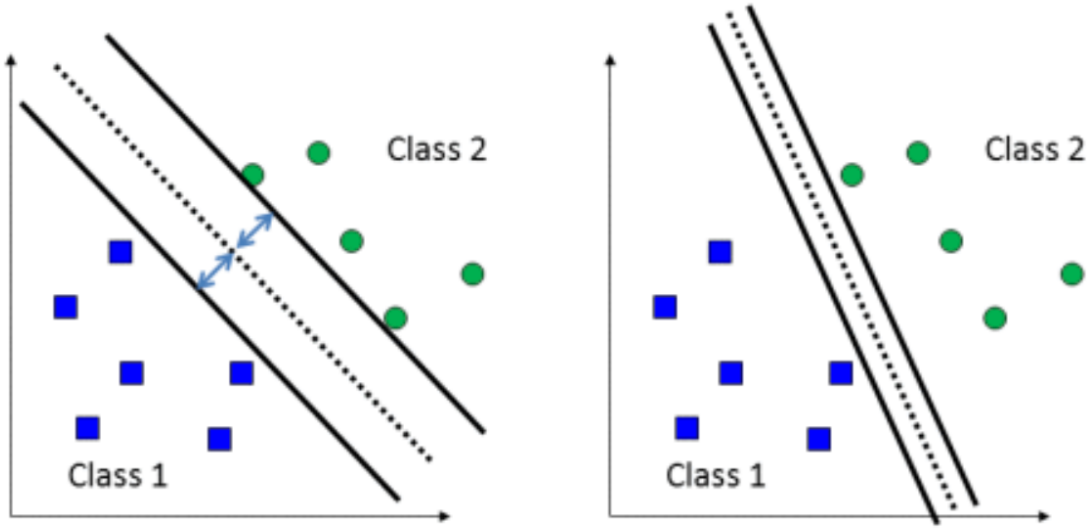
### 2.3.2 支持向量機 ( Support Vector Machine, SVM )

#### 2.3.2.1 SVM 介紹

支援向量機是在機器學期中的監督式學習模型，本身是一個二元 (binary) 的分類器，簡單來說就是一種演算法，試圖從資料中建構一個超平面(hyperplane)，將資料區分成兩個類別(2 classes)，最後進行預測/分類，而SVM除了能進行線性分類之外，還可以使用核技巧來有效地進行非線性分類。

假設給定的資料點屬於兩個類別，而目標是將新資料分類。對於支持向量機來說，資料點被視為p 維向量，目標利用p-1 維超平面來分開這些點，也就是所謂的線性分類器，當中可能有許多超平面可以把資料分類，因此最佳超平面的合理選擇是以最大間隔把兩個類別分開的超平面，也就是說，我們要選擇能夠讓

到每邊最近的資料點的距離最大化的超平面，下圖為二維的SVM模型，其中虛線為超平面，而實線之間的距離為margin。




。

2.3.2.2 Hard Margin、Soft Margin與非線性SVM

為了解決更複雜的問題，SVM又可以分為hard margin、soft margin，還有非線性SVM模型，Hard margin SVM 的假設是存在一超平面，完美將「所有」資料分成兩邊且具有最大margin，但由於hard-margin SVM，追求要將資料完美分好，因此很容易有overfitting的風險。

因此後來有人提出“soft-margin SVM”，讓SVM能容許一些被分錯的資料存在。在下圖中，soft-margin SVM的損失函數(loss function)中的C就是懲罰項，藉由該項可以給予那些被分錯的資料懲罰值，以控制support vectors(用來決定超平面的那些資料點)的影響力。

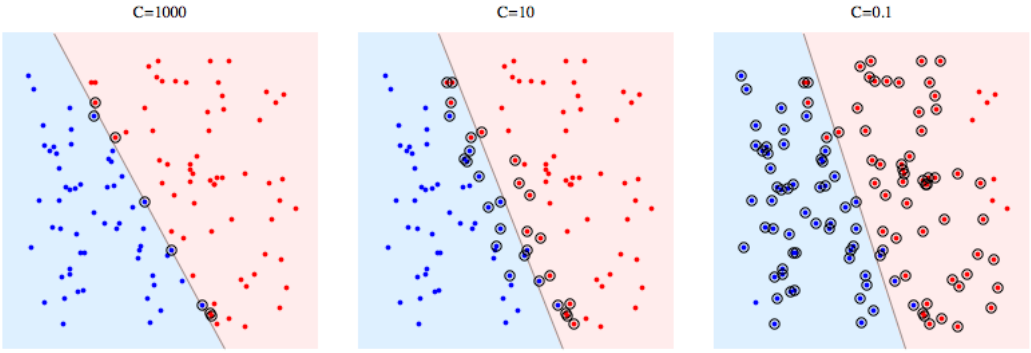
•

$$\text{minimize} \quad \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$\text{subject to} \quad y_i\left(w^T x_i - b\right) - 1 + \xi_i \geq 0 \quad \forall i$$

$$\xi_i \geq 0 \quad \forall i$$

換句話說，C越大，代表容錯越小，越少support vectors，越接近hard-margin SVM的概念，卻容易overfitting；C越小，代表容錯越大，越多support vectors，可以追求更大的margin。




如上圖所示，圈圈的點代表support vectors，用來決定margin的大小，當C=1000時，support vectors的點幾乎都發生在線上面，很接近hard-margin SVM的概念。當C越來越小，隨著support vectors的點越來越多，表示margin的範圍越來越大。

三者之間的差別，在於hard margin以及soft margin的kernel function 為線性，而非線性SVM模型的kernel function為非線性函數，其主要目標是尋找能夠將資料最佳分類的曲面。

# 三、結果與結論

## 3.1 線性SVM

### 3.1.1 Cross Validation決定最適模型

| Cost | Error | Dispersion |
|---|---|---|
| 0.001 | 0.4130435 | 0.0875583 |
| 0.01 | 0.4130435 | 0.0875583 |
| 0.1 | 0.4130435 | 0.0875583 |
| 1 | 0.2913043 | 0.05442024 |
| 5* | 0.2826087* | 0.05518676* |
| 10 | 0.2826087 | 0.05518676 |
| 100 | 0.2826087 | 0.05518676 |

*代表cross-validation error rate最小

所謂的Cost就是在soft-margin SVM的損失函數(loss function)中的容錯項，C越大，代表容錯越小，越少support vectors，越接近hard-margin SVM的概念，但卻容易over fitting；相反的當C越小，代表容錯越大，越多support vectors，可以追求更大的margin，而本研究採用Cross Validation的方法來選取最適的C，從上表可以看出當Cost=5時，有最小的cross-validation error rate，因此將之視為線性SVM中最適模型。

### 3.1.2 預測結果

藉由上述CV所決定之最適模型，將測試資料帶入模型預測，並與真實分類結果比較，其結果如下表：

| 預測分類 / 真實分類 | 反面 | 正面 |
|---|---|---|
| 反面 | 15 | 29 |
| 正面 | 1 | 13 |

我們可以從上表觀察出在線性SVM下總共有28筆資料被正確分類，準確度(Accuracy)為48.28%，特異度 (Specificity) 為30.95%，靈敏度 (Sensitivity) 為93.75%。

## 3.2 非線性SVM

### 3.2.1 Radial模型

#### 3.2.1.1 Cross Validation決定最適模型

| cost | gamma | error | dispersion |
|---|---|---|---|
| 0,01 | 0.5 | 0.4130435 | 0.0875583 |
| 0,1 | 0.5 | 0.4130435 | 0.0875583 |
| 1 | 0.5 | 0.4173913 | 0.08500219 |
| 10* | 0.5* | 0.3565217* | 0.06735623* |
| 100 | 0.5 | 0.3565217 | 0.06735623 |
| 0,01 | 1 | 0.4130435 | 0.0875583 |
| 0,1 | 1 | 0.4130435 | 0.0875583 |
| 1 | 1 | 0.4173913 | 0.0824942 |
| 10 | 1 | 0.4217391 | 0.07951227 |
| 100 | 1 | 0.4217391 | 0.07951227 |
| 0,01 | 2 | 0.4130435 | 0.0875583 |
| 0,1 | 2 | 0.4130435 | 0.0875583 |
| 1 | 2 | 0.4173913 | 0.0824942 |
| 10 | 2 | 0.4173913 | 0.0824942 |
| 100 | 2 | 0.4173913 | 0.0824942 |
| 0,01 | 3 | 0.4130435 | 0.0875583 |
| 0,1 | 3 | 0.4130435 | 0.0875583 |
| 1 | 3 | 0.4130435 | 0.0875583 |

| | | | |
|---|---|---|---|
| 10 | 3 | 0.4173913 | 0.0824942 |
| 100 | 3 | 0.4173913 | 0.0824942 |

*代表cross-validation error rate最小

從上表可以看出當Cost=10，gamma=0.5時，有最小的cross-validation error rate，因此將之視為非線性SVM (radial)中的最適模型。

3.2.1.2預測結果

| 預測分類 / 真實分類 | 反面 | 正面 |
|---|---|---|
| 反面 | 15 | 1 |
| 正面 | 36 | 6 |

我們可以從上表觀察出在非線性SVM(radial)下總共有21筆資料被正確分類，準確度(Accuracy)為36.21%，特異度 (Specificity) 為85.71%，靈敏度 (Sensitivity) 為29.41%。

## 3.2.2 Sigmoid模型

3.2.2.1 Cross Validation決定最適模型

| cost | gamma | error | dispersion |
|---|---|---|---|
| 0,01 | 0.5 | 0.4130435 | 0.0875583 |
| 0,1 | 0.5 | 0.4130435 | 0.0875583 |
| 1 | 0.5 | 0.3695652 | 0.08512565 |
| 10* | 0.5* | 0.2826087 | 0.06874517 |
| 100 | 0.5 | 0.2826087 | 0.06874517 |
| 0,01 | 1 | 0.4130435 | 0.0875583 |
| 0,1 | 1 | 0.4130435 | 0.0875583 |
| 1 | 1 | 0.2869565 | 0.04673773 |
| 10 | 1 | 0.2782609 | 0.05869118 |
| 100 | 1 | 0.2869565 | 0.07723428 |

| | | | |
|---|---|---|---|
| 0,01 | 2 | 0.4130435 | 0.0875583 |
| 0,1 | 2 | 0.4130435 | 0.0875583 |
| 1* | 2* | 0.2521739* | 0.05724179* |
| 10 | 2 | 0.326087 | 0.05123962 |
| 100 | 2 | 0.3695652 | 0.10299037 |
| 0,01 | 3 | 0.4130435 | 0.0875583 |
| 0,1 | 3 | 0.4130435 | 0.0875583 |
| 1 | 3 | 0.273913 | 0.06165811 |
| 10 | 3 | 0.3478261 | 0.05797101 |
| 100 | 3 | 0.3695652 | 0.07460602 |

*代表cross-validation error rate最小

從上表可以看出當Cost=1，gamma=2時，有最小的cross-validation error rate，因此將之視為非線性SVM (sigmoid)中的最適模型。

3.2.2.2預測結果

| 預測分類 / 真實分類 | 反面 | 正面 |
|---|---|---|
| 反面 | 13 | 3 |
| 正面 | 24 | 18 |

我們可以從上表觀察出在非線性SVM(radial)下總共31筆資料被正確分類，準確度(Accuracy)為53.45%，特異度 (Specificity) 為85.71%，靈敏度 (Sensitivity) 為35.14%。

## 3.3 結果比較

| 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% |

從上圖可以觀察出來利用Sigmoid 為核函數的模型準確度最佳，線性模型的準確度略低於Sigmoid模型，但靈敏度最高，而核函數為Radial的非線性模型分類結果準確度最差，靈敏度也比Sigmoid模型低，因此我們可以判斷以Radial為核函數的模型不適合此樣本的分類。

我們知道高靈敏度的模型很少忽略真陽性，而高特異度檢定則很少將不是檢定目標的其他東西鑑別為陽性，因此靈敏度(又稱真陽性率)可以作為避免假陰性的量化指標，而特異度(真陰性率)可以作為避免假陽性的量化指標，而在本文的模型設定中，陽性為負面消息，陰性為正面消息，因此靈敏度高代表真實分類為負面消息且預測結果也為負面的消息比率高，也就是該模型對負面新聞的判斷、分類結果越好，相對的，高特異度則代表實際為正面消息且預測結果也為正面消息的比率高，也就是說該模型對正面消息的判斷、分類結果越好。而本次研究的結果指出，若是對市場壞消息較敏感的投資人，通常希望對負面消息的判斷能夠越高，因此線性的SVM模型比較適合，而對好消息較敏感的投資人，則選擇Sigmoid模型較佳。

## 3.4 結論

1. 台灣投資人對外資相關的新聞關注度高，特別是美系的外商；針對主力動向的新聞也是財經媒體關注的重點，以及市場投資人十分重視與市場預期落差的部分，以此作為判斷的標準。
2. 本研究的結果：若是對市場壞消息較敏感的投資人，比較適合線性的SVM模型，而對好消息較敏感的投資人，則選擇Sigmoid模型較佳。
3. 本研究的硬傷為新聞的資料搜集不易，加上正負面新聞標題的判斷需要人為判斷，因此樣本數過少，會產生$p \gg n$的問題，而在模型預測的結果上，正確率大概只有50%，還有更多改進空間。

# 四、參考資料


1.上課講義

2.網路資源

https://zh.wikipedia.org/wiki/Tf-idf

http://www.hmwu.idv.tw/web/R/G03-hmwu_R-TextVis.pdf

https://pse.is/3jcxm5

https://pse.is/3hty5w

https://www.lexjansen.com/nesug/nesug10/hl/hl07.pdf

https://taweihuang.hpd.io/2018/12/28/imbalanced-data-performance-metrics/

https://pse.is/3jj2lt

https://pecu.gitbooks.io/-r/content/wen-zi-yun.html

https://bigdatafinance.tw/index.php/data-visualization/862-2019-05-26-14-56-40

https://rpubs.com/JJChiou/textmining_2

https://rpubs.com/JJChiou/textmining_1

https://zhuanlan.zhihu.com/p/131334741

https://pse.is/3gtrq9

https://tangshusen.me/2018/10/27/SVM/

https://rpubs.com/skydome20/R-Note14-SVM-SVR

https://rpubs.com/skydome20/R-Note14-SVM-SVR