**銘　　傳　　大　　學**

**資 訊 工 程 學 系**

**專 題 研 究 總 審 文 件**

本校一○五學年度 資訊工程學系

組員：　　周彥佑 、 陳勁瑋 、 張皓博

所提專題研究：　情緒分析系統之研究－以臉書粉絲團為例

指 導 教 授 ： 　　王豐緒 教授

**中華民國　一○五　年　十　月　十　日**

# 摘要

近年來社群網站的發展已趨成熟，社群網站已經成為大部分人最常使用的工具。如何善用社群網站的留言資訊，進行情緒分析以利決策的運行逐漸成為近幾年來受到熱烈研究的議題之一。本研究是採用趙妤瑄與王豐緒[9]所提出的TF-RF-IDF權重計算方式搭配SVM分群演算法來開發展情緒分析的系統。本系統將運算的結果以圖表呈現，來提供使用者了解政治人物特定時間內在FB粉絲專頁裡文章留言的正、負面評價的數量以及主要辭彙分佈，以利於決策者利用此系統做為決策參考的依據。本研究除延續[9]的研究對象與研究方法之外，也修改分類方法提升正面和負面文章的分類績效，並新增一位研究對象，以檢視其分類成效。預期本研究成果可以確認先前文獻[9]的結論，並實做出一套系統透過網頁平台的方式提供決策者使用。

目錄

[**摘要** II](#_Toc463979297)

[**目錄** III](#_Toc463979298)

[**圖目錄** VI](#_Toc463979299)

[**表目錄** VIII](#_Toc463979300)

[**第一章** **緒論** 1](#_Toc463979301)

[**第一節** **研究背景** 1](#_Toc463979302)

[**第二節** **研究動機** 2](#_Toc463979303)

[**第三節** **研究目的** 3](#_Toc463979304)

[**第二章** **文獻探討** 4](#_Toc463979305)

[**第一節** **網路文本情緒分析近況** 4](#_Toc463979306)

[**第二節** **情緒分析與意見探勘** 4](#_Toc463979307)

[**第三節** **辭彙情緒分析及建立情緒辭典** 5](#_Toc463979308)

[**第四節** **資料分類法** 7](#_Toc463979309)

[**第五節** **情緒分析應用** 7](#_Toc463979310)

[**第三章** **研究方法** 9](#_Toc463979311)

[**第一節** **研究流程** 9](#_Toc463979312)

[**第二節** **使用工具** 10](#_Toc463979313)

[**第三節** **語料收集** 10](#_Toc463979314)

[**第四節** **前處理與人工標記文章** 11](#_Toc463979315)

[**第五節** **詞彙篩選與建置情緒辭典** 12](#_Toc463979316)

[**第六節** **情緒加權值計算與留言情緒分類** 13](#_Toc463979317)

[**第四章** **可行性分析** 15](#_Toc463979318)

[**第一節** **經濟可行性** 15](#_Toc463979319)

[**第二節** **技術可行性** 15](#_Toc463979320)

[**第五章** **系統分析與設計** 16](#_Toc463979321)

[**第一節** **環境圖(**Context Diagram**)** 16](#_Toc463979322)

[**第二節** **使用案例圖(**Use Case Diagram**)** 17](#_Toc463979323)

[**第三節** **類別圖(**Class Diagram**)** 19](#_Toc463979324)

[**第四節** **實體關聯圖(**Entity Relationship Diagram**)** 19](#_Toc463979325)

[**第五節** **系統架構圖(**System Structure Diagram**)** 21](#_Toc463979326)

[**第六章** **研究結果** 22](#_Toc463979327)

[**第一節** **情緒辭典與模型的建立** 22](#_Toc463979328)

[**一、** **情緒加權值計算與分類結果** 22](#_Toc463979329)

[**二、** **新增停止詞與特殊詞彙前後差異** 23](#_Toc463979330)

[**三、** **回饋機制前後比較** 25](#_Toc463979331)

[**第二節** **預測新留言與模型回饋比較** 27](#_Toc463979332)

[**第七章** **操作畫面** 30](#_Toc463979333)

[**第八章** **結論** 36](#_Toc463979334)

[**參考文獻** 37](#_Toc463979335)

**圖目錄**

[圖 1 情緒分析詞彙分類 6](#_Toc463979336)

[圖 2 OpView政治民意 8](#_Toc463979337)

[圖 3 研究流程 9](#_Toc463979338)

[圖 4 粉絲團文章和留言 11](#_Toc463979339)

[圖 5 人工表記留言 11](#_Toc463979340)

[圖 6 斷詞詞彙 12](#_Toc463979341)

[圖 7 停止詞詞彙 12](#_Toc463979342)

[圖 8 環境圖(Context Diagram) 16](#_Toc463979343)

[圖 9 使用案例圖(Use Case Diagram) 18](#_Toc463979344)

[圖 10 類別圖(Class Diagram) 19](#_Toc463979345)

[圖 11 實體關聯圖(Entity Relationship Diagram) 20](#_Toc463979346)

[圖 12 系統架構圖**(**System Structure Diagram**)** 21](#_Toc463979347)

[圖 13 合併人物後的精準率 25](#_Toc463979348)

[圖 14 合併人物後的召回率 26](#_Toc463979349)

[圖 15 回饋前後的精準率 26](#_Toc463979350)

[圖 16 回饋前後的召回率 27](#_Toc463979351)

[圖 17 預測未知留言的回饋前後精準率 28](#_Toc463979352)

[圖 18 預測未知留言的回饋前後召回率 28](#_Toc463979353)

[圖 19 預測未知留言的回饋前後準確率 29](#_Toc463979354)

[圖 20 介面四大區塊頁面 30](#_Toc463979355)

[圖 21 圖表分析頁面 31](#_Toc463979356)

[圖 22 留言與回饋機制頁面 32](#_Toc463979357)

[圖 23 文章列表頁面 33](#_Toc463979358)

[圖 24 留言列表頁面 34](#_Toc463979359)

[圖 25 關鍵字頁面 35](#_Toc463979360)

**表目錄**

[表 1 斷詞前後範例 13](#_Toc463979361)

[表 2 對象一分析結果 22](#_Toc463979362)

[表 3 對象二分析結果 23](#_Toc463979363)

[表 4 對象一新增詞彙後 24](#_Toc463979364)

[表 5 對象二新增詞彙後 24](#_Toc463979365)

[表 6 對象一提升百分比 24](#_Toc463979366)

[表 7 對象二提升百分比 24](#_Toc463979367)

1. **緒論**
2. **研究背景**

最近由於無線網路的普及化，使用Facebook這類社群網站的人越來越多，網路更是知識的重要提供者之一。許多人不依靠電視、報紙，在臉書上獲取、分享訊息的動作也隨之增加，很多人也僅透過社群網站來了解最近的社會議題，由此可知社群網站已經成為部分人生活中的一部分。

隨著社群網站被越多人使用，社群網站所能影響的方面越來越廣，發揮出的力量也越來越不容小覷，這讓政府機關、社會企業也開始看重社群網站上所提出的意見。透過社群網站訊息的即時性、方便性，能讓使用者快速地取得他人對特定事件的看法，藉此了解民眾對文章、事件的看法，方便做出之後的決策。

1. **研究動機**

早期對網頁使用資料探勘技術進行文字探勘，演進至近年來熱門的研究領域－情緒分析，甚至已經有企業在使用相關技術的系統，以提高企業競爭力，可得知此研究有一定的深度以及重要性。

數量龐大的網路意見、民意資料如果使用傳統的人工來標記、分析文本類型所需耗費的人力、時間成本過高，加上閱讀大量資料可能造成的疲勞，將可能導致分類上的錯誤、速度上的減慢，同時資料的龐大數量也讓管理者也沒有足夠的時間一一去參考，但是對管理者來說每一份資料、意見都是珍貴的。

本研究繼承現有的論文[9]繼續延伸，根據其研究成果，將其作出一套系統並驗證其結果之正確性，若有錯誤我們希望能找出錯誤並加以改善。

1. **研究目的**

先前文獻[9]成果針對特定政治粉絲團，已經發展一套使用情緒分類技術中以TF-RF-IDF權重計算方式搭配SVM分群演算法來開發展情緒分析的系統。

本研究擬應用該項技術發展一套以情緒分析為基礎的輿情系統，來幫助使用者更容易做出正確的決策，並藉由使用者的回饋了解該分類技術是否可以有效套用至其他政治人物之粉絲團文章。

1. **文獻探討**
2. **網路文本情緒分析近況**

情緒分析研究的與日俱增， Pang and Lee 表示造成這種情況的理由有以下幾點：[6]

1. 資訊檢索的與自然語言處理的技術越來越成熟。
2. 全球資訊網的發展使網站與日俱增，讓機器學習演算法的訓練資料可以來自網站上的評論留言。

近年來社群網站的興起例如：FB、twitter等，使得人們的意見可以透過這些平台抒發給其他人知道，這些意見經由分析、歸納、推理後可以從中獲得具價值性的內容，例如：許多公司使用意見探勘和市場情緒分析去了解大眾輿論的趨勢，如產品評論，品牌認知度，和政治等相關問題[1]。

1. **情緒分析與意見探勘**

意見探勘與情緒分析的相似處都是要將原始資料歸納至特定的種類和兩者都能夠使用情緒辭典(Sentiment Lexicon) 來做為分類憑據，擁有一個合適的情緒辭典能有效提升分類準確度。[5]

意見探勘與情緒分析的差異:意見探勘(Opinion Mining)為某項特定物件各項屬性(Attributes)的意見之流程，而情緒分析則聚焦於判斷評論的情緒極性 (Polarity) 為正向 (Positive) 或負向 (Negative)，以整體文章或段落所表達的情緒作為分析目標，表明目標文本的情緒即可[5]。

外國學者Montoyo et al. (2012) 將情緒分析的發展現況整理，將此研究領域分成以下四種[7]：

1 來源建置(Creation of resources)：目的是建立具備字彙情緒標示的辭典或與料庫。

2.文字分類(Classification of text)：將文字依照所表達意見或情緒進行分類，可分成正向、負向、中性。

3.意見提取(Opinion extraction)：找出文件中表達情緒的部分，找出並決定該段落的情緒極性，最後歸納其來源之情緒。

4.情緒分析之應用(Applications of sentiment analysis):利用情緒分析開發一套系統，分析網站留言的情緒正負、趨勢預測系統等。

1. **辭彙情緒分析及建立情緒辭典**

此研究又包含以下子研究：主觀性分析（Subjectivity Analysis）、情緒極性分類（Polarity Classication）、情緒極性之強度辨識（Strength）。主觀性分析就是要先找出文本中個人情緒表達的主觀意見或是針對事實陳述的客觀描述，此很多研究可以從句子、段落或文件層級作為情緒判斷單位[6]。情緒極性分類目的是為了辨識詞語、句子、文章之感受，並將其依極性區分為正、負、中性三種，如圖1。

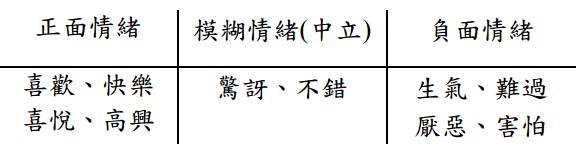


圖 1 情緒分析詞彙分類

目前建置情緒辭典多半是從資料庫蒐集大量辭彙，再以人工標記情緒詞及情緒詞彙。由於人工標記耗費時間、人力成本之高，因此發展出字典法與機器學習法兩種方法來建立情緒辭典，字典法是將人工標記完成 的情緒詞作為種子，再利用有語意架構的詞典（如WordNet、HowNet 等）進行 擴展比對，找出種子詞彙的同義詞或下位關係詞，全部標記為與種子詞彙相同的極性，便會產生收錄詞彙量較多的情緒辭典。

字典法能擴展的範圍有限，且相同 詞彙出現於不同主題，可能會產生不同的極性，字典法無法準確標記出極性。因此有研究開始嘗試利用機器學習法自動建置情緒辭典，以人工標記完成的情緒詞作為種子，蒐集網路上大量語料（如:新聞、部落格文章），再藉此計算各情緒辭彙在已人工標記分類好之文本中的權重值，若權重值為正數，歸類為正向情緒詞，反之則歸類為負向情緒詞，而趨近於零值則歸於中性情緒詞。然而，訓練資料不均衡容易造成機器學習結果偏頗。

1. **資料分類法**

目前常用的情緒分類方法可以分成監督式與非監督式兩種學習方法。使用在此領域的機器學習方法大多為監督式學習，因為處理的目標為自然語言以及人們的情緒，在非監督式下程式無法自主學習，必須要提供標註過的資料作為資料來源[5]。

此領域常用的方法有支援向量機（SVM）、最近鄰居法KNN、貝氏分類器 (Naïve Bayes)、最大熵學習法 (Maximum Entropy)，其中又以支援向量機（SVM）最為著名。

SVM的優點在解決小樣本、非線性及高維模式識別問題。Tan and Zhang（2008）對各種監督式學習法進行比較，發現支援向量機（SVM）分類效果較佳。Lan et al.（2005）提出了SVM的TF-RF權重計算方式並與其他方法比較，發現該方法具有最佳的準確率。Martineau and Finin（2009）也提出了SVM的DELTA TF-IDF權重計算方式，分類效能勝過了SVM的TF-IDF權重計算方法[3]。

1. **情緒分析應用**

近幾年來，各行各業逐漸重視民眾對特定議題的看法，在網路上也越來越多論壇能夠讓民眾表達自己的意見。如圖2，OpView是龍捲風公司所開發的產品[4]。情緒分析除了政治也能對產品、品牌、熱門話題等做出分析，範圍是相當多元的，由此可知情緒分析應用在這個充滿大量資訊的時代會更備受重視。



圖 2 OpView政治民意

1. **研究方法**
2. **研究流程**

首先以臉書兩位政治人物之粉絲團為例，本研究將陳菊稱作對象一，蔡英文稱作對象二，收集其粉絲團文章及留言當作我們的網路文本，並由人工標記其情緒類別作為後續研究。接下來針對區域語言特性，對所蒐集之文本建立特殊詞彙詞庫，結合R語言套件所提供的詞彙，接著運用R語言斷詞套件對文本進行斷詞分析。

本研究採用TF-RF-IDF權重計算方式對文本之情緒詞做權重計算，並以SVM分類法對網路文本進行情緒分類。藉由混淆矩陣探討研究結果，接著評估其精準率與召回率。最後將圖表、文本和情緒類別以網頁的方式呈現且系統會將使用者回饋後的留言詞彙與回饋前的留言詞彙合併並重新建立情緒辭典與模型，來增加系統準確率。

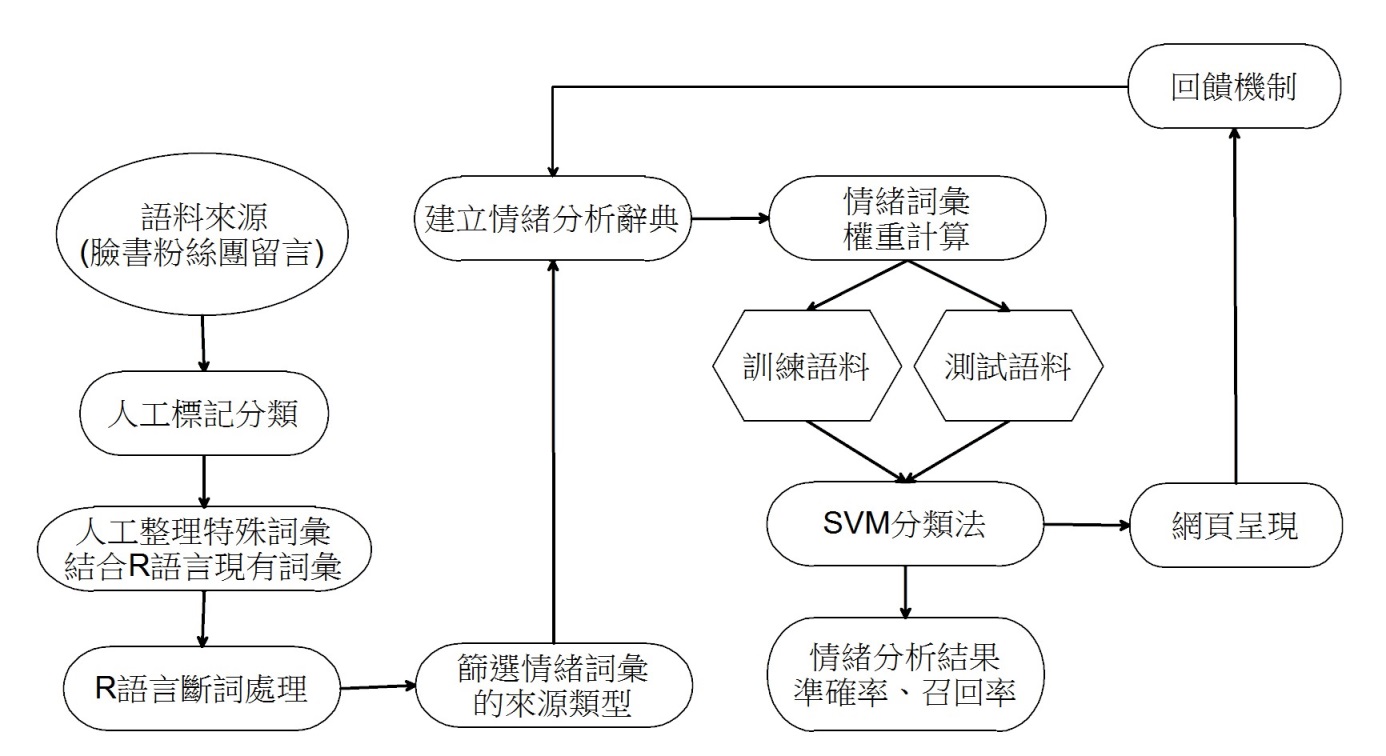


圖 3 研究流程

1. **使用工具**

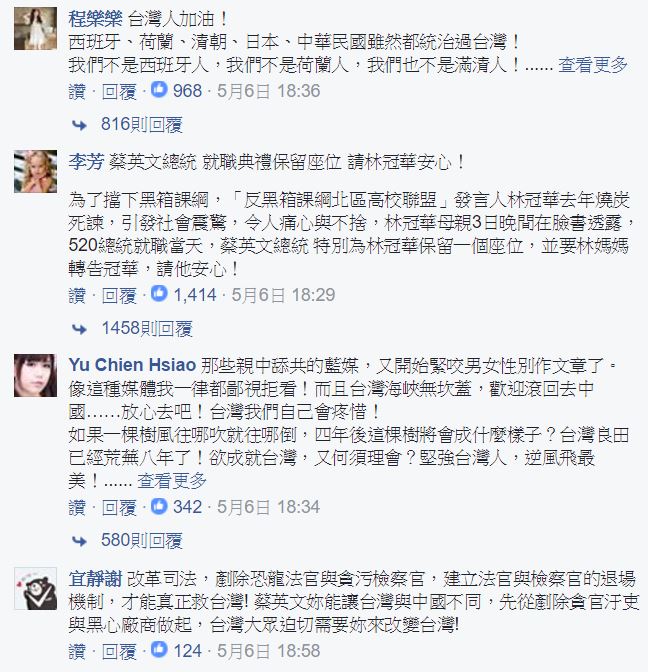
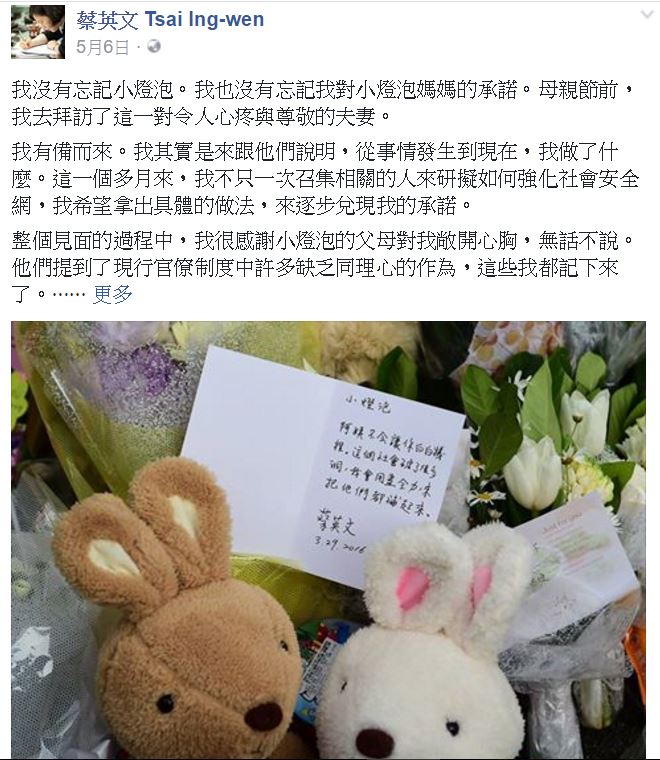
因R軟體具有強大的繪圖功能、多種分析套件等特色，因此本研究運用R語言作為資料分析的工具。我們選擇Python作為資料抓取的工具。網頁呈現則是以網頁伺服器Apache、資料庫軟體MySQL、後端程式語言PHP、Html、CSS、Javascript進行製作。

1. **語料收集**

大多數的人都會透過臉書發表自己的心情以及對某人、事、物的情緒與看法，因此本研究選用FACEBOOK作為語料來源，針對臉書粉絲團文章之留言分析其情緒傾向。選擇先前文獻[9]所研究的某政治粉絲團和當下總統的粉絲團相互驗證本系統對於政治領域的準確程度。

語料庫來源如圖4所示，以臉書粉絲團文章的留言作為本研究分析的語料庫。文獻[9]是從N篇貼文中，透過R語言程式篩選留言數最多的某篇貼文，再將其留言內容擷取下來作為語料庫，而本研究是從1月1日至7月31日的每一篇文章以500筆留言為上限並擷取至資料庫。擷取後對象一與對象二的粉絲團文章筆數與留言筆數分別為:對象一文章305筆，留言63106筆；對象二文章333筆，留言159144筆。

圖 4 粉絲團文章和留言



1. **前處理與人工標記文章**

首先將資料庫隨機挑選對象一1775筆留言與對象二1700筆留言並進行人工標記正、負面或無情緒傾向之類別，如圖5所示。再將留言中有出現英文、標點符號和空白字元過濾掉，以利後續分析。

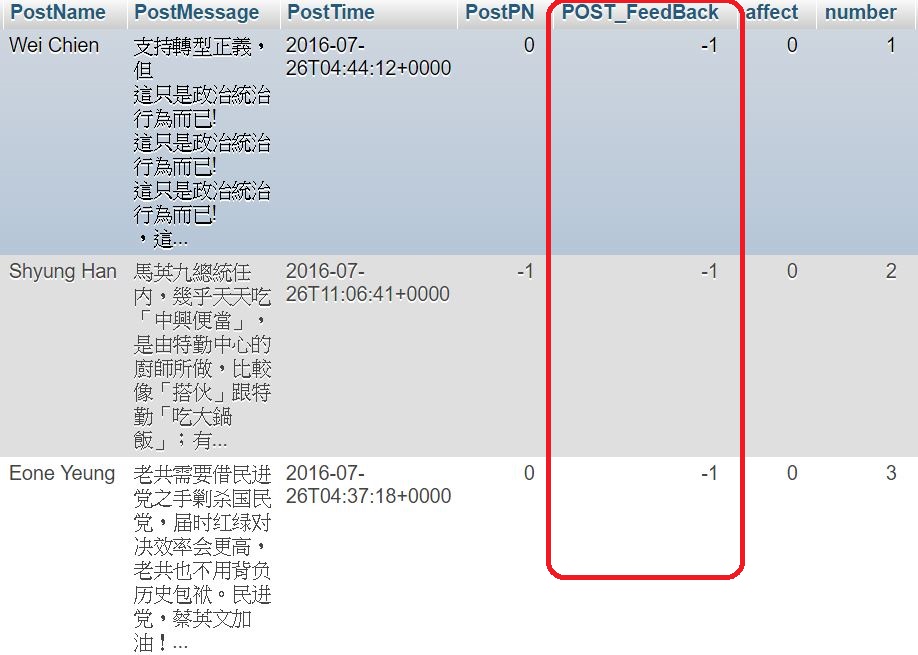


圖 5 人工表記留言

1. **詞彙篩選與建置情緒辭典**

由於R語言斷詞套件對於人名、地名、組織名和時事等未收錄之詞彙，會無法做出準確的斷詞結果。為了避免斷詞出來的詞彙不易辨識，故將R語言現有的詞彙加入部分特殊的詞彙一起斷詞，如圖6，再透過R語言斷詞套件將文本進行斷詞，以提升文本斷詞之準確度。

****

圖 6 斷詞詞彙

接下來，運用R語言中相關函數去除一些文本中出現頻率非常高，但對文本的意義沒有實質影響的冗詞贅字，如圖7，以利後續分析。



圖 7 停止詞詞彙

然而斷詞後的詞彙具有詞性，例如:動詞、名詞、形容詞、介係詞等等，為了有效擷取情緒詞彙，經由人工不斷的對詞彙的詞性進行檢視，本研究最後一元詞彙選用形容詞和名詞；二元或二元以上詞彙選用形容詞、動詞、名詞、副詞、名形詞、副形詞、動名詞作為情緒詞彙的可能來源，斷詞前後範例，如表1所示。

表 1 斷詞前後範例

|  |
| --- |
| 蔡總統：  死者為大！！  菩薩面前沒有國籍！！  身為領導人應有寬宏的度量！!  可是您於事發後，不僅表現對大陸罹難者的不聞不問也未出席，  更只對鄭姓臺灣導遊致送輓聯…是非不分的行為令人咋舌！！  可恥！！ |

|  |
| --- |
| 總統(n)  死者(n) 大(a)  菩薩(n) 沒有(v) 國籍(n)  領導人(n) 寬(a) 度量(n)  事發(v) 表現(v) 大陸(n) 罹難者(n) 不聞不問(i)  輓聯(n) 是非不分(userDefine) 行為(n) 人(n) 咋舌(v)  可恥(userDefine) |

1. **情緒加權值計算與留言情緒分類**

相同的詞彙在不同領域可能有不同的情緒，情緒分析的基本概念是當詞彙出現在正面文章的頻率高於其他文章情緒的頻率時，此詞彙便較偏向正面；反之則較偏向負面；或者為無情緒[9]。

本研究採用TF-RF-IDF權重方式

分析與計算文本中詞彙之情緒傾向程度。TF-RF-IDF權重方式是結合RF與TF-IDF的方法，()是以詞彙在該文章出現的數量除以該文章的詞彙總數，RF是以詞彙出現在正面文章數量()除以詞彙出現在負面文章()數量作為計算，IDF是文章的總數()除以含有該詞彙的文章數量()作為計算，使得IDF與RF相互影響。

每一個文本皆具有獨特的特徵向量，由詞彙的情緒權重值所組成。本研究藉由前述的權重計算方式對詞彙進行權重計算，建立文本特徵向量。再藉由支持向量機(SVM)分類演算法對文本進行分類，文本分類包含正面、負面、無情緒三種類別。將對象一1775筆留言和對象二1700筆留言各別從文本中隨機抓取80%作為訓練語料，20%作為測試語料，並將訓練語料導入SVM分類器進行模型訓練，最後藉由訓練出來的模型對測試語料進行分類。由於每次隨機取的測試語料都不盡相同，因此分類結果也會有所不同，故本研究採10次隨機取樣，對每次隨機取樣做分類，最後統計各組合方法之分類準確率的平均值。再評估何種組合方法對於文本情緒分類具有較高之精準率與召回率。所謂精準率是指預測為某一類別的文本中真正為該類別的比例；而召回率則表示真正為某一類別之文本有多少比例被預測出來。

1. **可行性分析**

此章會分析此系統各方面的可行性與風險。

1. **經濟可行性**

此研究皆使用免費軟體，有以下軟體：網頁伺服器Apache、資料庫軟體MySQL、 後端程式語言PHP、 統計分析軟體R語言、 Html、CSS、Javascript。硬體方面為使用自己的主機，作為網站伺服器。所以從經濟可行性角度考慮，本研究完全可行。

1. **技術可行性**

首先對此研究所用到的軟體進行分析，我們對於後端程式PHP與資料庫程式MySQL尚有經驗，但是統計軟體R語言未能運用自如、對於網路爬文技術尚無經驗、對R語言和PHP之間的整合尚未清楚了解。對於不熟悉的R語言軟體會更深入的學習，爬文技術則會上網搜尋相關研究。簡而言之，我們會針對R語言跟PHP語言更深入學習，因此技術層面是可行的。

1. **系統分析與設計**

在此針對使用者的需求做詳細的分析。我們使用環境圖(Context Diagram)定義本系統和外部實體的關聯性。案例圖（Use Case Diagram）定義此系統的功能性需求，界定功能範圍。類別圖(Class Diagram)了解系統中類別的屬性，以及屬性之間的關聯性。實體關聯圖(Entity Relationship Diagram)來描述系統資料間的關係。系統架構圖(System Structure Diagram)描述本系統和在實體機器和伺服器的關聯。

1. **環境圖(**Context Diagram**)**

由此環境圖如圖8 能更清楚瞭解到本系統和外部實體的關聯性，Facebook為外部實體，由本系統輸出人物id和文章數量，再由Facebook為外部實體傳回文章資訊。

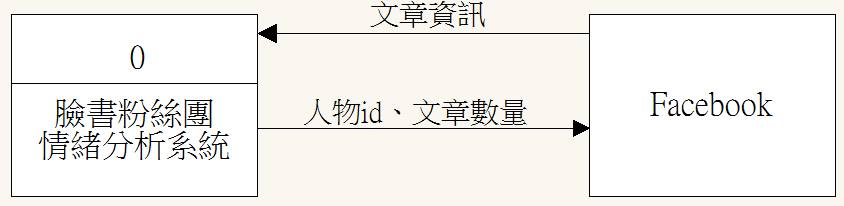


圖 8 環境圖(Context Diagram)

1. **使用案例圖(**Use Case Diagram**)**

我們透過使用案例圖（Use Case Diagram）對系統的功能進行說明，如圖9 所示。而功能的具體說明如下：

**選擇人物：**

使用者可以對網站所提供的人物選項進行點選來觀看圖表、文章、留言、關鍵字。例如：對象一、對象二。

**選取時間軸：**

使用者可以選取網站提供的時間或自訂義時間範圍，來查詢此人在不同時間點的分析結果。

**圖表輸出:**

以折線圖、長條圖、圓餅圖呈現，以便使用者能夠輕易了解所選擇的時間內正面、負面、無情緒留言的筆數。

**回饋機制:**

當使用者發現系統所預測的留言情緒有錯誤時，能夠選擇情緒並回饋給系統，幫助系統改善準確率。

**關鍵字：**

針對系統提供的Facebook粉絲團人物，統計所選的時間內詞彙的出現頻率，以文字雲圖的方式輸出至頁面上，並可點選關鍵字來觀看詞彙出現在那些留言。

**抓取資料：**

針對Facebook粉絲團人物，針對所選的時間去抓取貼文和留言。

**建立情緒模型：**

針對新的資料去做情緒的預測，分成正、負面和無情緒。

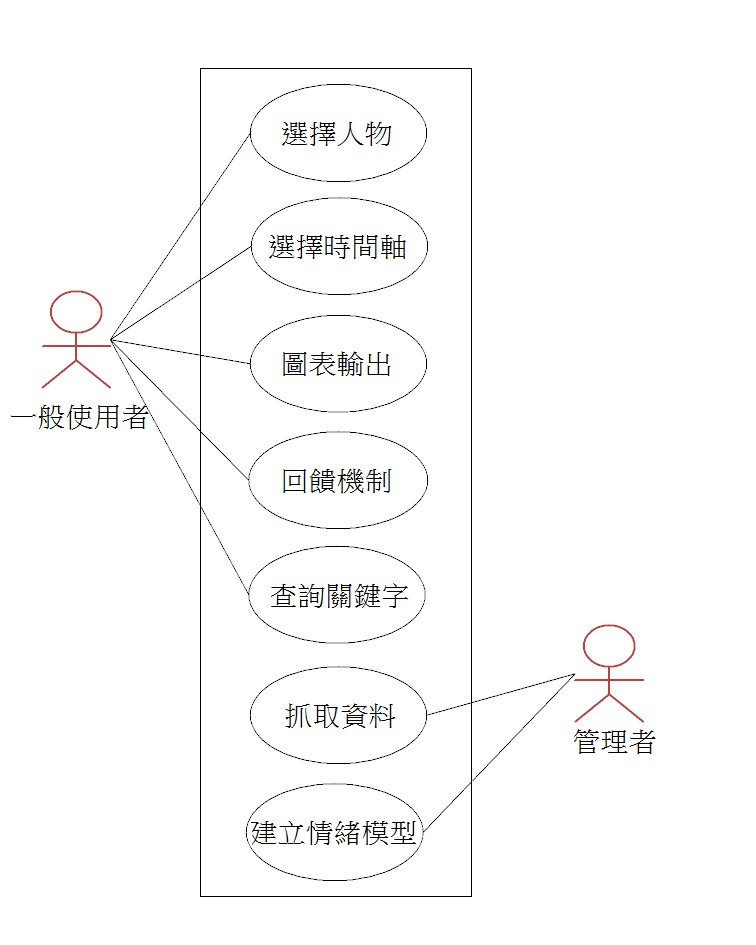
****

圖 9 使用案例圖(Use Case Diagram)

1. **類別圖(**Class Diagram**)**

本研究使用類別圖(Class Diagram)描述系統的類別集合、類別屬性以及類別之間的關係，為之後程式的撰寫提供參考。

如圖10 所示，頁面顯示負責接收資料並顯示於頁面上。文章類別、留言類別分別對資料庫讀取文章和留言資訊。留言正負無數量圖類別負責把使用者所選取的時間以折線圖、長條圖、圓餅圖，關鍵字顯示。關鍵字類別負責顯示每週留言裡詞彙出現次數並顯示排行，使用者能夠點選詞彙來觀看詞彙出現在那些留言裡。回饋機制類別負責接收使用者對於留言情緒的回饋並修改該留言的情緒值。

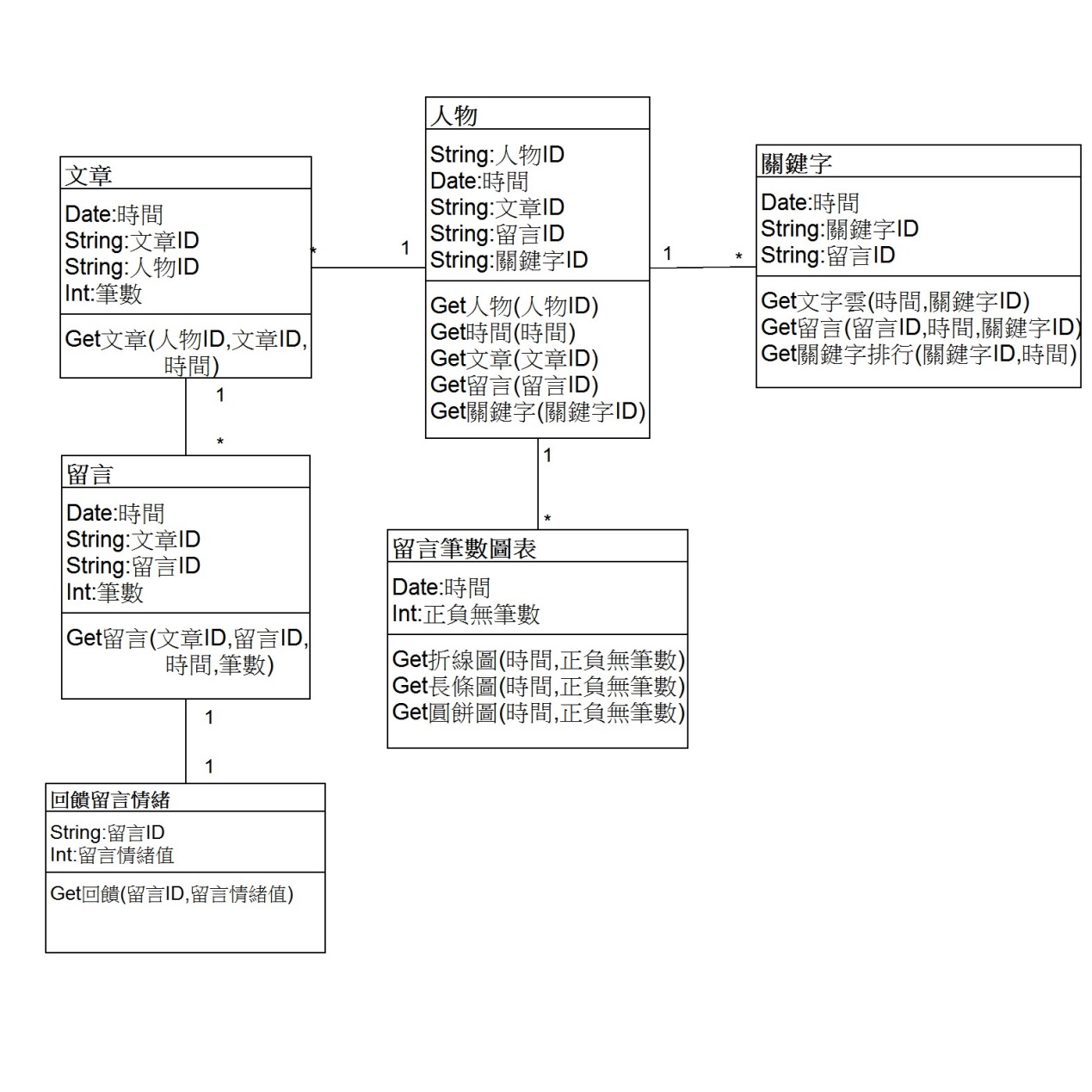


圖 10 類別圖(Class Diagram)

1. **實體關聯圖(**Entity Relationship Diagram**)**

本章節使用實體關聯圖(Entity Relationship Diagram)來描述系統資料間的關係，如圖11所示，為之後程式的撰寫提供參考。本系統共有六個實體，分別是Page、Post、Key\_Post、Keyword、idf\_、Date，Page存文章的屬性，Post存留言的屬性，以PageId當作Page和Post的Primary Key；Key\_Post存和每筆留言的關鍵字屬性，以PostId當作Key\_Post和Post的Primary Key；Keyword存每個關鍵字ID的詞性屬性，以WordId當作Key\_Post和Keyword的Primary Key；idf存關鍵字的情緒值的屬性，以WordId當作idf和Keyword的Primary Key；Date存每日情緒值的屬性，以PostName當作Date和Post的Primary Key。

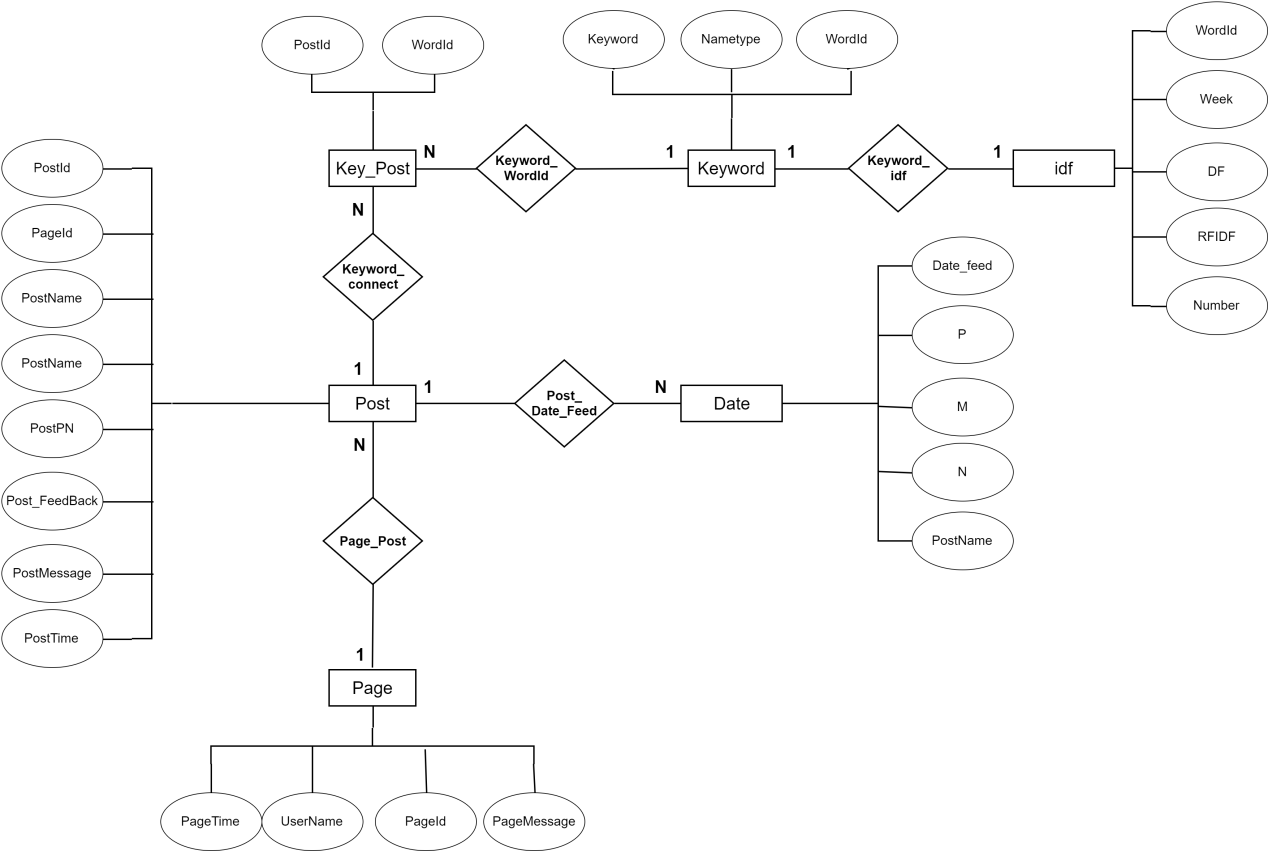


圖 11 實體關聯圖(Entity Relationship Diagram)

1. **系統架構圖(**System Structure Diagram**)**

如圖12所示，左半部為一般使用者和管理者，管理者可做一般使用者所擁有的功能、資料抓取和建立情緒模型，一般使用者可操作本系統一般使用者頁面；中間則是系統內部所使用到的語言；右半部為後端，資料庫存放系統資料，資料則是透過facebook伺服器端傳給本系統。

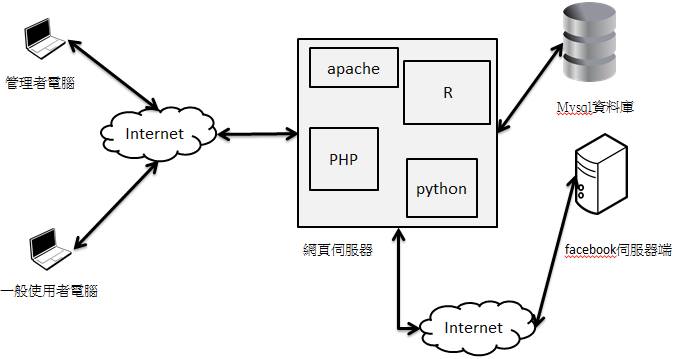


圖 12 系統架構圖**(**System Structure Diagram**)**

1. **研究結果**

在第一節將分析情緒加權值計算與分類結果和改善方式；第二節運用改善後的方法建立情緒辭典與模型，來預測新留言(未包含在情緒辭典中的詞彙)並分析結果。

1. **情緒辭典與模型的建立**
2. **情緒加權值計算與分類結果**

本章針對研究所得資料提出進一步的分析，首先透過TF-RF-IDF權重方式對文本詞彙進行權重計算，接著分別採用SVM分類法對文本進行情緒傾向分類。

文獻[9]是從對象二的N篇貼文中，透過R語言程式篩選留言數最多的某篇貼文，將其留言作為語料庫，而本研究則是針對全部的留言去作探討。文獻[9]只有針對對象二的199筆留言去做探討，筆數略有不足。因此本研究決定擴增筆數，最後以對象一1775筆留言和對象二1700筆留言去預測精準率、召回率、準確率和調和平均數，再以10次的平均值作為結果，如表2和表3。

因調和平均數可以看出精準率和召回率兩個數值的差異大小，也就是說當兩個數值的百分比不低且差異不高，調和平均數相對就比較高，反之亦然。因此在此本研究主要針對正面和負面去作探討，在第二節將會說明對象一負面的調和平均數和對象二正面的調和平均數做改善。

表 2 對象一分析結果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 精準率 | 召回率 | 調和平均數 | 準確率 |
| 正 | 80.36% | 92.25% | 85.91% | 76.65% |
| 負 | 62.84% | 62.31% | 62.57% |
| 無 | 64.18% | 40.72% | 49.82% |

表 3 對象二分析結果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 精準率 | 召回率 | 調和平均數 | 準確率 |
| 正 | 62.17% | 46.48% | 53.19% | 68.01% |
| 負 | 59.58% | 76.69% | 67.06% |
| 無 | 75.30% | 70.52% | 72.83% |

1. **新增停止詞與特殊詞彙前後差異**

在此節將說明新增停止詞和特殊詞彙對情緒值的影響和調和平均數改善後的結果，本系統為了提高正、負的精準率、召回率、準確率和調和平均數，因此決定觀察正面和負面的留言並對其留言找出相關停止詞和新詞彙進行改善，表4和表5為本研究新增停止詞與特殊詞彙結果。

停止詞:

針對留言內常出現卻沒有實質影響的冗詞贅字，例如:啦、吧、或等，找出原本詞庫中沒有的詞彙加入停止詞，並將中科院計算所中文自然語言處理開放平台的所提供的詞彙一起加入停止詞庫，共新增1500個詞彙，停止詞的總詞彙量共1747個。

特殊詞彙:

為了讓斷詞後的詞彙更加準確，除了新增停止詞之外，也將針對留言內的特殊詞彙做新增，例如:女童、割喉等，共新增50個時事詞彙，特殊詞彙的總詞彙量共10726個。

表 4 對象一新增詞彙後

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 精準率 | 召回率 | 調和平均數 | 準確率 |
| 正 | 82.13% | 91.73% | 86.66% | 78.62% |
| 負 | 68.25% | 65.50% | 66.85% |
| 無 | 67.94% | 47.55% | 55.94% |

表 5 對象二新增詞彙後

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 精準率 | 召回率 | 調和平均數 | 準確率 |
| 正 | 64.17% | 55.36% | 59.44% | 69.00% |
| 負 | 58.46% | 80.10% | 67.59% |
| 無 | 78.46% | 67.87% | 72.78% |

由表6和表7可觀察出新增停止詞和特殊詞彙後，對象一或對象二的精準率、召回率、準確率和調和平均數大多都有明顯的提升，因此停止詞和新詞彙對於判斷情緒有很大的影響。

表 6 對象一提升百分比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 精準率 | 召回率 | 調和平均數 | 準確率 |
| 正 | 2.20% | -0.56% | 0.87% | 2.57% |
| 負 | 8.61% | 5.12% | 6.84% |

表 7 對象二提升百分比

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 精準率 | 召回率 | 調和平均數 | 準確率 |
| 正 | 3.22% | 19.10% | 11.75% | 1.46% |
| 負 | -1.88% | 4.45% | 0.79% |

1. **回饋機制前後比較**

為了將本研究做成系統，必須克服人工判讀的方式去標記留言情緒值，因此決定將對象一和對象二的訓練語料集做合併，並以此訓練語料集去做回饋機制的前後比較。

如圖13和圖14所示，比較對象一、對象二和合併後差異，再以10次的平均值作為結果。合併後的精準率在正面比對象一和對象二更高，負面稍低一些，無情緒則是在對象一和對象二之間。合併後的召回率不論在正面、負面和無情緒的數值都非常接近對象一或對象二的最佳值。合併後的數值跟對象一和對象二並無太大的出入，因此本系統將對象一1775筆留言與對象二1700筆留言合併並作為訓練語料集，再隨機從資料庫抽取1100筆留言作為回饋資料，並以客觀的方式做情緒標記，觀察回饋機制對於系統精準率與召回率的影響。

圖 13 合併人物後的精準率

圖 14 合併人物後的召回率

如圖15和圖16所示，雖然新增回饋語料後正面的精準率有變低，但是在負面和無情緒有了明顯的提升。在召回率方面，負面和無情緒有稍微提升，正面是持平的狀態，在整體上精準率和召回率是提升的。由此可知，當語料集變大時，在預測時也能夠更加準確。

圖 15 回饋前後的精準率

圖 16 回饋前後的召回率

1. **預測新留言與模型回饋比較**

使用本研究的方法對新留言做正、負面和無情緒的預測，比較回饋前後對於預測新留言的差異。回饋前以對象一1775筆留言加上對象二1700筆留言隨機抽取80%作為訓練語料，回饋後以對象一1775筆留言加上對象二1700筆留言再加上回饋後的1100筆留言加總後隨機抽取80%作為訓練語料，預測對象一粉絲團1月~7月的留言。並且從1月~7月，每個月隨機抓取100筆留言，共700筆留言作為判斷精準率、召回率和準確率的依據，如圖17、18、19所示。

精準率如圖17所示，回饋後無情緒下降約15%，正面的精準率上升約10%，負面的精準率提升約30%。加入1100筆回饋後對於新留言的正面和負面片段的更加精準。

圖 17 預測未知留言的回饋前後精準率

召回率如圖18所示，回饋後不論正面、負面或無情緒都有所上升，因此加入1100筆對新留言的召回率改善許多。

圖 18 預測未知留言的回饋前後召回率

從圖19可看出回饋前後準確率的差異，做回饋前預測新留言準確率為53.9%，回饋後為61.4%。由此可知回饋後的語料集與情緒辭典對於精準率、召回率和準確率影響很大，因此回饋能夠幫助系統判斷情緒更加精準。

圖 19 預測未知留言的回饋前後準確率

1. **操作畫面**

此章節將說明本系統頁面的操作方式，分成四大區塊:首頁、圖表分析、文章列表、關鍵字，如圖20所示。



圖 20 介面四大區塊頁面

點選首頁的人物列表後進入圖表分析，使用者可選擇一段時間，來觀看正、負和無情緒的圖表，如圖21所示。使用者可點選折線圖上的點座標來觀看當天的留言內容和情緒值，也可使用回饋機制進行回饋，如圖22所示。

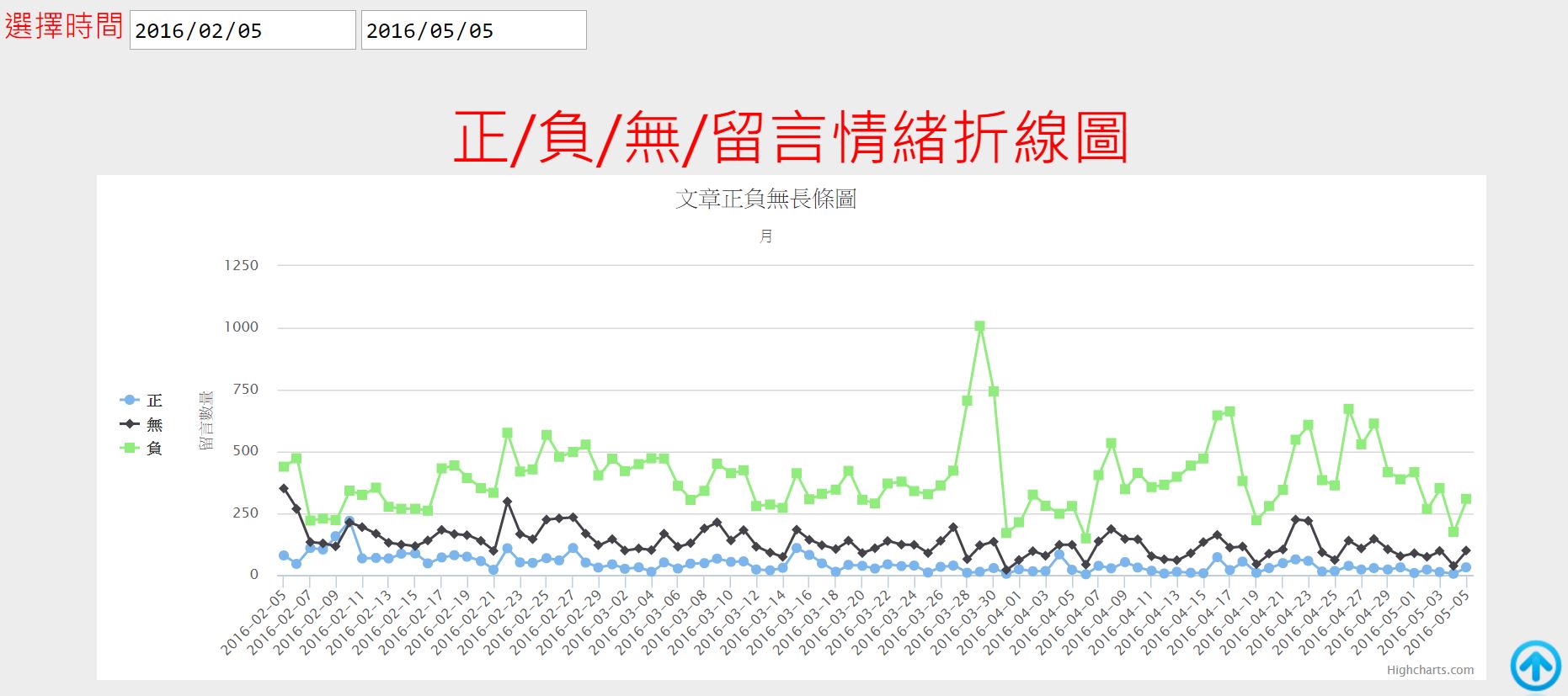
****

圖 21 圖表分析頁面



圖 22 留言與回饋機制頁面

如圖23所示，點選文章列表使用者可進行月數選擇，觀看該名人物的貼文、時間，點選右邊的留言按鈕可觀看該貼文的留言，如圖24所示，使用者也可在此使用回饋機制進行回饋。



圖 23 文章列表頁面



圖 24 留言列表頁面

點選關鍵字列表將可以對週數和情緒值進行選擇，如圖25所示，左邊以文字雲的方式呈現關鍵詞彙，右邊以表格方式呈現關鍵詞彙並可點選觀看文章留言，使用者也可在此使用回饋機制進行回饋。



圖 25 關鍵字頁面

1. **結論**

本研究選擇輿論多樣性最高的Facebook，透過其粉絲專業文本中的訪客留言，作為研究分析語料的語料來源。我們將這些留言斷詞之後並使用TF-RF-IDF對文字做情緒詞權重計算，接著使用SVM分群演算法將語句分類。且經過改善詞彙篩選的方法、新增斷詞詞彙、新增停止詞、新增訓練語料集和回饋機制的方式，本系統對於留言情緒判斷的準確度有明顯的改善。

我們的系統目標是能夠清楚了解公眾人物的言論在大眾眼中的看法，於是我們在網頁中加入了許多功能，例如：依據時間跟人物進行繪製留言情緒筆數圖、文字雲等。本系統目前只有抓取對象一和對象二的**粉絲專頁做分析，期望未來能夠新增更多政治人物給使用者作選擇。**最後，期望未來能夠以本研究結果為基礎，**應用本方法於其他領域之粉絲專頁**，並改善抓取文章和分析文章的時間，且能夠新增更多功能來完善本系統。

**參考文獻**

1. 黃信華。以FACEBOOK塗鴉牆文本分析情緒文字的關係。樹德科技大學資訊工程系，頁1-8，2013。
2. 蘇怡仁、余碩文、黃皇瑋、陳岳群。基於字典樹改善中文情緒分類效能之研究。TANET，頁1-6，2014。
3. 游和正、黃挺豪、陳信希 (2012)。領域相關詞彙極性分析及文件情緒分類之研究。http://www.cs.cmu.edu/~tinghaoh/pdf/2012\_ijclclp.pdf
4. 張建文(2013)。OpView社群口碑資料庫。

http://www.cybersoft4u.com/cindy/%E6%84%8F%E8%97%8D%E9%BE%8D%E6%8D%B2%E9%A2%A8\_OpView%E6%9C%8D%E5%8B%99%E7%B0%A1%E4%BB%8B\_20130731.pdf

1. 李柏霖。社群事件情緒分析 與商品評論見解探勘之研究。國 立 高 雄 第 一 科 技 大 學，頁18，2015。
2. B. Pang and L. Lee. “Opinion mining and sentiment analysis,“ Foundations and Trends in Information Retrieval, 2008
3. Montoyo, A., Martínez-Barco, P., and Balahur, A. (2012), “Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments,” Decision Support Systems, 53(4), pp. 675-679.
4. Tan, S. & Zhang, J. (2008). An empirical study of sentiment analysis for Chinese documents. Expert System with Applications, 34(4), 2622-2629.
5. 趙妤瑄，王豐緒。情緒詞權重計算與分類演算法對於情緒分析結果之影響－以臉書粉絲團議題分析為例。第八屆臺灣商管與資訊研討會，頁1-18，2016