

國立臺灣師範大學 資訊工程研究所碩士論文

指導教授：侯文娟 博士

利用臉書資訊探討網路新聞的吸引度
及極性分析

News Attraction and Polarity Analysis
Using Facebook Information

研究生： 楊登堯 撰

中華民國 一〇六年 七月

摘要

過去人們獲取資訊的途徑只有從談話、書籍、報章雜誌等媒體，資訊量的收集速度緩慢且數量有限，然而現今網路的發達以及科技改良所賜，網路的方便性及發達帶給了這個社會資訊化。

社群網站的興起(例如:facebook、twitter)，讓許多人開始透過這些網路平台，迅速傳播新聞資訊或就生活上的知識進行交流與溝通。報紙雜誌等傳統媒體，也開始透過網路平台進行發佈。

然而在資訊爆炸的時代，人們該如何從這些大量的報導中獲取想要或者喜歡的資訊，而媒體又該如何從新聞內文中適當的撰寫以便吸引閱聽人，讓人們能夠喜歡閱讀該報導，並且可以從這些報導當中，發掘現今人們的新聞喜好傾向，這些都是目前值得研究者探討且著墨的地方。

本研究將首先利用情緒分析的技術，分析現在網路新聞內文經常使用哪些詞彙或語句，可激發閱聽人的情緒反應以增加其閱讀興趣。其次，為了更進一步了解新聞極性的趨勢，也就是正向的新聞比較受歡迎還是負向的新聞比較受歡迎，會先進行斷詞之後，利用 TF-IDF 值尋找出關鍵字，然後利用語料庫進行比對，得到正向詞與負向詞的資訊，接著再利用 Facebook 提供的讚數當作佐證，就可以看出現在人們是喜歡哪一類的新聞。

研究結果發現，閱聽者比較常關注負向新聞，並且本研究利用圖斯勒(Marc

Trussler)和索羅卡(Stuart Soroka)在加拿大麥基爾大學(McGill University)的實驗結果相比對，顯示和從心理系角度所做的研究，有相符的結果，進而佐證本研究的可信度。



關鍵字:自然語言處理，情感分析，中文剖析器，語意字典

Abstract

In the past, the way of people obtaining information is only from the conversation, books, newspapers and other media information collection, which is slow and limited in number. However, due to the development of network technologies, in the present, the vast amount of information can be retrieved conveniently from internet.

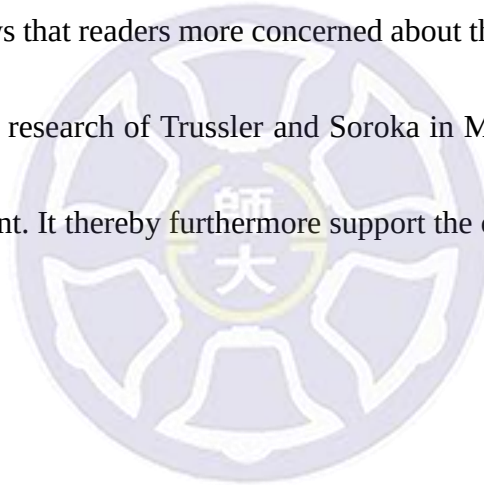
Some community web sites (*such as* facebook , twitter)make many people start with these network platforms for the rapid dissemination of news or for exchange of knowledge on life. Traditional newspapers, magazines and other traditional media also begin to publish their reports on the network platform.

In an era of information explosion, how people can get information that they want or like from these extensive reports and how news can attract readers are worthy of investigation. Furthermore, the preference tendency of negative news or positive news is also very important.

The study will first take advantage of the sentiment analysis technology to analyze network news by extracting the frequently-used words or phrases in order to increase

people's interest in reading. Second, to further understand the trend of news polarity, that is, whether positive news is more popular than negative news or not. The study segments words, finds keywords using TF-IDF values, and then matches keywords with a semantic dictionary in order to get the polarity informations. Finally, use the number of 'like' provided by Facebook as corroboration, the trend of news polarity that people like is shown.

The study shows that readers more concerned about the negative news. Comparing to the psychological research of Trussler and Soroka in Macgill University in Canada, the result is consistent. It thereby furthermore support the confidence of this study.



Keywords: NLP, sentiment analysis, Chinese parser, semantic dictionary

謹獻給

我的家人，指導教授，以及所有幫助過的人



致謝

本論文得以完成，首先要感謝我的指導教授侯文娟老師，在這兩年來引導我做研究。讓我能從零開始，一步步踏進自然語言處理的領域中；在研究遇到瓶頸時，老師也給我許多幫助和提點，讓我能突破困難，見識到不同的眼界。此外老師也提供了良好的研究環境讓我們能夠心無旁騖的投入研究，獲得許多寶貴的知識和追求學問的方法，此外向老師致上最高的謝意。

感謝我的家人這兩年來的陪伴與鼓勵，讓我能夠無後顧之憂的完成學業，讓我在迎向未來的挑戰實更有勇氣，更有自信面對。



目錄

摘要.....	I
Abstract	III
目錄.....	VII
附表目錄.....	VIII
附圖目錄.....	IX
第一章 簡介.....	1
第一節 研究背景.....	1
第二節 研究動機.....	1
第三節 研究目的.....	2
第四節 論文組織.....	3
第二章 相關研究探討.....	4
第一節 大數據.....	4
第二節 斷詞系統.....	6
第三節 極性分類 (polarity classification)	8
第四節 情感分析 (Sentiment Analysis)	9
第五節 NTUSD	10
第六節 TF-IDF	11
第三章 研究方法.....	14
第一節 緒論.....	14
第二節 研究資料.....	14
第三節 研究方法與架構.....	29
第四節 研究方法描述.....	32
第四章 實驗結果與分析.....	36
第一節 抓取 facebook 上蘋果新聞資料.....	36
第二節 內文進行斷詞.....	39
第三節 新增高吸引度與低吸引度語料庫以及擴充字典.....	40
第四節 TF-IDF 在新聞內文裡面的關鍵字	56
第五章 結論與未來展望.....	65
參考文獻.....	67

附表目錄

表 1 NTUSD positive 部分詞彙.....	11
表 2 NTUSD negative 部分詞彙.....	11
表 3 中文斷詞系統詞性表.....	22
表 4 高吸引度所用的詞彙以及極性(節錄).....	42
表 5 低吸引度所用的詞彙以及極性(節錄).....	43
表 6 擴充新聞詞彙.....	44
表 7 計算高吸引所有詞彙出現的詞性機率.....	46
表 8 計算低吸引所有詞彙出現的詞性機率.....	47
表 9 高吸引度實驗結果.....	53
表 10 低吸引度實驗結果.....	55
表 11 TF-IDF 之新聞判斷正負向(取出高吸引新聞部分).....	58



附圖目錄

圖 1 大數據定義.....	5
圖 2 蘋果粉絲專業.....	15
圖 3 爬蟲程式(節錄).....	15
圖 4 爬取新聞內文.....	16
圖 5 從粉絲專頁得知讚數多寡.....	17
圖 6 graph API 資料的指令.....	18
圖 7 graph API 資料的指令(讚數).....	18
圖 8 中文斷詞系統之輸出範例.....	21
圖 9 斷詞完之後的詞彙與詞性特徵及頻率次數.....	24
圖 10 教育部重編國語辭典修訂本.....	27
圖 11 相似詞相反詞的結果輸出.....	27
圖 12 "完美"一詞在 NTUSD 所歸類的極性.....	28
圖 13 擴充辭典並且找出吸引度高低所使用的詞性為何.....	30
圖 14 現代人喜歡負面新聞或者正向新聞.....	31
圖 15 抓取 facebook 上的資料.....	37
圖 16 抓取新聞內文.....	38
圖 17 新聞內文進行斷詞.....	39
圖 18 教育部重編國語辭典修訂本輸出結果.....	41
圖 19 NTUSD 負向詞彙.....	41
圖 20 NTUSD 正向詞彙.....	42
圖 21 新聞寫作之寫作技巧(ref:吳建華老師之新聞寫作技巧).....	48

第一章 簡介

第一節 研究背景

過去人們獲取資訊的途徑只有從談話、書籍、報章雜誌等媒體，資訊量的收集速度緩慢且數量有限，然而現今網路的發達以及科技改良所賜，網路的方便性及發達帶給了這個社會資訊化。社群網站的興起(例如:facebook、twitter)，讓許多人開始透過這些的網路平台，迅速傳播新聞資訊或一些生活上知識的交流進行溝通。傳統報紙雜誌等一些傳統媒體，也開始透過網路平台去發佈。人們開始透過手機看線上新聞報導，將這些新聞報導資訊化、數位化，每一個資訊、每一則新聞、每一張圖片、每一個影片，都可以藉著網頁呈現。透過網際網路，人們可以隨時隨地閱讀文章或者新聞，不用到超商買各家的報紙以便了解生活時事，不用到書局買一本本的書籍獲取知識。

然而在資訊爆炸的時代，人們該如何從這些大量的報導中獲取想要或者喜歡的資訊，而媒體又該如何從新聞內文中適當的撰寫以便吸引閱聽人，讓人們能夠喜歡閱讀你這家報導，並且可以從這些報導當中，發掘現今人們是喜歡比較負面的新聞或者正向新聞，這些都是目前值得研究者探討且著墨的地方。

第二節 研究動機

內文是整個文章的脊髓，一則新聞想要表達些什麼，必須從文章內容得知，

針對新聞媒體的數位化以及資訊的巨量化，如何在內文表達時，能夠兼顧引起讀者的相關興趣，以及知道現在讀者喜歡怎樣的報導，是負向報導比較有興趣或者是正向報導比較有興趣，都是新聞內文的作用。

一篇網路新聞(以蘋果新聞為例)字數大約落在 300 字左右，一篇網路新聞不可能寫得很冗長，畢竟版面有限，那如何在有限的字數中，用一些吸引的字詞在新聞文章中吸引閱聽人，激發閱聽人的興趣，是值得研究的方向。此外，閱讀線上新聞的人越來越多，已經不分年幼老少，那麼年輕人或年長者是喜歡偏正向的新聞或者喜歡一些負向新聞，這也是一大課題。

如果在一個新聞平台中，報導的文章可吸引更多讀者，將讓他們自家的品牌報導更加受歡迎。所以許多新聞媒體採用的手法便是運用一些誇大性、或者情緒性的字詞，嘗試誘發讀者的情緒反應，激發讀者的閱讀興趣，以便增加網路新聞的點閱率與訂閱率，進而曝光該網路新聞媒體的存在性。

第三節 研究目的

現在網路上，各家的報導齊聚一堂，每天網路新聞量十分的龐大，如果是每日發行報紙的新聞來說，可能就有好幾百則，更不用說隨時可能會有更新或是即時新聞。所以現在新聞媒體對於文章的撰寫也開始漸漸誇大，且越來越趨於習慣使用一些情緒激發性字眼，來增進網路新聞的點擊率，進而增加他們的曝光度。

再者，網路新聞的主題種類繁多，所以針對不同主題的內容，新聞媒體慣用的情緒煽動詞彙也會不太一樣。因此本研究將首先利用情緒分析的技術，分析現在網路新聞內文經常使用哪些詞彙或語句，可激發閱聽人的情緒反應以增加其閱讀興趣。其次，為了更進一步了解新聞極性的趨勢，也就是正向的新聞比較受歡迎還是負向的新聞比較受歡迎。會先進行斷詞之後，利用 TF-IDF 值，尋找出關鍵字，然後利用語料庫進行比對，得到是正向詞或者負向詞的資訊，接著再利用 Facebook 提供的讚數當作佐證，就可以看出現在人們是喜歡哪一類的新聞。

故本研究旨在探討網路新聞內文是否能夠誘發閱聽人的情緒來激發他們的興趣，進而再探討現在人們是喜歡哪一面向的新聞(正向或負向)，最後探討網路新聞與商業之間的關係。

第四節 論文組織

本論文的組織架構如下:第一章為簡介，說明研究動機與目的;第二章為相關研究的討論，介紹該領域相關的研究背景;第三章則是提出本次研究的架構，並詳細說明本論文所提出的研究方法;第四章為實驗過程與實驗結果;第五章將會說明本論文的結論以及未來發展方向。

第二章 相關研究探討

本章節整理相關文獻，作為本研究方法的理論基礎，第一節為大數據，介紹大數據帶給怎樣的商機以及帶動科技的風潮。第二節將介紹斷詞，說明如何運用斷詞系統斷出的句子或者詞。第三節將介紹極性分類 (polarity classification)，用於分類文章為正極性 (贊同)或是負極性 (反對)，第四節將介紹情感分析，說明根據研究目標制定的需求，所使用的方法與結果。第五節將介紹語料資料庫 NTUSD。第六節將介紹 TF-IDF，用於評估一字詞對於一個文件及或者一個語料庫中其中一份文件的重要程度。

第一節 大數據

隨著網際網路的發展以及行動載具的普及，網路上的數據內容也跟著大量成長，成長的幅度甚至到了巨量的地步。而這些大量數據內容通常被稱為「大數據」(Bollier, 2010)。而大數據又可以稱為海量資料、巨量資料、大資料，普遍的定義是在於儲存數量的價值(value)，可以由大量(volume)、速度(velocity)、和種類(variety)構成全面性的定義(Russim, 2011)(如圖 1)。

一般而言，大數據的定義是 Volume (容量)、Velocity (速度) 和 Variety (多樣性)，但也有人另外加上 Veracity (真實性) 和 Value (價值) 兩個 V。但其實不論是幾 V，大數據的資料特質和傳統資料最大的不同是，資料來源多元、種類繁多，大多是非結構化資料，而且更新速度非常快，導致資料量大增。而要用大數

據創造價值，不得不注意數據的真實性。

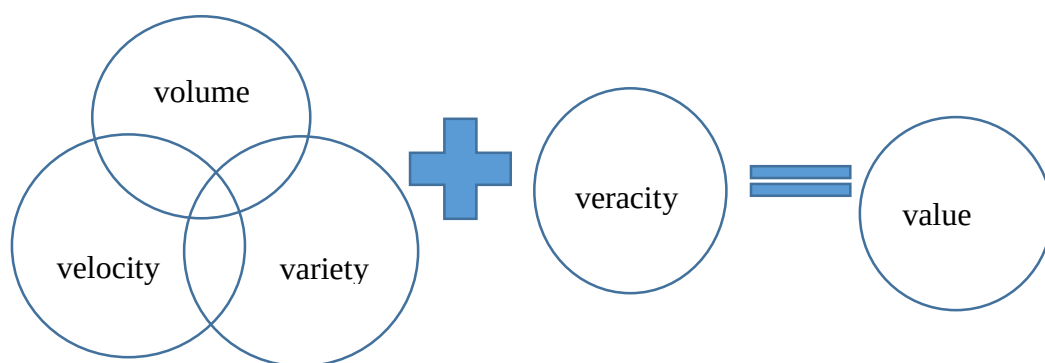


圖 1 大數據定義

隨著大數據降臨，如何對於這些數據資訊進行篩選便是值得討論的課題。因為藉由相關資料的分析，可以篩選出許多有用的資訊，探索網路資訊與社會的脈動連結，甚至是消費者行為相關趨勢(Bollier, 2010; 葉乃嘉, 2013)。大數據為資訊社會帶來全面性的影響，甚至捲起一個大數據風潮，包括個人行為、商業模式、生活型態、策略與管理改變、社會與國家改變(趙麗慧, 2014)。因此許多政府組織、領域專家學者也都開始進行大數據分析的研究，以便了解社會趨勢或者民眾行為，藉此進行策略上的制定。

在商業研究上，Boyd 及 Crawford(2012)的研究指出，大數據對於市場有非常大的幫助，一般公司的行銷人員或決策者如果能夠有效的利用大數據的資料而獲得目標族群的行為，可進而確定一些商業上決策的可行性。針對大數據的行銷策略運用，如果決策者能有效的看到精髓所在，便能夠對於消費者產生直接或者間接的影響，然後就可以創符合市場上的需求，再使用一些歸納法的方式歸納策略上的分析模式，藉此刺激消費者行為與市場潛在需求來滿足行銷行為(呂瑞男，

2015)。游綉雯(2015)針對線上旅遊網站評論資料作情緒分析，將評論分類並計算情緒分數，藉此了解消費者滿意與不滿意的項目。

此外，關鍵字廣告已經成為行銷工具的大宗，網頁搜尋龍頭 Google 的關鍵字廣告(Google Adwords)收入就佔了其營收的絕大部分，因為關鍵字廣告具有針對性，可依據特定客戶的喜好、產品特性來設計合適的廣告文案，以此達到吸引客戶的目的(李凱斌，2011)。

由上述文獻可以得知，大數據時代的來臨與網路密不可分，也直接導致了商業行為網路化，因此如何將這些資訊數據篩選分類與分析，便是學者們重視的研究領域。

第二節 斷詞系統

在自然語言處理上，最基本的處理單位通常是詞，此處的詞指的是語言學家所定義的「能夠獨立運用，具有完整語意的最小語言成分」。很多自然語言處理應用的研究，例如：文件檢索、中文輸入、語意辨識等等，都需要先將本文切割，以詞為單位後才能進行後續處理。在英文中，每個單字(word)就可成詞，且大都以空白當作字與字之間的符號，因此無須進行斷詞的前置作業。反之中文字詞和詞之間並無空白或特定符號區隔，因此將正確的詞切分出來，就成為自然語言處理的最基礎工作。同樣地，文章的摘要或者文章本身也都是得經過斷詞處理，才進行後續的處理。

傳統用在中文上的斷詞方法一般分為兩大類別，分別是 N-gram 與詞彙法 (Word based)。近年來中文斷詞的研究主要以詞彙法為主導，但 N-gram 仍為實務所重視。

(一) 中研院斷詞系統(CKIP)

中研院資訊所、語言所於民國七十五年成立一個跨所合作的中文計算語言研究小組，共同合作建構中文自然語言處理的資源與研究環境，為國內中文自然語言處理及相關研究提供基本的研究資料與知識架構。中文斷詞系統 (<http://ckipsvr.iis.sinica.edu.tw/>) 為其研究之一，其特色為：

- 包含一個約十萬詞的詞彙庫及附加詞類、詞頻、詞類頻率、雙連詞類頻率等資料。分詞依據為此一詞彙庫及重疊詞等構詞規律及線上辨識的新詞，並解決分歧義問題。
- 採用的「中央研究院平衡語料庫」，是世界上第一個有完整詞類標記的漢語平衡語料庫。1997 年開放的研究院語料庫 3.0 版已達到五百萬目詞的規模，目前正朝向一千萬詞的目標邁進。

(二) Jieba 斷詞系統

斷詞中最著名的便是「結巴(Jieba)中文斷詞」(<https://github.com/fxsjy/jieba>)，Jieba 的核心演算法為使用語料庫建立 Trie 字典樹，Trie 字典樹的結構為單詞以及其機率值，首先利用正規表示式將語句中符號與文字切割，接著把屬於文字

之部分與 Trie 字典樹比對，計算共有幾種斷詞組合並且產生出現的機率值，而根據這些單詞組合與機率值組合表示成一個有向無環圖(DAG)。若連續出現某個單字詞，則將這些單字詞組合匯入隱馬可夫模型(Hidden Markov Model)，再使用 Viterbi 演算法計算是否能組合成新的單詞，提升斷詞組合之滿意度。結巴(Jieba)中文斷詞在處理詞語歧義性與新詞識別有很好的效能。

上述所提供的兩者斷詞方法，本研究使用中研院斷詞系統，因 Jieba 斷詞適合用在簡體文字上，但本研究的新聞用字全部為繁體中文，所以如果使用 Jieba 斷詞，會有語意上斷詞的問題，影響此研究之結果。

第三節 極性分類 (polarity classification)

語意分析涉及正反意見在文字語句中的判定，因此必須使用多方視角看待問題與答案，與此相關的研究包含意見導向的資訊萃取、資訊摘要。語意分析通常分為三個階段：校準、辨識、與分類所有已經讀取的文句。本研究探討文件階層(document-level)的分析，將會著重分析特定類型的評論文件。首先須處理的問題就是極性分類 (polarity classification)，對於目標決定正極性 (贊同)或者是負極性(反對)，部分研究也包括中性的極性，雖然研究成果相當多且廣，極性分類仍然是自然語言處理系統中重大的挑戰。

接著介紹極性分類的方法。在語言學中，建立一個高效的極性分類可透過以下特徵：

- high order n-grams
- 複合形容詞，例如 happy 被視為正，而 terrible 視為負面。
- 詞彙的相依關係
- 來自於中立文件中所描述的詞組

經由極性分類，可以反映正反兩方兩種評論，為了增加精準度，其中一種做法是將單純闡述事實的評論去除、以及在中性評論用的用詞特別處理，接著對於形容詞與關聯名詞做統計，確保面向的評論對象是研究所需。

第四節 情感分析(Sentiment Analysis)

情感分析主要目的是分類文件或句子意見表達的極性為正向、負向或中立。常使用的方法是利用情感字典，或是訓練已經標記好的評論，用於建立情感模型(Sentiment model)。在情感詞典部分，可以是廣義的情緒詞彙，或者是特定領域的情緒詞彙(Lu and Tsou, 2010)。情感分析的技術是關於辨識和整合使用者表達的情緒或意見，在過去大部分情感分析研究方法中，可以分為監督式(Supervised)、非監督式(Unsupervised)和半監督式(Semi-Supervised)。

監督式學習可以由訓練資料中學到或建立一個模式，並依此模式推測新的實例，例如使用以 Maximum Entropy 為主的機器學習演算法(Li et al., 2009);使用 Support Vector Machines(SVM)分類微博文章的情感極性(Sui et al., 2012);而在 Yang 和 Zhou(2011)所提的論文中，作者分析中文句子的語意結構，接著將分析

的語意結構依照動詞的句法結構，分成直接和間接的意見，再使用 Maximum Entropy 演算法學習這兩種特徵，對中文句子進行情緒分析。文章情感分類的非監督式學習方法則是以辭典為主(Wang and Lee, 2011; Zhai et al., 2010)，Zhang 等人(2012)提出改善自動化產生情感辭典的方法，並且給予詞彙分數。首先利用已知的正負向情感詞彙建立初始的種子詞彙，然後定義代表詞彙的語意極性分數的分數向量，並建立種子詞彙和其他詞彙之間的關聯圖，最後根據向量和關聯圖擴充種子詞彙；Wang 等人(2013)使用以字典為基 lexicon-based 的方法並結合模糊理論，首先將情感辭彙進行情緒量化，接著定義分類正向和負向的模糊集合和歸屬函數，然後透過實驗訓練資料集決定歸屬函數的最佳參數，最後利用模糊分類器對評論進行分類。另外，Zhuo 等人(2014)應用模糊理論進行情感分析，他們提出根據情感程度詞彙和中文模糊語意模型對中文文章進行情感分析。而半監督式學習則是介於監督式與非監督式學習之間，有一部分使用已經標好的資料(labeled data)，而剩下尚未被發現的特徵為未標記的資料分類到正確的特徵類別，例如 Zhai 等人(2012)應用半監督式學習於辨識領域的特徵，某些特徵在不同領域有不同意義，例如"picture" 這個詞在相機中代表相片，在電影領域則代表影片。

第五節 NTUSD

NTUSD(National Taiwan University Semantic Dictionary)(Ku&Chen, 2007)是台灣大學意見詞典。此詞典包含了常用的意見詞詞彙，以及該詞彙的極性。將這

些詞彙分類成正向極性與負向極性，分別屬於 NTUSD positive、NTUSD negative，本研究所使用的是舊版的 NTUSD，一共收錄約 2810 個正向詞彙以及 8276 個負向詞彙。在 NTUSD 中，除了意見詞本身之外，也包含了該意見詞的實際應用、例句等。以下截錄部分 NTUSD positive 內容於表 1；部分 NTUSD negative 內容於表 2。可以看出某些情緒詞有發生跨區出現的情形，例如兩張表中的詞彙「入迷」、「入迷的」。

表 1 NTUSD positive 部分詞彙

了不起	了不起的	了解	人情	人情味
入迷	入迷的	上等	大膽	大膽的
小天使	才氣	才能	才智	才幹
不任性	不任性的	中肯	井然有序	互相密和著的

表 2 NTUSD negative 部分詞彙

一下子爆發	一巴掌	一再	一再叮囑	一陣緊張
入迷	入迷的	入迷的人	刀刀	刁難
匕首	下垂	下垂度	下流	下流的
小心	小卒	小氣	小氣的	小偷

第六節 TF-IDF

TF-IDF (term frequency-inverse document frequency) 是一種用於資訊檢索與

文本挖掘的常用加權技術。TF-IDF 是一種統計方法，用以評估字詞對於一個文件集或一個語料庫中的其中一份文件的重要程度。字詞的重要性隨著它在文件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。TF-IDF 加權的各種形式常被搜尋引擎應用，作為文件與用戶查詢之間相關程度的度量或評級。除了 TF-IDF 以外，網際網路上的搜尋引擎還會使用基於連結分析的評級方法，以確定文件在搜尋結果中出現的順序。

TF-IDF 的主要概念是：如果某個詞或短語在一篇文章中出現的頻率 TF 高，並且在其他文章中很少出現，則認為此詞或者短語具有很好的類別區分能力，適合用來分類。TF-IDF 實際上是： $TF * IDF$ ，其中 TF (Term Frequency) 代表詞頻，IDF (Inverse Document Frequency) 代表逆向文件頻率。TF 表示詞條在文檔 d 中出現的頻率。

在一份給定的文件裡，詞頻 (term frequency, TF) 指的是某一個給定的詞語在該文件中出現的頻率，這個數字是對詞數 (term count) 的歸一化，以防止它偏向長的文件，因為同一個詞語在長文件裡可能會比短文件有更高的詞數，而不管該詞語重要與否。對於在某一特定文件裡的詞語 t_i 來說，它的重要性可表示為公式(1)：

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

其中 j 是「某一特定文件」，i 是該文件中所使用單詞或單字的「其中一種」， $n_{i,j}$ 就是 i 在 j 當中的「出現次數」，那麼 $tf_{i,j}$ 的算法就是 $n_{i,j} /$

$(n(1,j)+n(2,j)+n(3,j)+\dots+n(i,j))$ 。例如第一篇文件中，被篩選出兩個重要名詞，分別為「健康」、「富有」，「健康」在該篇文件中出現 70 次，「富有」出現 30 次，那「健康」的 $tf = 70/(70+30)=70/100=0.7$ ，而「富有」的 $tf=30/(70+30)=30/100=0.3$ ；在第二篇文件裡，同樣篩選出兩個名詞，分別為「健康」、「富有」，「健康」在該篇文件中出現 40 次，「富有」出現 60 次，那「健康」的 $tf = 40/(40+60)=40/100=0.4$ ，「富有」的 $tf = 60/(40+60)=60/100=0.6$ ， tf 值愈高，其單詞愈重要。

逆向文件頻率 (inverse document frequency, IDF) 是一個詞語普遍重要性的度量。某一特定詞語的 IDF，可以由總文件數目除以包含該詞語之文件的數目，再將得到的商取對數得到，如公式(2)

$$idf_i = \log \frac{|D|}{|\{j: t_i \in d_j\}| + 1} \quad (2)$$

其中 $|D|$ 為語料庫中的文件總數， $|\{j: t_i \in d_j\}|$ 則是包含詞語 t_i 的文件數目 (即 $n_{i,j} \neq 0$ 的文件數目)。如果該詞語不在語料庫中，就會導致分母為零，因此一般情況下使用 $1 + |\{j: t_i \in d_j\}|$ 。

因此 TF-IDF 公式如下：

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad (3)$$

從公式(3)可看出，某一特定文件內的高頻率詞語，以及該詞語在整個文件集合中的低文件頻率，可以產生出高權重的 TF-IDF。因此，TF-IDF 傾向於過濾掉常見的詞語，保留重要的詞語。

第三章 研究方法

第一節 緒論

在資訊爆炸的時代，人們該如何從這些大量的報導中獲取想要或者喜歡的資訊，而媒體又該如何從新聞內文中適當的撰寫以便吸引閱聽人，讓人們能夠喜歡閱讀你這家報導，並且可以從這些報導當中，發掘現今人們是喜歡比較負面的新聞或者正向新聞，這些都是目前值得研究者探討且著墨的地方。

本研究目標為分析現在網路新聞內文經常使用哪些詞彙或語句，可激發閱聽人的情緒反應以增加其閱讀興趣，以及處理新聞極性的趨勢，也就是正向的新聞比較受歡迎還是負向的新聞比較受歡迎，進而可以讓各家報章可以利用本研究去撰寫出一篇吸引度高的新聞報導。

第二節 研究資料

(一) 研究材料

新聞內文是取自於蘋果新聞，為何選取蘋果新聞而不選取其他家新聞，因為蘋果新聞是經由統計出來，現今的人們，不管男女老少，都是比較偏好於蘋果新聞。因此，本研究題材就選取蘋果新聞內文，來做為此研究資料，因為根據專門進行網站流量等數據分析的市調公司 comScore 最新調查，《蘋果日報》四月份於全裝置新聞類別，不重複訪客數量排名居各媒體第一；其中，在手機裝置包括手機網和 App 排行，也都高居第一。本研究挑取一個禮拜的新聞(時間為 2017 年的

5/3~5/9)，總共約 500 篇新聞，先到 facebook 蘋果新聞的粉絲專頁(如圖 2)去找尋新聞，尋找完之後，利用爬蟲程式(如圖 3)將所需要的資料，截取出來(如圖 4)。



圖 2 蘋果粉絲專業



圖 3 爬蟲程式(節錄)

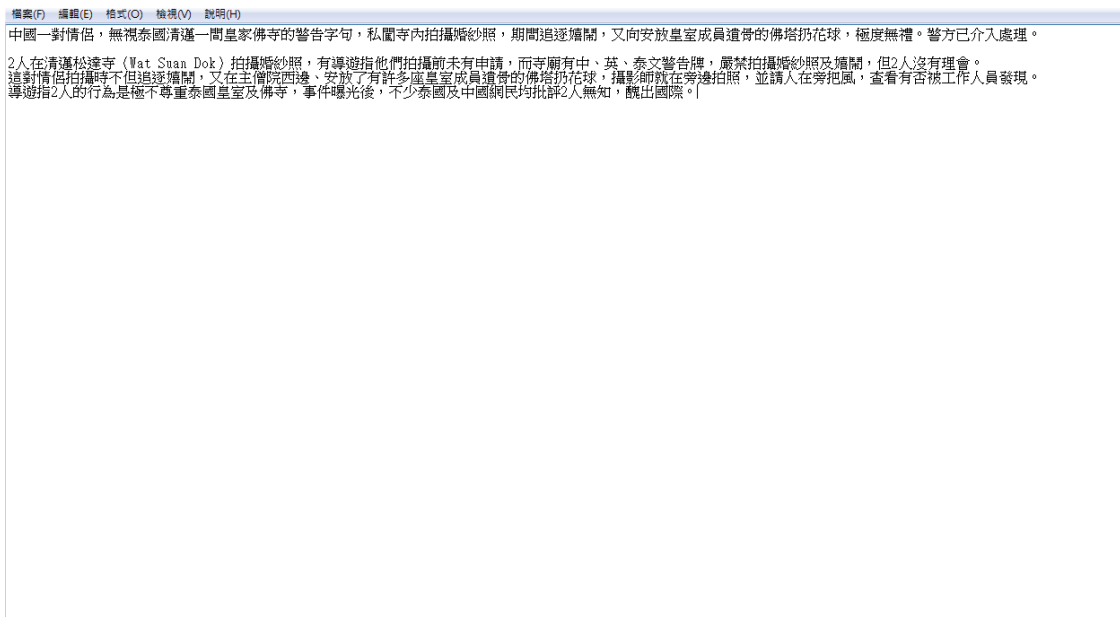


圖 4 爬取新聞內文

接下來，需要取得每一篇新聞的讚數多寡，因為要怎麼知道這一篇新聞是否很多人去點閱，就必須靠讚數多寡判斷。兩種方法可以進行抓取某篇新聞的讚數數量。方法一:直接到粉絲專頁，挑取剛剛所抓的新聞內文的 po 文，進而可以看到讚數數量的多寡(如圖 5)，可以看到這篇的讚數為 1011。方法二:利用 Facebook 所提供的 facebook for developer¹平台，並且透過平台所提供的 graph API，輸入要抓取資料的指令，進而可以得到 json 資料，json 資料裡面包含新聞標題、新聞內文、新聞發布時間以及新聞按讚的個數等資訊，並且可以得知是誰給這篇新聞一個讚的肯定(如圖 6、圖 7)。而且利用 facebook 所提供的開發者人員(facebook for developer)，可以直接下取指令，快速的找到想要的資料，並不需到實際的粉絲專業，去慢慢搜尋想要的資訊。例如:想要得知五月一日的新聞，就可以利用指令的方式，直接抓取五月一日所有的新聞資訊。不然現在是七月，若想要往回抓取六

¹ <https://developers.facebook.com/>

月份的資料，就必須在粉絲專頁，慢慢的往下滑，找到六月份所有的新聞資料，這樣子需要花費很長的一個時間，所以建議是利用方法二的方式，可以讓快速去抓取想要的資料，節省時間並且可以正確的抓取資料。



圖 5 從粉絲專頁得知讚數多寡

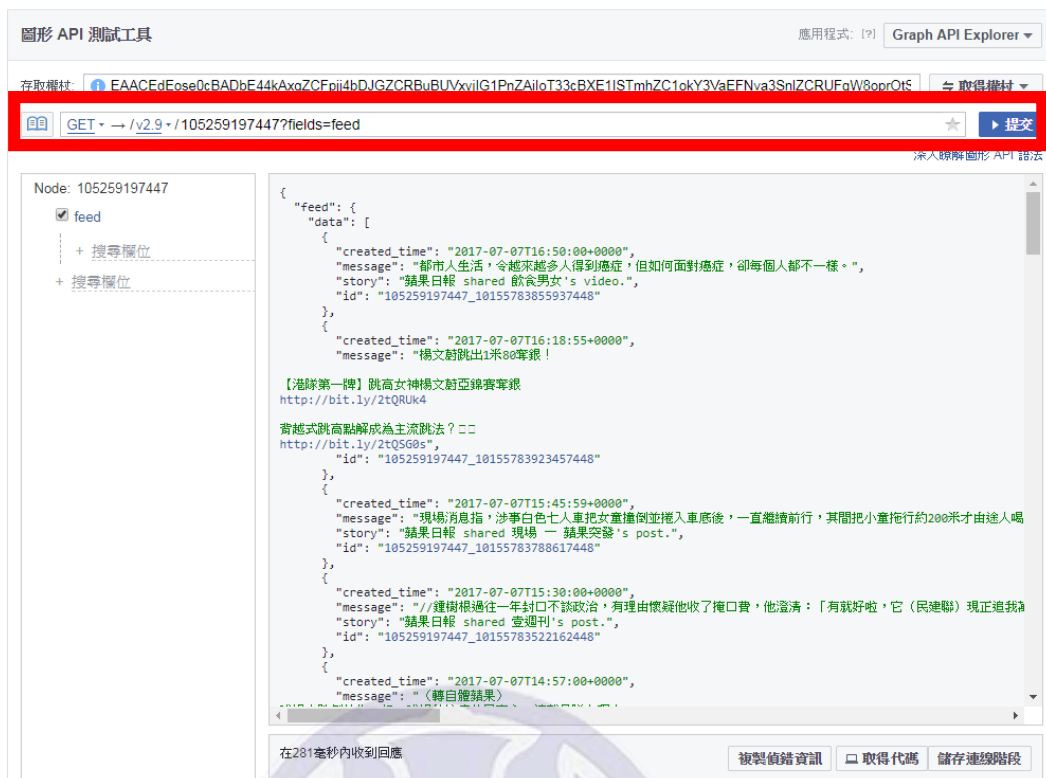


圖 6 graph API 資料的指令

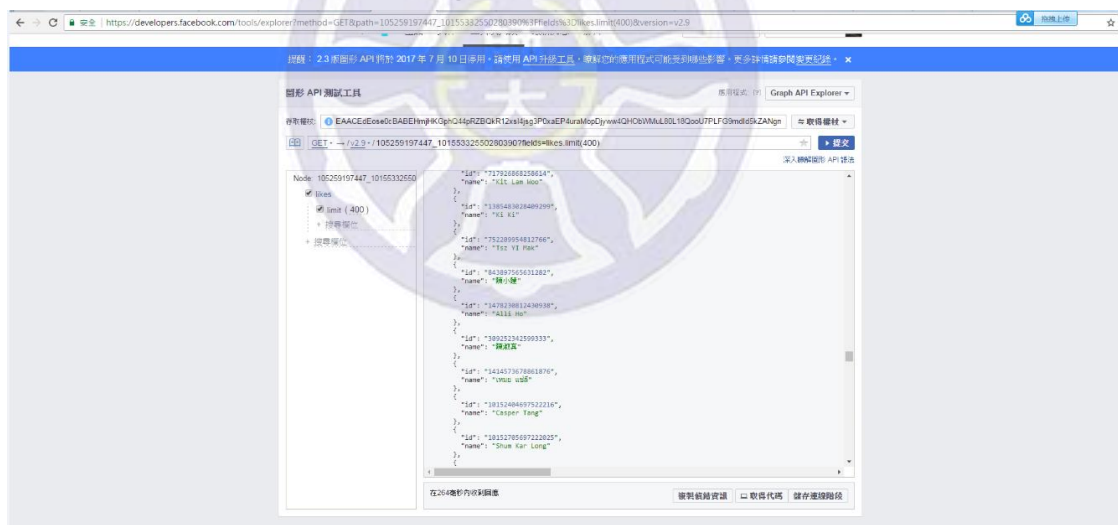


圖 7 graph API 資料的指令(讚數)

由上圖 6 的紅色框可以發現，那一條指令其實就是在抓取蘋果新聞粉絲專頁裡面的所有資料。可以發現當中有一串數字 105259197447，這一串數字就是粉絲專頁的 id 代碼，這串 id 代碼要從使用者資料先去找找到蘋果日報的名稱，然後再對應到 id，即可知道。接下來後面那串指令，就是利用 facebook 所提供給的指令，

去抓取粉絲專頁的內容。指令如下: 105259197447?fields=feed, 就可以看到結果的呈現。呈現方式是利用 json 的格式去輸出, 那麼裡面可以看到有 created_time(建立日期)、message(新聞標題)、story(新聞的內聞)以及每篇新聞所屬的 id。接下來看到上圖 7, 可以再進一步的去查看這篇新聞有幾個人按讚, 一樣可以去下指令, 抓取按讚個數。抓取指令如下: 105259197447_10155332550280390? fields=likes, 資料格是一樣是以 json 的格式輸出, 再去計算總共按讚個數即可。

(二) Facebook API 簡介

Facebook API(<https://developers.facebook.com/>)可以應用在多種環境下, 包括網站(Website)、Facebook 應用程式(Apps on Facebook.com)以及行動應用程式(Mobile Apps)。根據不同的平台, Facebook 也提供了不同的 SDK(Software Development Kit)給開發者使用, 包含了網站上使用的 JavaScript SDK、PHP & Python SDK, 以及行動裝置上的 iOS SDK(iPhone & iPad)、Android SDK 等, 可以說無論在哪個平台上皆可看到 Facebook API 的身影。

使用 Facebook API 並不困難, 甚至有越來越簡易的趨勢。在過去一兩年間, 由於將 Facebook API 用在網站上的使用人數大幅增加, Facebook 也將常用到的一些服務包裝成套件供開發者直接套用 (Facebook 稱之為 Social Plugin), 其中最常見的『按讚』、『朋友的動態』、『留言系統』…等。若一般網站開發者需要使用這些功能, 只要到 Facebook 開發者網頁填寫一些必要的資訊 (appID、plugin 的寬度、回傳 url…等), Facebook 便會自動替開發者產生相對應的 plugin 程式碼,

接著開發者只需要將這些程式碼嵌入網頁中就大功告成了！而在筆者截稿之前，Facebook 甚至開放『留言』(comment) API，進一步將 Facebook 的互動融入在各個網站中。

但無論 Facebook 提供各種 API，追根究柢，其實大家想要使用 Facebook API 的目的終究是希望能夠擷取使用者以及與使用者相關的一些物件（朋友、網頁…等）之間的鏈結，也因此 Facebook 將這個概念濃縮成為一個核心概念，稱之為 Social Graph。而存取這些關係的 API，就稱之為 Graph API。

(三) Graph API

Graph API(<https://developers.facebook.com/tools/explorer/145634995501895/>)提供了可以一窺社群網路的方法，並提供物件本身的資訊以及物件間彼此的鏈結關係。本處所提到的物件，其實在 Facebook 裡的定義是非常廣泛的，舉凡人、照片、事件、網頁都可以是一個物件。而鏈結關係就是這些物件間彼此的關係，例如人跟人之間的關係（朋友/非朋友）、人跟照片的關係（人擁有某些照片、照片上的標籤屬於某些人等）、人跟事件與網頁的關係（讚）…等。因此其實也可以說，任何研究人員需要 Facebook 提供的資訊，Graph API 幾乎都能滿足研究人員的需求。

(四) 中文斷詞系統

本研究利用 CKIP 的中文斷詞系統(<http://ckipsvr.iis.sinica.edu.tw/>)進行蘋果日報新聞內文的斷詞。從中研院斷詞系統中，把一篇文章斷完詞之後，然後找出詞性的特徵。因為是先就有把 500 篇新文抓下來，在這個步驟中，直接把內文貼到

中文斷詞系統

- 簡介
- 未知詞與收錄方法
- 詞類標註列表
- 線上顯示
- 線上服務申請
- 線上資源
- 公告
- 聯絡我們

隱私權聲明 | 版權資訊



Copyright © National
Digital Archives Program,
Taiwan.
All Rights Reserved.

[\[版面說明 \]](#)
[\[錯誤回報 \]](#)
[\[中文詞彙特性使用指南 \]](#)
[\[語料查詢 \]](#)
[\[語法查詢 \]](#)
[\[詞類標註 \]](#)
[\[字彙查詢 \]](#)

自置(Hc)	將係(Na) { PARENTHESISCATEGORY } 3(0 Neu) 公升(MF) ↓ PARENTHESISCATEGORY 臨盆(Na) 比比皆是(VH) + COWPACATEGORY	
恒(Cbb) { PARENTHESISCATEGORY }	一樣(VH) 非(Na) 微(VC) 實(Na) 耀(Na) (Na) ↓ PARENTHESISCATEGORY + COWPACATEGORY	
一(Neu) 樓(MF) 陰籃(Na) 也(D) 有(V_2) 4(Neu) 取(Fu) 基(Na) + COWPACATEGORY		
美裔(Nc) 總辦(Nc) 泌尿科(Nc) 醫術(Na) 鮮時菜(Nb) { PARENTHESISCATEGORY } David(Fu) Shuterman(Fu) ↓ PARENTHESISCATEGORY 郭(D)	整頓出(VC) 4(Neu) 大(VH) 陰籃(Na) 嬌型(Na) + COWPACATEGORY	
4(NuV) 飲(Na) 是(SHI) 擊打(VT) 喝(Hop) —(Nu) 耀(Hu) + PERIODCATEGORY		
1(Neu) { PERIODCATEGORY } 燈罩型(Na) + PERIODCATEGORY		
鮮時菜(Nb) 指出(VE) + COWPACATEGORY		
如果(Cbb) 男性(Na) 是(SHI) 彎彎型(Na) + COWPACATEGORY		
可觀(D) 票(D) 注意(VK) 平(Hu) + COWPACATEGORY		
尤顯(D) 是(SHI) 彎彎(VH) 的(DE) 角度(Na) 出處(VH) 3(0 Neu) ~/(Fu) 6(0 Neu) 度(Hu) + COWPACATEGORY		
短(Dfa) 有(V_2) 可短(D) 是(SHI) 電風(VT) { PARENTHESISCATEGORY } 佩治尼氏(Na) ↓ PARENTHESISCATEGORY Peyronies(Fu) disease(Fu) + COWPACATEGORY		
文(D) 積(VG) 曲甲座(Na) { PARENTHESISCATEGORY } + COWPACATEGORY		
變(D) 齒牙(D) 表(D) 有髮型(Na) + PERIODCATEGORY		
2(Nu) { PERIODCATEGORY } 鐘(Na) 小(VH) 身(Na) 大型(A) + PERIODCATEGORY		
其(Na) 嬌型(Na) 煎(DE) 陰籃(Na) { PARENTHESISCATEGORY } 鼻籠(Na) ↓ PARENTHESISCATEGORY 軟(Dfa) 大(VH) + COWPACATEGORY		
恒(Cbb) 前隔(Ncd) 軟(Dfa) 小(VH) + COWPACATEGORY		
3(Neu) { PERIODCATEGORY } 曲線型(Na) + PERIODCATEGORY		
宛型(VG) 爹(D) 惹(Caa) 驚(Na) 煎(DE) 長尾(Na) + PERIODCATEGORY		
鮮時菜(Nb) 表示(VE) + COWPACATEGORY		
曲線型(Na) 和(Caa) 彎彎型(Na) 最(Dfa) 大(VH) 的(DE) 不同(VH) 之(DE) 處(Na) 是(SHI) + COWPACATEGORY		
曲線型(Na) 煎(DE) 陰籃(Na) 是(SHI) 帶頭(D) 上橫(VH) + COWPACATEGORY		
恒(Cbb) 彎彎型(Na) 的(DE) 陰籃(Na) 是(SHI) 下蓋(VH) 吹氣(Na) + COWPACATEGORY		
大約(Da) 有(V_2) 1(0 Neu) ~(Fu) 2(0 Neu) 男性(Na) 擁有(VT) 該(Nes) 嬌型(Na) + PERIODCATEGORY		
4(Neu) { PERIODCATEGORY } 隔犬(VH) 身(Na) 小型(Na) + PERIODCATEGORY		
陰籃(Na) 宛型(VG) 香結型(Na) + COWPACATEGORY		
前隔(Ncd) 比(VC) { PARENTHESISCATEGORY } 身籠(Na) ↓ PARENTHESISCATEGORY 壓犬(VH) + COWPACATEGORY		
蔚(Da) 有(V_2) 2(0 Nuq) 陰籃(Na) 屬於(VT) 該(Nes) 嬌型(Na) + PERIODCATEGORY		

可能產生的詞性列表，亦即就利用這些特徵延伸發展，創造屬料庫統計值與全球資訊網統計

表 3 中文斷詞系統詞性表

精簡詞類	簡化標記	對應的CKIP詞類標記 ²	
A	A	A	/*非謂形容詞*/
C	Caa	Caa	/*對等連接詞，如： 和、跟*/
POST	Cab	Cab	/*連接詞，如：等等 */
POST	Cba	Cbab	/*連接詞，如：的話 */
C	Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	/*關聯連接詞*/
ADV	Da	Daa	/*數量副詞*/
ADV	Dfa	Dfa	/*動詞前程度副詞*/
ADV	Dfb	Dfb	/*動詞後程度副詞*/
ASP	Di	Di	/*時態標記*/
ADV	Dk	Dk	/*句副詞*/
ADV	D	Dab, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
N	Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
N	Nb	Nba, Nbc	/*專有名稱*/
N	Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
N	Ncd	Ncda, Ncdb	/*位置詞*/
N	Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
DET	Neu	Neu	/*數詞定詞*/
DET	Nes	Nes	/*特指定詞*/
DET	Nep	Nep	/*指代定詞*/
DET	Neqa	Neqa	/*數量定詞*/
POST	Neqb	Neqb	/*後置數量定詞*/
M	Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
POST	Ng	Ng	/*後置詞*/
N	Nh	Nhaa, Nhab, Nhac, Nhbc, Nhc	/*代名詞*/
Nv	Nv	Nv1, Nv2, Nv3, Nv4	/*名物化動詞*/
T	I	I	/*感嘆詞*/
P	P	P*	/*介詞*/
T	T	Ta, Tb, Tc, Td	/*語助詞*/
Vi	VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/
Vt	VAC	VA2	/*動作使動動詞*/
Vi	VB	VB11,12,VB2	/*動作類及物動詞*/

¹ 斜體詞類，表示在技術報告#93-05中沒有定義，即後來增列的。

Vt	VC	VC2, VC31,32,33	/*動作及物動詞*/
Vt	VCL	VC1	/*動作接地方賓語動詞*/
Vt	VD	VD1, VD2	/*雙賓動詞*/
Vt	VE	VE11, VE12, VE2	/*動作句賓動詞*/
Vt	VF	VF1, VF2	/*動作謂賓動詞*/
Vt	VG	VG1, VG2	/*分類動詞*/
Vi	VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
Vt	VHC	VH16, VH22	/*狀態使動動詞*/
Vi	VI	VI1,2,3	/*狀態類及物動詞*/
Vt	VJ	VJ1,2,3	/*狀態及物動詞*/
Vt	VK	VK1,2	/*狀態句賓動詞*/
Vt	VL	VL1,2,3,4	/*狀態謂賓動詞*/
Vt	V_2	V_2	/*有*/
T	DE	/*的, 之, 得, 地*/	
Vt	SHI	/*是*/	
FW	FW	/*外文標記*/	
COLONCATEGORY			/* 冒號 */
COMMACATEGORY			/* 逗號 */
DASHCATEGORY			/* 破折號 */
ETCCATEGORY			/* 刪節號 */
EXCLAMATIONCATEGORY			/* 驚嘆號 */
PARENTHESISCATEGORY			/* 括弧 */
PAUSECATEGORY			/* 頓號 */
PERIODCATEGORY			/* 句號 */
QUESTIONCATEGORY			/* 問號 */
SEMICOLONCATEGORY			/* 分號 */
SPCHANGECATEGORY			/* 雙直線 */

接著如圖 9 所示，會看到 Na、Nf、VH、VC、Neu、D、FW、VE、Nh、VJ、Dfa、Nd 等等，這些都是由中文斷詞系統所產生的詞性中挑選。在中文斷詞系統特徵詞中，會選這些詞性做為使用，主要原因是因為中文斷詞系統特徵中會出現那些詞性是沒有意義的。像是 T 這個詞性代表語助詞，而研究實驗中無法依語助

詞找出是否會影響閱聽者對一篇報導吸引度高低的判斷，故無法用上。去除這些沒有意義的詞性後，本研究就可以再進一步分析與探討。

	A	B	C	D	E	F	G	H	I	J
1	鄉民	(Na)	1							
2	公分	(Nf)	1							
3	比比皆是	(VH)	1							
4	一樣	(VH)	1							
5	米	(Na)	1							
6	養	(VC)	1							
7	百	(Neu)	1							
8	樣	(Nf)	1							
9	人	(Na)	1							
10	也	(D)	1							
11	款	(Nf)	1							
12	型	(Na)	1							
13	美國	(Nc)	1							
14	紐約	(Nc)	1							
15	泌尿科	(Nc)	1							
16	醫師	(Na)	1							
17	David	(FW)	1							
18	Shustermar	(FW)	1							
19	就	(D)	1							
20	整理出	(VC)	1							
21	看看	(VE)	1							
22	你	(Nh)	1							
23	1	(Neu)	1							
24	指出	(VE)	1							
25	如果	(Cbb)	1							
26	注意	(VK)	1							
27	一下	(Nd)	1							
28	尤其	(D)	1							
29	彎彎	(VH)	1							
30	角度	(Na)	1							
31	出現	(VH)	1							
32	60	(Neu)	1							
33	度	(Nf)	1							
34	很	(Dfa)	1							
35	罹患	(VJ)	1							

圖 9 斷詞完之後的詞彙與詞性特徵及頻率次數

(二) 利用 NTUSD 的擴充

此部分中應用 NTUSD positive 以及 NTUSD negative 兩份語料做進一步的擴充，試圖納入更多新聞內文撰寫方法之高低吸引度的情感詞彙。

NTUSD 為一份詳細標記極性的語料庫。除了情緒詞本失之外，甚至包含該辭會的相關用法、可能出現的句型。例如:詞彙「不凡」被歸類為正向極性詞彙;相關的詞彙「不凡的人」同樣被歸類於正向極性詞彙中。「不凡的人」中多出了「的」以及「人」兩個修飾詞彙。需要的關鍵是極性詞彙本身，而非相關的修飾詞。雖然一個詞彙越是明確、描述的越仔細，判斷時的正確率越高;但另一方面修飾詞卻可能影響擴充時的效率。

為了取出具有明確性的詞彙，於是將 NTUSD positive 以及 NTUSD negative 分別經過斷詞系統做分析。由於不確定斷詞後的情緒詞彙將被判定為哪些類型的詞性，為了盡可能增加擴充詞彙，同時不漏掉和習慣上認知類型不同的詞彙，因此本研究加入所有詞性的詞彙，盡可能達到本研究擴充詞典的完整性。

如果與 NTUSD 語料庫進行 positive 和 negative 的比對，如果該詞彙沒有在該字典裡面，就必須再透過教育部新增詞典，進行相似詞的比對。如果該詞彙的相似詞如果在 NTUSD 有出現，那麼就將該詞彙歸類為一樣的極性。

(三) 教育部重編國語辭典修訂本

教育部重編國語辭典修訂本網路版收錄單字一萬一千多字，詞目十六萬條。可以字詞、注音、釋義查詢。

需要靠這個工具去做出辭典擴充的判斷，今天有一個詞彙為出現在 NTUSD 與料庫裡面，那麼就必須到該詞典修訂本來查詢該詞彙的相似或者相反詞，來幫助將該詞彙歸類為何種極性。

假設今天有"完善"一詞，於是就到 NTUSD 進行比對，可是在 NTUSD 並沒有找到"完善"這個詞的存在，所以這時候，就必須到教育部重編國語辭典修訂本找與"完善"一詞的相似詞，如果"完善"的相似詞有在 NTUSD 裡面的話，就將該詞歸類一樣的極性。

如圖 10 將"完善"輸入到該詞典搜尋引擎裡面，就會發現會跑出與"完善"這個詞的相似詞與相反詞以及他的釋義(如圖 11)。



圖 10 教育部重編國語辭典修訂本



圖 11 相似詞相反詞的結果輸出

由如圖 11 可以發現，"完善"這個詞彙的相似詞有"完美"、"完滿"、"圓滿"這三種相似詞，於是分別將這三個相似詞與 NTUSD 去進行比對，可以發現"完美"這個詞彙，在 NTUSD 裡面是屬於正向詞彙(如圖 12)。於是就可以將"完善"一樣歸類為正向詞彙，並且將"完善"一詞新增到自己的詞典裡，新增一套新聞高低吸引度的情感字典。

圖 12"完美"一詞在 NTUSD 所歸類的極性

第三節 研究方法與架構

本研究提出的方法分為兩大部分，即方法一和方法二，方法一主要處理利用情緒分析的技術，分析現在網路新聞內文經常使用哪些詞彙或語句，可激發閱聽人的情緒反應以增加其閱讀興趣。且在方法一中擴充一套屬於新聞高低吸引度的詞彙字典，會與 NTUSD 進行比對的動作，並且藉由教育部字典做為輔助，以便將詞彙新增到自建的語料庫中。所抓下來的新聞，利用 Facebook 的粉絲專業讚數當做驗證，並且用讚數平均的方式找出哪些是高吸引度新聞、哪些是低吸引度新聞。將 500 篇新聞的讚數計算出平均值，如果讚數超過平均值，就把它歸類成高吸引度，如果低於平均值，就把它歸類為低吸引度。

方法二主要處理新聞極性的趨勢，也就是正向的新聞比較受歡迎還是負向的新聞比較受歡迎。會先進行斷詞之後，利用 TF-IDF 值尋找出關鍵字，然後利用語料庫進行比對，得到正向詞或負向詞的資訊，接著再利用 Facebook 提供的讚數當作佐證。

方法一：

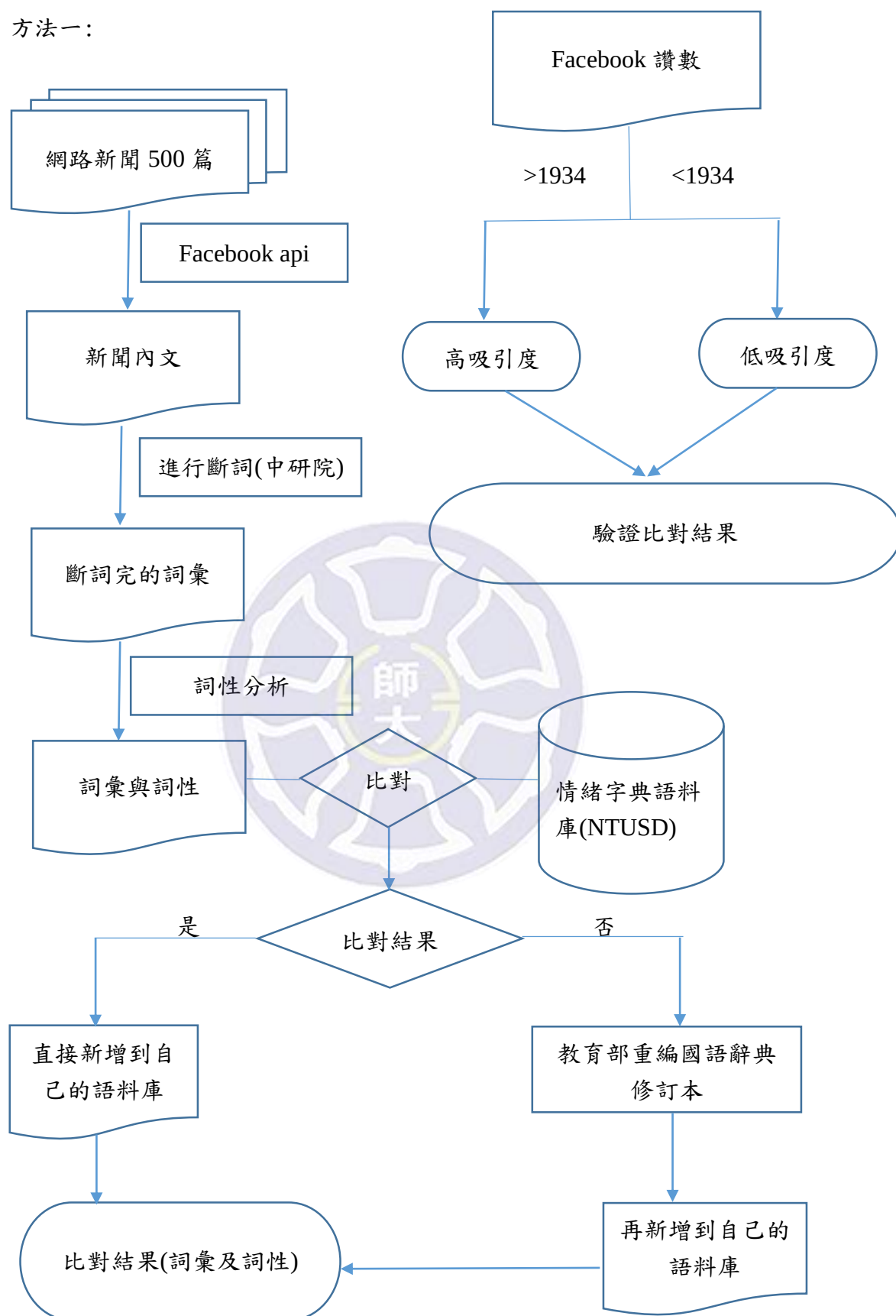


圖 13 擴充辭典並且找出吸引度高低所使用的詞性為何

方法二:

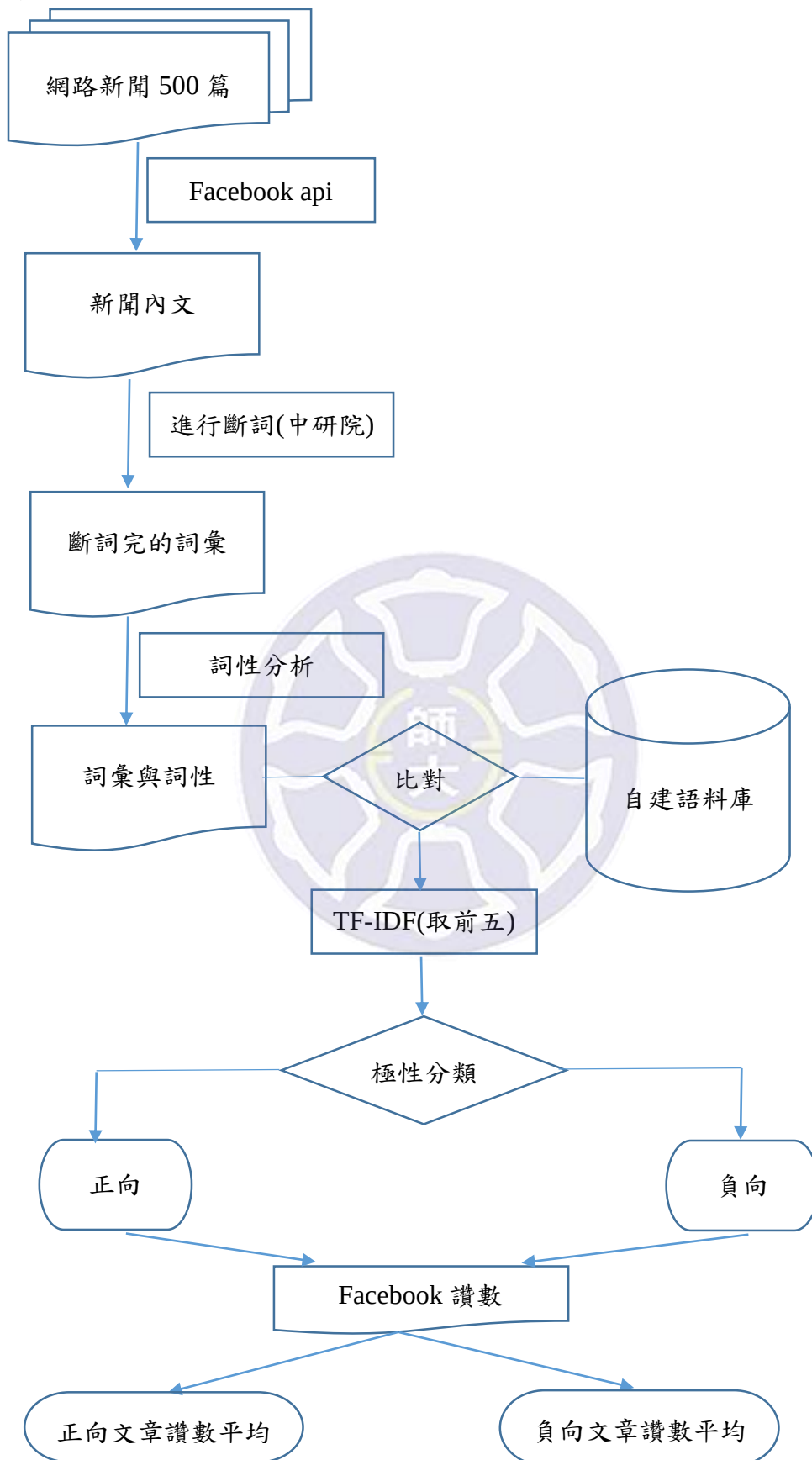


圖 14 現代人喜歡負面新聞或者正向新聞

第四節 研究方法描述

本研究方法可以歸類成以下幾個步驟說明:(一)利用斷詞系統將新聞內文斷詞，(二)新增自己的語料庫，分成高吸引力語料庫以及低吸引力語料庫，(三)利用 TF-IDF 找出一篇文章的關鍵詞，各步驟詳細如下:

(一) 利用斷詞系統將新聞內文斷詞

將從蘋果新聞所得到的新聞內文經由中文斷詞系統斷詞完之後，會得到已經切割完的詞彙，然後將這些詞彙存成 excel 檔，之後統計字頻時候會用到，以及在看哪些詞性是會影響整篇新聞的吸引力。

例如:

新北市一名黃姓男子上周三晚間 8 時許，騎車行經板橋縣民大道與民權路口，未依路口機車須兩段式左轉，而逕行左轉至民權路，正巧被停一旁停等紅燈的板橋分局警備隊警員發現將他攔下盤查。未料黃男下車出示證件供查時，卻疑因自己剛假釋出獄，擔心被警方「誤會」，一時心急下竟在大馬路上將全身衣服脫光，想證明自身清白，黃男突如其來的舉動也讓警方一度錯愕，直言他實在不必如此。過程中，黃男在一旁不斷打電話給家人及房東，表情盡顯不耐，警方也表示查明無誤後會盡速讓他離去。警方後來在機車置物箱內發現一個咖啡包及一罐裝有液體的瓶子，多問了黃男兩句，黃男為證明自己清白，竟在大馬路上將全身脫個精光，除讓警方錯愕外，也引起一旁路人側目，警方雖感錯愕，仍耐心仔細將可疑處一一查明。《蘋果》記者騎車行經路口發現黃男突兀的將衣服脫光，立即將過程拍下並待黃男將衣服穿上後上前詢問。警方表示，當天發現黃男交通違規，遂立即上前將他攔下。黃男拿出證件供查時，態度起先鎮定，但口中一直要求：「可以快點嗎？我跟房東有約」讓人感覺黃男有心虛的感覺。加上查打資料發現黃男不久前才剛因毒品罪假釋出獄，臉上有吸毒後易生的爛瘡，手上也有針孔的痕跡，機車踏板上又有數包物品，才會在黃男同意下查看他身上及車上的物品。黃男解釋咖啡包是他之前去

旅館住時順手帶走的，裝有液體的瓶子則是工作用消毒水；由於警方並未在黃男身上及車上找到明確違法的事證，除要求他趕緊將衣服穿上，否則將依妨害風化送辦外，也告誡他以後別再未依兩段式左轉而交通違規後讓其離去。

斷詞後：

新北市(Nc) 一(Neu) 名(Nf) 黃姓(Nb) 男子(Na) 上(Ng) 周三(Nb) 晚間(Nd) 8時(Nd) 許(Neqb) 騎車(VA) 行經(VCL) 板橋(Nc) 縣民(Na) 大道(Na) 與(Caa) 民權(Na) 路口(Nc) 未(D) 依(P) 路口(Nc) 機車(Na) 須(D) 兩段式(A) 左轉(VA) 而(Cbb) 逕行(D) 左轉(VA) 至(Caa) 民權路(Na) 正巧(D) 被(P) 停(VHC) 一(Neu) 旁(Ncd) 停(VHC) 等(Cab) 紅燈(Na) 的(DE) 板橋(Nc) 分局(Nc) 警備隊(Na) 警員(Na) 發現(VE) 將(P) 他(Nh) 攔下(VC) 盤查(VC) 未料(VK) 黃男(Nb) 下車(VA) 出示(VC) 證件(Na) 供(VF) 查(VE) 時(Ng) 卻(D) 疑因(Na) 自己(Nh) 剛(D) 假釋(VB) 出獄(VA) 擔心(VK) 被(P) 警方(Na) 誤會(Na) 一時(Nd) 心急(VH) 下(Ng) 竟(D) 在(P) 大(VH) 馬路(Na) 上將(Na) 全身(Nc) 衣服(Na) 脫光(VC) 想(VE) 證明(VE) 自身(Nh) 清白(VH) 黃男(Nb) 突如其來(A) 的(DE) 舉動(Na) 也(D) 讓(VL) 警方(Na) 一(Neu) 度(Nf) 錯愕(VH) 直言(VE) 他(Nh) 實在(D) 不必(D) 如此(VH) 過程(Na) 中(Ng) 黃男(Nb) 在(P) 一(Neu) 旁(Ncd) 不斷(D) 打電話(VB) 給(VD) 家人(Na) 及(Caa) 房東(Na) 表情(Na) 盡(D) 顯(VK) 不耐(VJ) 警方(Na) 也(D) 表示(VE) 查明(VE) 無誤(VH) 後(Ng) 會(D) 盡速(Na) 讓(VL) 他(Nh) 離去(VA) 警方(Na) 後來(Nd) 在(P) 機車(Na) 置物箱(Na) 內(Ncd) 發現(VE) 一(Neu) 個(Nf) 咖啡(Na) 包(VC) 及(Caa) 一(Neu) 罐裝(A) 有(V_2) 液體(Na) 的(DE) 瓶子(Na) 多(D) 問(VE) 了(Di) 黃男(Nb) 兩(Neu) 句(Nf) 黃男(Nb) 為(P) 證明(VE) 自己(Nh) 清白(VH) 竟(D) 在(P) 大(VH) 馬路(Na) 上將(Na) 全身(Nc) 脫(VC) 個(Nf) 精光(VH) 除(P) 讓(VL) 警方(Na) 錯愕(VH) 外(Ng) 也(D) 引起(VC) 一(Neu) 旁(Ncd) 路人(Na) 側目(VI) 警方(Na) 雖(Cbb) 感(VK) 錯愕(VH) 仍(D) 耐心(D) 仔細(VH) 將(P) 可疑處(Na) 查明(VE) 蘋果(Na) 記者(Na) 騎車(VA) 行經(VCL) 路口(Nc) 發現(VE) 黃男(Nb) 突兀(VH) 的(DE) 將(P) 衣服(Na) 脫光(VC) 立即(D) 將(P) 過程(Na) 拍下(VC) 並(D) 待(P) 黃男(Nb) 將(P) 衣服(Na) 穿上(VC) 後(Ng) 上前(VA) 詢問(VE) 警方(Na) 表示(VE) 當天(Nd) 發現(VE) 黃男(Nb) 交通(Na) 違規(VA) 遂(D) 立即(D) 上前(VA) 將(P) 他(Nh) 攔下(VC) 黃男(Nb) 拿出(VC) 證件(Na) 供(VF) 查(VE) 時(Ng) 態度(Na) 起先(D) 鎮定(VH) 但(Cbb) 口(Na) 中(Ng) 一直(D) 要求

(VF) 可以(D) 快點(D) 嗎(T) 我(Nh) 跟(P) 房東(Na) 有約(VH) 讓(VL) 人(Na) 感覺(Na) 黃男(Nb) 有(V_2) 心虛(VH) 的(DE) 感覺(Na) 加上(Cbb) 查打(VC) 資料(Na) 發現(VE) 黃男(Nb) 不久(Nd) 前(Ng) 才(Da) 剛(D) 因(Cbb) 毒品罪(Na) 假釋(VB) 出獄(VA) 臉(Na) 上(Ncd) 有(V_2) 吸毒(VA) 後(Ng) 易(VH) 生(VH) 的(DE) 爛瘡(Na) 手(Na) 上(Ng) 也(D) 有(V_2) 針孔(Na) 的(DE) 痕跡(Na) 機車(Na) 踏板(Na) 上(Ng) 又(D) 有數(VH) 包(VC) 物品(Na) 才(Da) 會(D) 在(P) 黃男(Nb) 同意(VK) 下(Ng) 查看(VE) 他(Nh) 身(Na) 上(Ncd) 及(Caa) 車(Na) 上(Ncd) 的(DE) 物品(Na) 黃男(Nb) 解釋(VE) 咖啡(Na) 包(VC) 是(SHI) 他(Nh) 之前(Ng) 去(VCL) 旅館(Nc) 住(VCL) 時(Ng) 順手(VH) 帶走(VC) 的(DE) 裝有(VC) 液體(Na) 的(DE) 瓶子(Na) 則(D) 是(SHI) 工作(Na) 用(P) 消毒水(Na) 由於(Cbb) 警方(Na) 並(D) 未(D) 在(P) 黃男(Nb) 身(Na) 上(Ncd) 及(Caa) 車(Na) 上(Ncd) 找到(VC) 明確(VH) 違法(VA) 的(DE) 事(Na) 證(Na) 除(P) 要求(VF) 他(Nh) 趕緊(D) 將(P) 衣服(Na) 穿上(VC) 否則(Cbb) 將(D) 依(P) 妨害(VJ) 風化(Na) 送辦(VC) 外(Ng) 也(D) 告誡(VF) 他(Nh) 以後(Nd) 別(D) 再(D) 未(D) 依(P) 兩段式(A) 左轉(VA) 而(Cbb) 交通(Na) 違規(VA) 後(Ng) 讓(VL) 其(Nep) 離去(VA)

(二) 建立自建語料庫

將斷完詞的詞彙，與原本有的 NTUSD 情緒字典做比對，如果詞彙在 NTUSD 裡面的話就直接新增到自建語料庫，如果沒有在 NTUSD 裡面的話，就必須到教育部詞典那裡，找出該詞彙的相似詞，拿該詞彙的相似詞再與 NTUSD 做比對，如果該相似詞有出現在 NTUSD 裡面的話，就將該詞彙列為與該詞的相似詞一樣的極性，然後再新增到自建語料庫。

因為要觀察哪些詞彙以及詞性是影響整篇新聞的吸引度高低，因此自建語料庫有分為兩種，一種是高吸引度語料庫，另外一種是低吸引度語料庫。

接下來就是統計哪些詞性是比較可能會影響吸引度的，然後把該詞性的詞彙

一一列出來，可以幫助報家能夠參考這些詞彙，適時的運用，達到閱聽者想要看的程度，以及哪些詞彙可能會影響整篇新聞的吸引度，盡量不要用到那些詞彙。因為同一個意思，或許很多詞彙都可以用，但是往往就是寫的最吸引的詞彙才可能會影響整個吸引度的關鍵。

(三) 計算收集字作為特徵詞

由第一步中文斷詞完之後，收集到的詞彙進行字頻的統計，經過正規化頻率(TF)以及 TF*IDF 計算後，得知每一篇新聞的關鍵字。

其中 TF、IDF、TF*IDF 公式如 p13、p14 所示。公式(1)中的 $n_{i,j}$ 為斷詞後每一個詞彙所出現的個數， $\sum_k n_{k,j}$ 為每一個詞彙頻率的加總， $tf_{i,j}$ 為每一篇斷詞後詞彙的正規化頻率(TF);公式(2)中 $\log$ 以 10 為底， $|D|$ 設為 500，因為本研究使用 500 篇蘋果新聞， $|\{j: t_i \in d_j\}|$ 為每一篇斷完詞彙在這 500 篇蘋果新聞內文中所出現的次數;公式(3)是將公式(1)及(2)的結果相乘，得到的結果可以作為新聞撰寫策略使用。

第四章 實驗結果與分析

第一節 抓取 facebook 上蘋果新聞資料

經過臉書所提供的開發者人員，可以方便快速的抓取想要的資料。每天的新聞好幾百則，如果是利用人工的方式，必須花上好幾天的時間去整理想要的資料。所以在科技資訊進步的時代，必須在有效的時間內，抓取有用的資料，來幫助作研究。所以必須有自動化的方式，能夠一次抓取想要的資料。

本研究因為是做新聞吸引度的極性分析，所以必須抓取新聞內文的部分，但是單單只有抓取內文是不夠的，因為要把所抓取下來的新聞，要分出高吸引度以及低吸引度，所以必須要有方法，可以去佐證是高吸引度或低吸引度，所以我想到利用 facebook 的讚數來當作佐證的證據。所以抓取下來的資料裡面有包含:新聞內文、新聞讚數、新聞標題、新聞的分享數、新聞的日期以及時間及新聞的留言數。不過最主要的是要看讚數以及內文的部分。

本研究抓取新聞為 500 篇，抓取一個禮拜的分量做為本研究的範本，來做為本研究的實驗結果及分析。

	A	B	C	D	E	F
1	樣本編號	日期	貼文時間	文章日期時間	新聞標題	按讚數
2	a001	5/3	23:59	2017年05月03日10:15	媒體遭嗾使只敢A女C師 朱學恒轟：政府憑什麼	475
3	a002	5/3	23:59	2017年04月19日12:05	刺殺凱撒的大密謀	315
4	a003	5/3	23:59	2017年05月03日14:34	陰莖類型有4種 你是哪一種？	2222
5	a004	5/3	23:00	2017年05月03日8:31	把小孩獨自留車上 華裔媽被痛罵：單親很辛苦	199
6	a005	5/3	23:00	2017年05月03日19:59	蚊子小黑蚊大軍來襲 用這些方式分辨兇手	273
7	a006	5/3	23:00	2017年05月03日21:38	【幼稚片】陸代表國際會議鬧台灣 澳媒批「令人厭惡」	763
8	a007	5/3	23:00	2017年05月03日23:59	【獨家】謝金燕守護豬哥亮消瘦露臉 神隱333天終現身	1050
9	a008	5/3	23:00	2017年05月03日20:55	帝王蟹吃到飽餐廳無預警關門 消費者轟惡意吸金	877
10	a009	5/3	23:00	2017年05月03日22:46	連生九仙女曾哭20年 現在幸福得不得了	8598
11	a010	5/3	23:00	2017年05月02日23:59	粵警「肥成這樣」 醉漢擋警車挑釁被壓制	384
12	a011	5/3	23:00	2017年05月03日17:14	砸53億贈女豪宅 賭王四太來台只吃路邊攤	1467
13	a012	5/3	23:00	2017年05月03日13:57	【毀容片】30強國大學生誠收被毀容 商場稱是喪屍妝	279
14	a013	5/3	22:00	2017年05月03日19:32	賓士送修「試車」58公里 老闆接小孩拒道歉	940
15	a014	5/3	21:50	2017年05月03日20:15	補教老師用本名教課？ 呂捷打算這樣做	435
16	a015	5/3	21:40	2017年05月03日21:30	【獨家】余筱萍通殺麻將百家樂賺673萬 揭余天老賭輸原因在…	408
17	a016	5/3	21:20	2017年05月03日11:15	他po文徵人「追走女友」 成功懸賞1萬元	278
18	a017	5/3	21:10	2017年05月03日16:55	陳氏太極拳傳人：魏雷是打太極操	6658
19	a018	5/3	21:00	2017年05月03日16:17	愛河支線變牛奶河 「魚發情噴X啦」網友挖苦	1184
20	a019	5/3	20:40	2017年05月03日18:11	批傳統武術後 陸格鬥狂人再批：「中醫都是騙子」	1779
21	a020	5/3	20:20	2017年05月03日20:00	【獨家】醫為理念投身司法 卻寧賠150萬堅辭檢察官	1711
22	a021	5/3	20:10	2017年05月03日12:00	【健康問蘋果】燕窩要這麼吃 肉肉才消得掉	261
23	a022	5/3	20:00	2017年05月03日19:56	狂徒致電警局詐個資 2衰警自首洩密	1293
24	a023	5/3	19:40	2017年05月03日19:50	報導林奕含案符合社會公益 北市觀傳局：可揭露名字	481
25	a024	5/3	19:20	2017年05月03日18:44	才女作家殞落 醫師募資拍電影惹議	175
26	a025	5/3	19:10	2017年05月02日13:09	【實用片】用鹽巴洗菜除農藥咁有效？真正厲害是這招	186
27	a026	5/3	19:00	2017年05月03日23:58	【百想電視】孔劉出道16年首奪戲王 《鬼怪》獲最大賞	642
28	a027	5/3	18:40	2017年05月03日18:21	台電施工不慎影響用路 北市罰6萬到15萬元	244
29	a028	5/3	18:20	2017年05月03日18:25	KO太極拳師 陸格鬥狂人找館長「交流」	3619
30	a029	5/3	18:11	2017年05月03日17:58	年改審查表決通過 公務員退休金採計最後15年	865
31	a030	5/3	17:58	2017年05月03日17:45	林奕含案事關社會公益 衛福部：此案不須隱去被害人姓名	1452
32	a031	5/3	17:40	2017年05月03日20:36	【不倫動畫】補習班名師變狼師又一例 誘高中女生車震遭逮	1885
33	a032	5/3	17:24	2017年05月03日16:50	許瑋甯「塑膠袋」上身 網友只證實一件事	1189
34	a033	5/3	17:15	2017年05月03日15:55	花13萬參加義大利團 吃不飽反映被領隊稱大食怪	802
35	a034	5/3	17:04	2017年05月03日20:23	【揭密片】補教界不能說的秘密 為什麼名師都愛化名	309

圖 15 抓取 facebook 上的資料

上圖 15 可以發現有六個欄位分別為:1.樣本編號 2.日期 3.貼文時間 4.文章日期時間 5.新聞標題 6.按讚數。

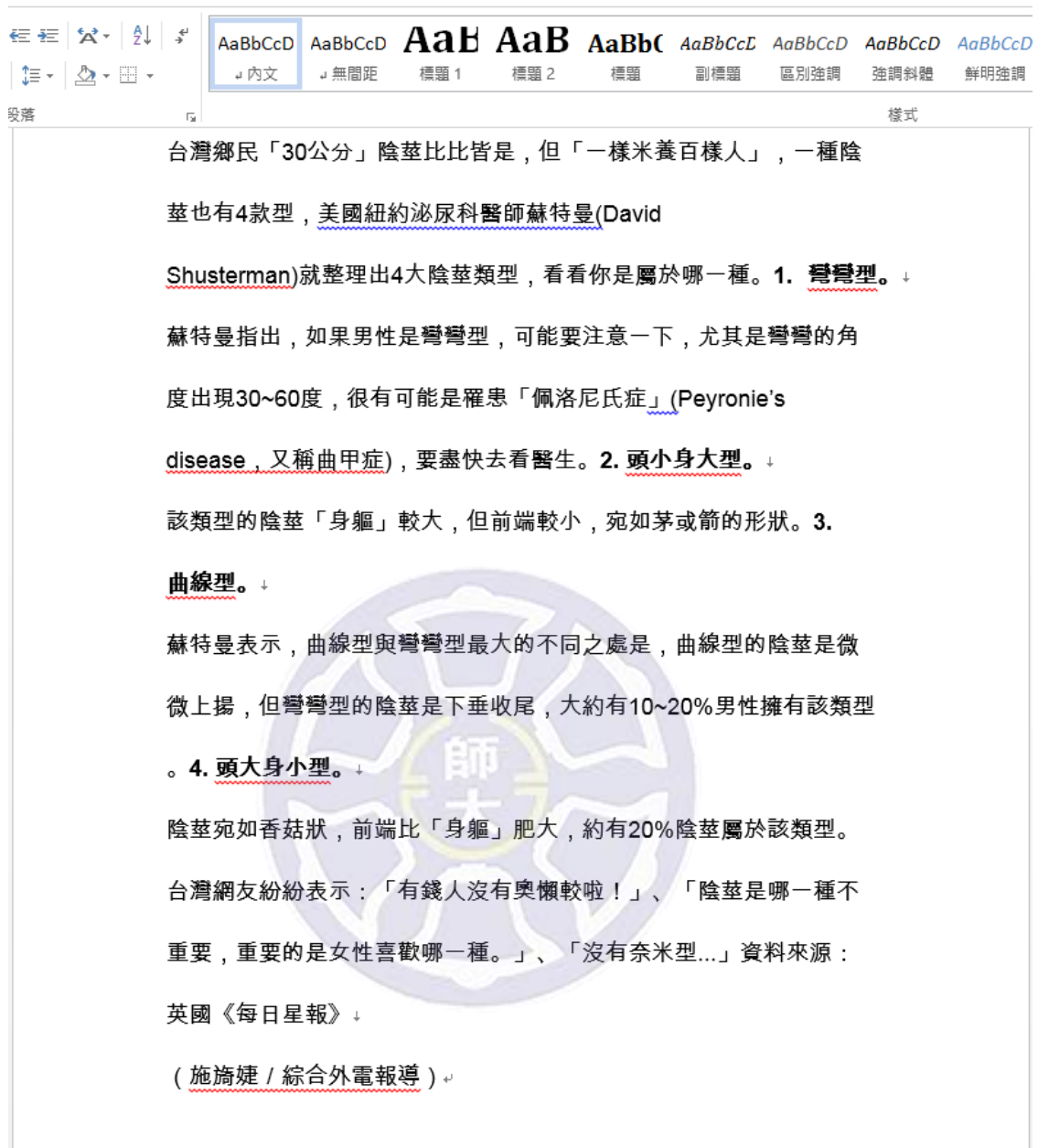


圖 16 抓取新聞內文

上圖 16 抓取整篇新聞的內文，因為之後要進行斷詞的部分，所以必須抓取整篇的內文。

第二節 內文進行斷詞

經過上述的抓取內文之後，就必須將內文開始進行斷詞，因為實驗結果的目的就是想要找出哪些詞彙是可能會影響整篇內文的吸引度。所以就必須到中研院斷詞系統，進行內文斷詞。運用此系統的主要原因是目前中文斷詞系統已經做的相當完善，可以準確且非常詳細分析本次實驗資料的所有詞性，本研究經由中研院斷詞系統後，得以順利的進行後續分析，分析出新聞高低吸引度的詞彙，然後找出什麼詞性會影響整篇新聞的吸引度。

台灣(Nc) 鄉民(Na) 「(PARENTHESISCATEGORY) 3 0 (Neu) 公分(Nf) 」(PARENTHESISCATEGORY) 陰莖(Na) 比比皆是(VH) ，(COMMACATEGORY)

但(Cbb) 「(PARENTHESISCATEGORY) 一梗(VH) 米(Na) 糞(Vc) 百(Neu) 梗(Nf) 人(Na) 」(PARENTHESISCATEGORY) ，(COMMACATEGORY)

一(Neu) 種(Nf) 陰莖(Na) 也(D) 有(V_2) 4 (Neu) 款(Nf) 型(Na) ，(COMMACATEGORY)

美國(Nc) 紐約(Nc) 泌尿科(Nc) 醫師(Na) 蘇特曼(Nb) ((PARENTHESISCATEGORY) David(Fw) Shusterman(Fw))(PARENTHESISCATEGORY) 就(D) 整理出(Vc) 4 (Neu) 大(VH) 陰莖(Na) 類型(Na) ，(COMMACATEGORY)

看看(VE) 你(Nh) 是(SHI) 屬於(VJ) 哪(Nep) 一(Neu) 種(Nf) 。(PERIODCATEGORY)

1 (Neu) -(PERIODCATEGORY) 彎彎型(Na) 。(PERIODCATEGORY)

蘇特曼(Nb) 指出(VE) ，(COMMACATEGORY)

如果(Cbb) 男性(Na) 是(SHI) 彎彎型(Na) ，(COMMACATEGORY)

可能(D) 要(D) 注意(VK) 一下(Nd) ，(COMMACATEGORY)

尤其(D) 是(SHI) 彎彎(VH) 的(DE) 角度(Na) 出現(VH) 3 0 (Neu) ~ (Fw) 6 0 (Neu) 度(Nf) ，(COMMACATEGORY)

很(Dfa) 有(V_2) 可能(D) 是(SHI) 罹患(VJ) 「(PARENTHESISCATEGORY) 佩洛尼氏症(Na) 」(PARENTHESISCATEGORY) ((PARENTHESISCATEGORY) Peyronies(Fw) disease(Fw)) ，(COMMACATEGORY)

又(D) 稱(VG) 曲甲症(Na))(PARENTHESISCATEGORY) ，(COMMACATEGORY)

要(D) 盡快(D) 去(D) 看醫生(VA) 。(PERIODCATEGORY)

2 (Neu) -(PERIODCATEGORY) 頭(Na) 小(VH) 身(Na) 大型(A) 。(PERIODCATEGORY)

圖 17 新聞內文進行斷詞

第三節 新增高吸引度與低吸引度語料庫以及擴充字典

將每一篇新聞內文斷完詞之後開始進行比對的動作，因為是做新聞的情感分析，在現有的 NTUSD 裡面，不一定會有新聞情感的詞彙，那麼在這個時候，就必須開始擴充屬於新聞的語料庫。

本研究因為是做高低吸引度的分析與探討，所以會自建兩個語料庫，一個為高吸引度語料庫，另一個為低吸引度語料庫。把 500 篇新聞，每篇新聞的讚數做一個加總平均，大於 1934 為高吸引度，低於 1934 為低吸引度，標準差值為 41.61。

在高低吸引度語料庫裡面，包含了有原本在 NTUSD 裡面所出現的詞彙，且新增了屬於新聞的詞彙。

本研究擴充新聞語料庫將以教育部重編國語辭典修訂本做為輔助，將新聞用辭的詞彙輸入到教育部重編國語辭典修訂本裡面，那麼就會出現相似詞與相反詞。如果相似詞的詞彙有出現在 NTUSD 裡面，就將原詞彙歸類成與相似詞一樣的極性，並且新增到本研究增列新聞語料庫裡面。

教育部重編國語辭典修訂本
李榮芳

網站導覽 | 其他辭典 | 常見問題
首頁 | 使用說明 | 編輯說明 | 版權頁 | 附錄 | 語文叢書 | 本版說明
國家教育研究院 維護

基本檢索 進階檢索 注音索引 筆畫索引 部首索引 學習筆記

:::歡迎將本頁固定網址 (<http://dict.revised.moe.edu.tw/cbdic/search.htm>) 設為書籤，其他問題請詳閱「本版說明」。

字詞 全文 查詢

加入學習筆記 回結果列表 意見反應 列印

字級設定: 小 中 大 巨

字詞	【完善】
注音	ㄨ ㄣ ˊ ㄕ ㄢ ˋ
漢語拼音	wán shàn
相似詞	①完美、完滿、圓滿
相反詞	①破綻、簡陋
釋義	① 完美、完好。《史記·卷一一〇·匈奴傳》：「其得漢繒絮，以馳草棘中，衣褲皆裂敝，以示不如旃裘之完善也。」 ② 完好無缺。《初刻拍案驚奇》卷五：「從陸倍道，忽遇虎負愛女至，驚逐之頃，虎去而人不傷。今完善在舟，希示進止！」

本頁網址: <http://dict.revised.moe.edu.tw/cgi-bin/cbdic/gswweb.cgi?o=dcbdic&searchid=Z0000016>

加入學習筆記 回結果列表 意見反應 列印

圖 18 教育部重編國語辭典修訂本輸出結果

檢視 台大正負向.docx [相尋模式] - Word

AaBbCcD AaBbCcD AaBbCcD AaBbCcD AaBbCcD AaBbCcD AaBbCcD AaBbCcD AaBbCcD AaBbCcD

一般 段落 樣式

一下子爆發
 一下子爆發的一連串
 巴掌
 再
 再叮囑
 拳
 般般人罪
 陣狂風
 陣緊張
 掌
 團糟
 二撥
 黑點
 蹶不振
 人事不省
 人為
 人為的
 人迷
 人迷的
 人迷的人
 刀刃
 刁難
 力盡
 匕首
 下地獄
 下垂
 下態度
 下流

圖 19 NTUSD 負向詞彙

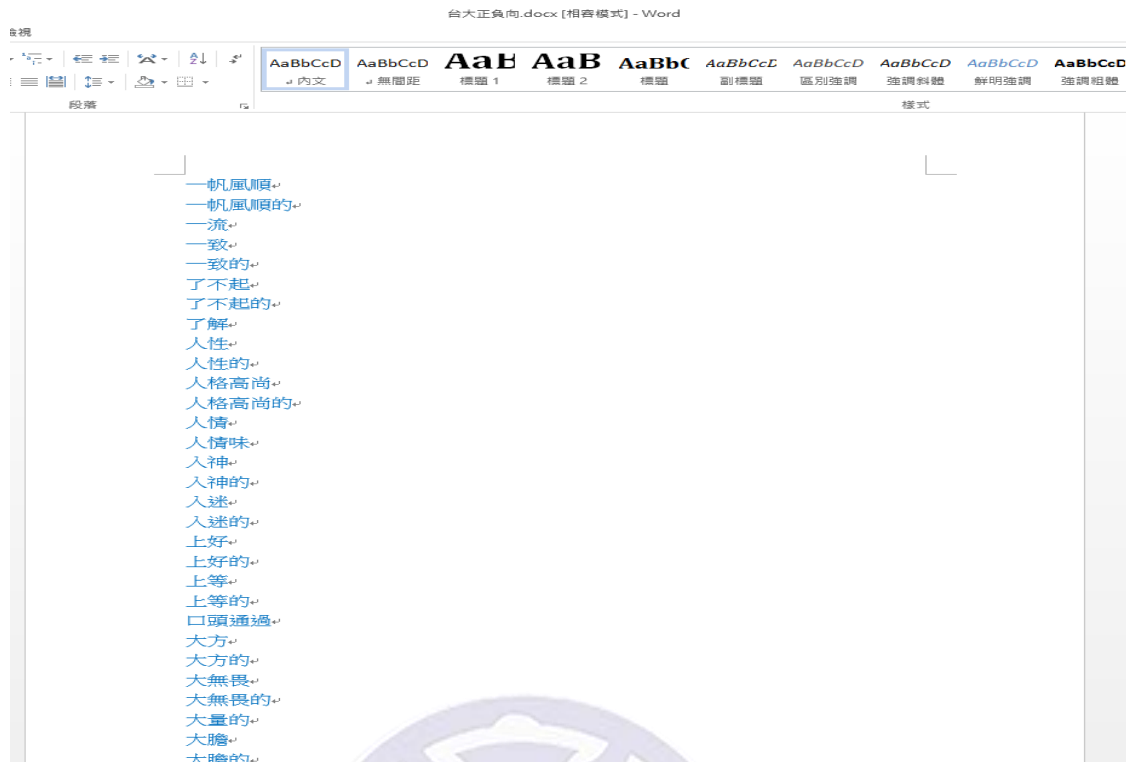


圖 20 NTUSD 正向詞彙

A. 高吸引度詞彙(節錄)

表 4 高吸引度所用的詞彙以及極性(節錄)

詞彙	極性
養(VC)	正向
罹患(VJ)	負向
下垂(VH)	負向
肥大(VH)	負向
奈米型(A)	負向
沒有(D)	負向
重要(VH)	正向
表揚(VC)	正向
壓力(Na)	負向
陰霾(Na)	負向
不得了(VH)	正向
輕生(VA)	負向
義務(Na)	正向
眼淚(Na)	負向
沒地位(A)	負向

我們可以從表 4 發現負向偏多，由此表可以為方法二帶出一點端倪，可能蘋果新聞知道現代人喜歡看負面新聞，所以撰寫的新聞負面新聞偏多，所以可以看出高吸引度文章裡面，負向偏多。且可以經由表 4 發現，負向詞彙在高吸引度詞彙表中佔了 255 個，而正向詞彙在高吸引度詞彙表中佔了 142 個，所以負向詞彙佔整個高吸引度詞彙表為 64%，正向詞彙佔整個高吸引度詞彙表為 36%。

B. 低吸引度詞彙(節錄)

表 5 低吸引度所用的詞彙以及極性(節錄)

詞彙	極性
去世(VH)	負向
參與(VC)	正向
密謀(VF)	負向
救(VC)	負向
自稱為(VG)	負向
單獨(D)	負向
超過(VJ)	負向
逮捕(VC)	負向
忽略(VJ)	負向
危險(VH)	負向
歇斯底里(VH)	負向
告誡(VF)	負向
防範(VE)	負向
傷口(Na)	負向
發炎(VH)	負向

我們可以從表 5 發現負向偏多，由此表可以為方法二帶出一點端倪，可能蘋果新聞知道現代人喜歡看負面新聞，所以撰寫的新聞負面新聞偏多，所以可以看出低吸引度文章裡面，負向偏多。且可以經由表 5 發現，負向詞彙在低吸引度詞彙表中佔了 332 個，而正向詞彙在高吸引度詞彙表中佔了 101 個，所以負向詞彙

佔整個高吸引度詞彙表為 76%，正向詞彙佔整個高吸引度詞彙表為 24%。

C. 本研究擴增的新聞詞彙

本研究擴充之新聞詞彙如表 6 所示。

表 6 擴充新聞詞彙

新聞使用的詞彙(與 NTUSD 詞彙相似)	極性	NTUSD 的詞彙
陰霾(Na)	負向	陰鬱
不得了(VH)	正向	可靠度
輕生(VA)	負向	自殺
想不開(VH)	負向	討好
浸潤(VB)	正向	討好
完善(VH)	正向	完美
防身(VA)	正向	保護
勝(VH)	正向	勝利
揭發(VC)	負向	透露
慕名而來(VA)	正向	嚮往
耐心(D)	正向	有耐性
風險(Na)	負向	危險
差別待遇(Na)	負向	不平等
精英(Na)	正向	優秀
打敗(VC)	正向	戰勝
寂靜(VH)	負向	寂寞
懸掛(VC)	負向	擔心
警戒(VA)	負向	警惕
嚴禁(VE)	負向	禁止
傻眼(VH)	負向	呆滯
肥成(VG)	負向	肥
風趣(VH)	正向	幽默
迷惘(VH)	負向	迷惑
珍惜(VJ)	正向	珍貴
霸氣(VH)	負向	專橫
衝到(VCL)	負向	衝出
錯(VH)	負向	錯誤
受損(VH)	負向	受損失
知名(VH)	正向	著名

轉型(VH)	負向	改變成
超愛(Na)	正向	喜歡
毆人(Na)	負向	毆打
樸實(VH)	正向	樸素
春風(Na)	正向	滿面春風
幫忙(VC)	正向	幫助
垂危(VH)	負向	危及
傳播(VC)	負向	會傳播的
推崇(VJ)	正向	敬重
激情(Na)	正向	熱情
抓狂(VH)	負向	失控
呵護(VC)	正向	庇護
照顧(VC)	正向	庇護
老羞成怒(VH)	負向	大發雷霆
杜絕(VC)	負向	阻止
知悉(VK)	正向	知道
解聘(VC)	負向	解僱
抱不平(VI)	負向	打抱不平
甜蜜蜜(VH)	正向	甜甜蜜蜜
猛(VH)	負向	猛烈
新(VH)	正向	新鮮
爆氣幹譙(Na)	負向	幹譙
沒有通知(D)	負向	未被通知
隱忍(VJ)	負向	忍耐
根本(D)	負向	根本就
痛毆(VC)	負向	毆打
軟趴趴(VH)	正向	柔軟
參與(VC)	正向	參加
單獨(D)	負向	孤單
攔截(VC)	負向	阻攔
關切(VJ)	正向	關心
礙於(VJ)	負向	有礙
譁然(VH)	負向	吵鬧
名氣(Na)	正向	有名
帥氣(VH)	正向	英俊
抱不平(VI)	負向	打抱不平
蠢豬(Na)	負向	笨豬

尖銳化(VHC)	負向	尖銳
淒淒慘慘(VH)	負向	淒慘
詐欺(VC)	負向	欺騙
瘦弱(VH)	負向	消瘦
意味濃厚(VH)	負向	意味深長
小心翼翼(VH)	正向	小心謹慎
致歉(VB)	負向	道歉
想念(VJ)	負向	牽掛
止不住(VL)	負向	抑制
欠一屁股(D)	負向	欠錢
拖油瓶(Na)	負向	累贅
起疑(VH)	負向	疑心
粗勇(Nb)	正向	強壯
開罰(VC)	負向	處罰
偷吃(VC)	負向	出軌
墊背(Na)	正向	犧牲
躲起來(VA)	負向	躲藏
溝通(VC)	正向	協調
標榜(VE)	正向	稱讚
典範(Na)	正向	規範
璀璨(VH)	正向	鮮豔
脾氣差(VH)	負向	壞脾氣
退化(VHC)	負向	衰退
勸說(VF)	負向	勸告
割掉(VC)	負向	割斷
人品(Na)	正向	品行
無辜(VH)	正向	無罪
汙蔑(VC)	負向	歪曲
喧擾(VA)	負向	吵雜

D. 計算出哪些詞性是會影響新聞內文撰寫

● 高吸引度詞性

表 7 計算高吸引所有詞彙出現的詞性機率

DK	1 個	0.002
Ncd	1 個	0.002
NF	1 個	0.002

Nh	2 個	0.005
VCL	2 個	0.005
VI	3 個	0.007
A	4 個	0.010
VB	4 個	0.010
VF	5 個	0.012
VL	5 個	0.012
VHC	7 個	0.017
VE	16 個	0.040
VA	17 個	0.043
D	18 個	0.046
VK	20 個	0.051
VJ	27 個	0.069
Na	60 個	0.153
VC	78 個	0.199
VH	120 個	0.315

可以從表 7 可以看出，一篇高吸引度的新聞，所使用的詞彙當中，如果他的詞性是 VH 的詞彙，會影響到新聞的撰寫。

● 低吸引度詞性

表 8 計算低吸引所有詞彙出現的詞性機率

VG	1 個	0.002
VCL	1 個	0.002
Dfa	1 個	0.002
A	2 個	0.004
NB	2 個	0.004
VL	3 個	0.004
VI	4 個	0.009
VB	5 個	0.011
VF	8 個	0.018
VHC	10 個	0.023
VE	13 個	0.030
D	14 個	0.032
VK	18 個	0.042
VJ	29 個	0.068

VA	35 個	0.082
NA	63 個	0.147
VC	72 個	0.169
VH	145 個	0.340

可以從表 8 可以看出，一篇高吸引度的新聞，所使用的詞彙當中，如果詞性是 VH 的詞彙，將會影響到新聞的撰寫。

上述做出來的實驗結果，可以發現到東莞台商子弟學校的老師吳建華(www.td-school.org.cn/3ws0608/edit/文宣編輯：新聞寫作.ppt)他教授的課程：文宣編輯與行銷課程中可以發現，吳老師所說的新聞寫作技巧中說到，新聞報導使用的動詞、名詞較多，形容詞、副詞較少(如圖 21)。

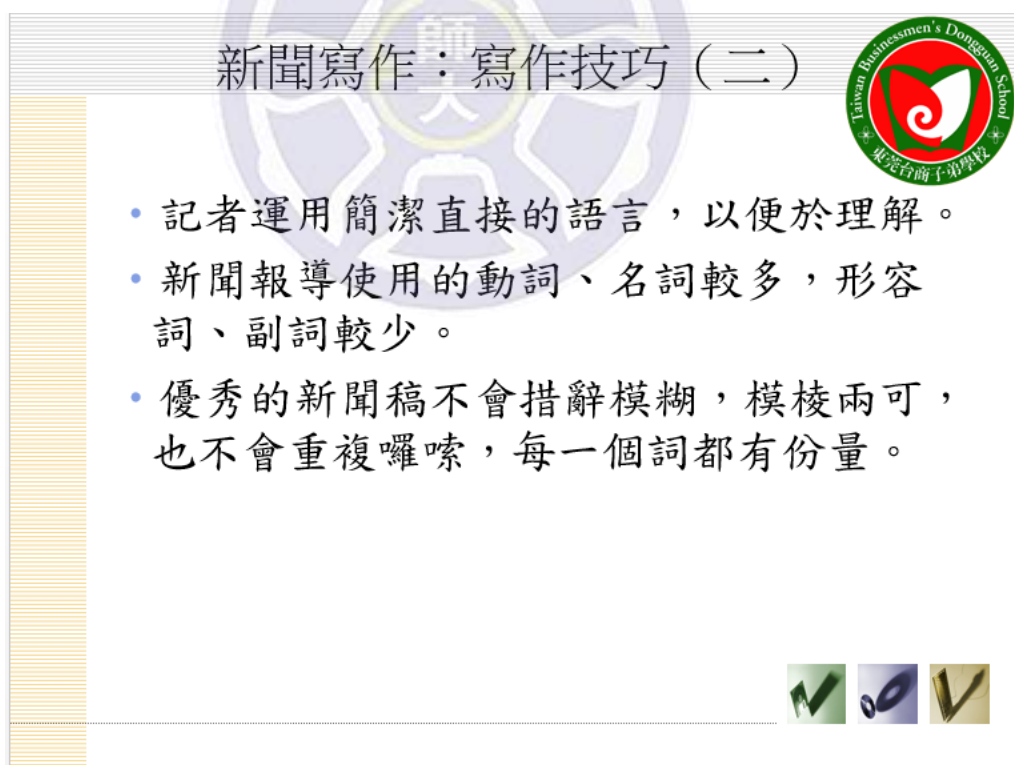


圖 21 新聞寫作之寫作技巧(ref:吳建華老師之新聞寫作技巧)

如果不是寫新聞內文的話，探討一般小說類之寫作技巧如下。參考巴哈姆特

作家亞蘇之資料，寫不同題材的作品，在人物描述上一定要做出差異。以東方小說為例，寫東方古代小說最大的優勢，就是「所有的」中文「動詞」都能夠輕易的使用，動詞變化絕對是極大的優勢，例如古時候的妝容、強調寫意美與神態美的情況會多過於形容體態或是實際的容貌，由此顯示中文寫作上動詞之重要性。

E. 高吸引度(VH)用的詞彙

本研究高吸引度(VH)用的詞彙如下所示：

下垂(負向)、肥大(負向)、重要(正向)、不得了(正向)、想不開(負向)、福氣(正向)、幸福(正向)、心酸(負向)、獨立(正向)、完善(正向)、敗(負向)、健康(正向)、突出(正向)、勝(正向)、獲勝(正向)、高傲(負向)、不斷(負向)、偉大(正向)、詳細(正向)、合作(正向)、假(負向)、莽撞(負向)、親密(正向)、可怕(負向)、可愛(正向)、寂靜(負向)、失敗(負向)、自豪(正向)、相反(負向)、清晰(正向)、不佳(負向)、不適(負向)、不切實際(負向)、過世(負向)、協力(正向)、安全(正向)、傻眼(負向)、隨便(負向)、不道德(負向)、清潔(正向)、無奈(負向)、怒(負向)、狼藉(負向)、傷心(負向)、爆破(負向)、不幸(負向)、罹難(負向)、殘酷(負向)、爆炸(負向)、僵持(負向)、懊悔(負向)、華麗(正向)、風趣(正向)、腦弱(負向)、低潮(負向)、順利(正向)、樂觀(正向)、沮喪(負向)、迷惘(負向)、卑微(負向)、霸氣(負向)、慢(負向)、沒關係(負向)、錯(負向)、受損(負向)、突發(負向)、知名(正向)、不便之處(負向)、轉型(負向)、嚇人(負向)、安安穩

穩(正向)、快快樂樂(正向)、痛(負向)、溫文儒雅(正向)、標緻(正向)、樸實(正向)、垂危(負向)、成功(正向)、不足(負向)、足夠(正向)、憤(負向)、真實(正向)、平靜(正向)、哽咽(負向)、抓狂(負向)、悲傷(負向)、受害(負向)、酒醉(負向)、老羞成怒(負向)、噙(負向)、有種(負向)、超爽(正向)、憂心(負向)、重傷(負向)、發狂(負向)、骨折(負向)、束手無策(負向)、清秀(正向)、情不自禁(正向)、可惜(負向)、戀愛(正向)、甜蜜蜜(正向)、圓滿(正向)、猛(負向)、新(正向)、失調(負向)、無力(負向)、容易(正向)、充分(正向)、受驚(負向)、失望(負向)、狼狽(負向)、過分(負向)、一面倒(負向)、可笑(負向)、舒服(正向)、大方(正向)、愚蠢(負向)、軟趴趴(正向)、過量(負向)

F. 低吸引度(VH)詞彙

本研究低吸引度(VH)用的詞彙如下所示：

去世(負向)、危險(負向)、歇斯底里(負向)、發炎(負向)、不易(負向)、充足(正向)、粗暴(負向)、病逝(負向)、消瘦(負向)、倒店(負向)、忍無可忍(負向)、奢華(負向)、盡興(正向)、逗趣(正向)、免費(正向)、潰爛(負向)、氣憤(負向)、歡樂(正向)、尷尬(負向)、誠摯(正向)、光明正大(正向)、假惺惺(負向)、譁然(負向)、純潔(正向)、狂妄(負向)、完蛋(負向)、熱忱(正向)、優秀(正向)、聰明(正向)、大樞呆(負向)、故障(負向)、聰慧(正向)、早逝(負向)、憤慨(負向)、驚醒(負向)、不好意思(負向)、薄弱(負向)、過量(負向)、激動(負向)、浪漫(正

向)、幸運(正向)、沉重(負向)、有趣(正向)、徬徨(負向)、無意義(負向)、緊急(負向)、混亂(負向)、重大(正向)、帥氣(正向)、堅決(正向)、精緻(正向)、不耐煩(負向)、積極(正向)、進取(正向)、狼狽為奸(負向)、燦爛(正向)、恐怖(負向)、不肖(負向)、分崩離析(負向)、淒淒慘慘(負向)、艱難(負向)、荒謬(負向)、愁眉苦臉(負向)、心有不甘(負向)、瘦弱(負向)、疲憊(負向)、意味濃厚(負向)、嚴密(正向)、失血(負向)、熱烈(正向)、謙虛(正向)、雙喜臨門(正向)、小心翼翼(正向)、不得已(負向)、生氣(負向)、嚴格(負向)、通貨膨脹(負向)、自由(正向)、好奇(正向)、不負責任(負向)、迅速(正向)、瘋狂(負向)、甜蜜(正向)、詭異(負向)、毛骨悚然(負向)、破裂(負向)、開心(正向)、成長(正向)、懂事(正向)、主動(正向)、超現實(負向)、緩慢(負向)、起疑(負向)、明顯(正向)、無知(負向)、莫名其妙(負向)、暴怒(負向)、出軌(負向)、起疑心(負向)、呆(負向)、曖昧(負向)、頻繁(負向)、亂搞(負向)、合理(正向)、受傷(負向)、發燒(負向)、不適(負向)、身亡(負向)、幽默(正向)、努力(正向)、患得患失(負向)、璀璨(正向)、不便(負向)、脾氣差(負向)、孤僻(負向)、想不開(負向)、變態(負向)、平安(正向)、包藏禍心(負向)、正確(正向)、無奈(負向)、值得(正向)、無辜(正向)、自私(負向)、昏迷(負向)、不睦(負向)、正常(正向)、惡劣(負向)、複雜(負向)、受損(負向)、傷天害理(負向)、破碎(負向)、細緻(正向)、稀薄(負向)、頭暈(負向)、惡化(負向)、斷裂(負向)、邪惡(負向)、猥褻(負向)、倒閉(負向)、失業(負向)、輕率(負向)、激烈(負向)、進退兩難(負向)、瞎眼(負向)

從高吸引度 VH 裡面可以發現"過世"這個詞彙與低吸引度 VH 裡面"去世"這個詞彙兩者是相似詞，但是如果使用"過世"這個詞的話，可以發現得到的讚數是比使用"去世"的讚數來的高，所以如果要用"去世"這個詞彙的話，新聞記者可以試著用"過世"這個詞彙，或許可以提高吸引度來吸引閱聽者。

G. 擴充高吸引度語料庫的準確率

從隨機新聞當中，隨機抽 20 篇高吸引度新聞做測試，目的為了看出本研究所建立的高吸引度語料庫的準確度為何。本研究會先計算高吸引度文章裡頭，平均一篇出現幾個高吸引度詞彙(自建語料庫)，實驗結果顯示，平均一篇大概會有 16 個詞彙會出現在自建語料庫裡。

本研究提出，三個規則以便判斷一篇新聞是否為高吸引度文章，以下為三個步驟：

1. 如果高吸引度詞彙數>平均一篇高吸引度詞彙數，則判斷為高吸引度文章。
2. 低吸引度詞彙數<高吸引度詞彙數，則判斷為高吸引度文章。
3. 如果高吸引度數並未大於平均一篇高吸引度詞彙數，但低吸引度詞彙數為 0 或者小於高吸引度詞彙數，則判斷為高吸引度文章。

表 9 高吸引度實驗結果

新聞	高吸引度詞彙 數	低吸引度詞彙 數	平均一篇高吸 引度詞彙數	實驗結果
1	17	0	16	是
2	22	3	16	是
3	10	0	16	是
4	26	5	16	是
5	21	1	16	是
6	31	6	16	是
7	23	2	16	是
8	27	3	16	是
9	33	4	16	是
10	27	3	16	是
11	11	15	16	否
12	7	11	16	否
13	16	1	16	是
14	19	2	16	是
15	5	10	16	否
16	21	4	16	是

17	13	19	16	否
18	23	2	16	是
19	9	9	16	否
20	22	1	16	是

從表 9 當中可以發現紅色部分，總共有五筆資料是有問題的，從這些錯誤資料當中可以發現，雖然高吸引度新聞使用低吸引度詞彙數偏多，但可以從這些低詞彙當中發現負向極性偏多，可以從圖斯勒(Marc Trussler)和索羅卡(Stuart Soroka)這兩位學者指出，人們會偏好負向新聞較多，所以才會導致資料錯誤，雖然使用低吸引度詞彙，卻一樣可以達到高吸引度。本研究高吸引度之準確率為： $(20-5)/20$ ，為 75%。其中分子為 20 篇新聞當中扣掉錯誤資料 5 篇，分母為隨機抽樣的 20 篇新聞。

H. 擴充低吸引度語料庫的準確率

從隨機新聞當中，隨機抽 20 篇低吸引度新聞做測試，目的為了看出本研究所建立的低吸引度語料庫的準確度為何。本研究會計算低吸引度文章裡頭，計算平均一篇出現幾個低吸引度詞彙(自建語料庫)，實驗結果顯示，平均一篇大概會有 11 個詞彙會出現在自建語料庫裡。

會有三個規則會去判斷，一篇新聞是否為低吸引度文章，以下為三個步驟：

1. 如果低吸引度詞彙數>平均一篇低吸引度詞彙數，則判斷為低吸引度文章。

2. 低吸引度詞彙數>高吸引度詞彙數，則判斷為低吸引度文章。
3. 如果低吸引度數並未大於平均一篇低吸引度詞彙數，但高吸引度詞彙數為 0 或者小於低吸引度詞彙數，則判斷為低吸引度文章。

表 10 低吸引度實驗結果

新聞	高吸引度詞彙 數	低吸引度詞彙 數	平均一篇低吸 引度詞彙數	實驗結果
1	2	15	11	是
2	4	19	11	是
3	12	7	11	否
4	1	16	11	是
5	3	21	11	是
6	7	32	11	是
7	13	3	11	否
8	17	7	11	否
9	3	14	11	是
10	5	27	11	是
11	13	3	11	否
12	2	17	11	是
13	1	13	11	是

14	4	25	11	是
15	3	19	11	是
16	11	2	11	否
17	4	21	11	是
18	17	11	11	否
19	5	27	11	是
20	3	25	11	是

從表 10 當中可以發現紅色部分，總共有六筆資料是有問題的，從這些錯誤資料當中可以發現，雖然低吸引度新聞使用高吸引度詞彙數偏多，但可以從這些低詞彙當中發現正向極性偏多，可以從圖斯勒(Marc Trussler)和索羅卡(Stuart Soroka)這兩位學者指出，人們會偏好負向新聞較多，所以才會導致資料錯誤且導致文章為低吸引度。本研究高吸引度之準確率為： $(20-6)/20$ ，為 70%。其中分子為 20 篇新聞當中扣掉錯誤資料 6 篇，分母為隨機抽樣 20 篇的新聞。

第四節 TF-IDF 在新聞內文裡面的關鍵字

從高吸引度文章當中，去看閱聽者是比較偏好負面新聞或者正向新聞，所以利用 TF-IDF 的方式，取出每篇文章當中前五個的關鍵字，然後看這五個是正向詞偏多或者負向詞偏多，如果正向詞偏多，就把他歸類為正向新聞，如果是負向詞偏多，就把他歸類為負向新聞。

為了探討原因，本研究從研究者圖斯勒(Marc Trussler)和索羅卡(Stuart Soroka)在加拿大麥基爾大學(McGill University)實驗可以與本論文相呼應。

之前其實也有研究探討人類是怎麼理解新聞的，但圖斯勒和索羅卡認為這些研究結果並不能滿足他們，舉例來說，有的研究讓受試者直接在家裡看新聞，這讓學者根本沒辦法確定到底是誰在用電腦，整個實驗過程沒有受到控制；有的實驗則是不切實際，邀請受試者到研究室選出想看的新聞，這讓受試者清楚知道他們選出來的新聞會受到學者嚴格的檢查，影響實驗結果。所以研究團隊決定試試另一個新方法：詐騙。

詐騙研究法

圖斯勒和索羅卡邀請他們學校的人來研究室進行『眼球追蹤研究』，自願參加的人會先被要求從新聞網站上選一些政治新聞來讀，這樣鏡頭才可以做一些基本的眼球追蹤測量。研究人員會告訴受試者這階段非常重要，他們要真的讀新聞，這樣才可以準確測量，不過實際上，他們讀哪篇新聞並不重要。

在『準備』完成之後，受試者會開始看一支短片，受試者會認為這支短片才是最主要的實驗，但短片其實只是個幌子。最後受試者會再根據他們想看的政治新聞，回答問題。

實驗結果和他們選擇閱讀的故事可以說是有點『憂鬱』，受試者常會選擇用負面角度報導的新聞來讀，像是有貪污、倒退、虛偽等負面詞的新聞，而比較少選中性或是正面的故事，對政治和時事特別有興趣的人，更會去選擇負面新聞。

有趣的是，當你問受試者他們喜歡哪種新聞，他們會說自己比較愛正面的新聞。另外，平均來說，受試者大多認為媒體太常報導負面新聞。

研究團隊認為他們的實驗可以說是證明人類有「負面偏好」(negativity bias)的證據，負面偏好是心理學上的一個詞，意思是人類都特別喜歡聽到並記下壞消息。

人會喜歡收集壞消息，並不是出自幸災樂禍的心態，而是因為已經演化成對潛在危機要能夠快速反應。壞消息可能會是一個信號，告訴該改變現在在做的事情，才能避開危險。

結果發現比起『寶寶』、『微笑』、『有趣』這類快樂又更常見的單字，人在看到『癌症』、『炸彈』、『戰爭』等負面單字時，可以用更快的速度按下按鈕。另外，比起正面的字，能夠更快認出負面單字，甚至是在確實認出下個單字前就預先知道單字是負面的。

由上面學者的實驗探討，可以發現現代人們會因為一些比較負面性的詞彙，選擇去觀看，反而一些正向性的詞彙，比較不會想要特別的關注。現代人的心態都處於負面的，所以一旦有負面的新聞，就會特別去關注報導。

表 11 TF-IDF 之新聞判斷正負向(取出高吸引新聞部分)

編號	TF-IDF	正向或負向	讚數
3	下垂 負向 肥大 負向 奈米型 負向 沒有 負向 重要 正向	負向	2222
9	幸福 正向 心酸 負向 辛苦 負向	正向	8598

	福氣 正向 表揚 正向		
17	信任 正向 打架 負向 鬥毆 負向 尊重 正向 肯定 正向	正向	6658
28	慕名而來 正向 挑戰 負向 贏 正向 輸 負向 合作 正向	正向	3619
38	耐心 正向 親密 正向 可愛 正向 可怕 負向 差別待遇 負向	正向	21000
40	精英 正向 打敗 正向 自豪 正向 清晰 正向 掌握 正向	正向	6702
43	抱怨 負向 吵 負向 傻眼 負向 隨便 負向 不道德 負向	負向	6217
45	爆破 負向 炸死 負向 罹難 負向 陰謀 負向 批評 負向	負向	3459
53	挑釁 負向 肥成 負向 破口大罵 負向 尊嚴 正向 妨害 負向	負向	6677
54	誘姦 負向	負向	2827

	腦弱 負向		
55	樂觀 正向 沮喪 負向 期盼 正向 珍惜 正向 夢想 正向	正向	2177
56	霸氣 負向 大吼 負向 滿意 正向 受損 負向 道歉 負向	負向	10001
64	知名 正向 轉型 負向 誇張 負向 不便之處 負向 嚇人 負向	負向	8479
67	虐待 負向 引誘 負向 誘姦 負向 批露 負向 符合 正向	負向	8015
76	教唆 負向 毆人 負向 爆出 負向 臭臉 負向 春風 正向	負向	4654
85	幫忙 正向 成功 正向 穩定 正向 肆虐 負向 暴雨 負向	正向	3877
87	打倒 負向 騙子 負向 喜愛 正向 誇大 負向 推崇 正向	負向	3503
91	激情 正向 憤 負向	正向	2943

	保護 正向		
92	呵護 正向 照顧 正向 平靜 正向 加害 負向 受害 負向	正向	2237
94	假裝 負向 酒醉 負向	負向	4156
98	靠北 負向 老羞成怒 負向 不理 負向 壞人 負向 超爽 正向	負向	2366
115	強暴 負向 色狼 負向 包庇 負向 不適任 負向 憂心 負向	負向	3534
118	咬 負向 重傷 負向 狠狠 負向 咬住 負向 發狂 負向	負向	2134
119	震驚 負向 戀愛 正向 甜蜜蜜 正向 圓滿 正向 分手 負向	正向	2254
139	適度 正向 增加 正向 幫助 正向 失調 負向 癌症 負向	正向	4244
209	妨礙 負向 隱忍 負向 受驚 負向 爆氣幹譙 負向 逼不得已 負向	負向	2475

210	正義魔人 負向 不如 負向 漠視 負向 可笑 負向 鼓勵 正向	負向	2574
136	老實說 正向 相信 正向	正向	2035
140	凝聚 正向 反對 負向 欣賞 正向 大方 正向 奪冠 正向	正向	2068
147	痛毆 負向	負向	3714
159	愚蠢 負向	負向	2362
173	疾病 負向 障礙 負向 過量 負向 改善 正向 戒除 正向	負向	3023
109	淚崩 負向 辛苦 負向 累 負向 心疼 負向 開心 正向	負向	1392
110	抱怨 負向 隱瞞 負向 拖油瓶 負向 傷人 負向 隱藏 負向	負向	144
111	搖晃 負向 擠滿 負向 緩慢 負向 傻眼 負向 開罰 負向	負向	231
112	爆破 負向 死 負向 殘酷 負向 陰謀 負向	負向	1183

	罹難 負向		
113	騙 負向 無知 負向 嚴重 負向 感嘆 負向 死 負向	負向	231
114	莫名其妙 負向 出軌 負向 疑心 負向 頻繁 負向 抓狂 負向	負向	1372
164	守法 正向 優質 正向 優先 正向 怒批 負向 幫助 正向	正向	7103
182	謝謝 正向 大讚 正向 感恩 正向 善良 正向 挑戰 負向	正向	9094

從表 11 中，從高吸引度文章裡面去做一個統計，為了要看出現代人是喜歡正向新聞或者負向新聞。將高吸引度文章每篇都取 TF-IDF，找出每一篇的關鍵詞，本研究是取出前五名的 TF-IDF 的關鍵詞，取完之後，去觀察這五個詞彙的極性，如果是正向極性偏多的話，就將這篇新聞歸類為正向新聞，如果是負向極性偏多的話，就將這篇新聞歸類為負向新聞。

方法二一樣會參考 facebook 的讚數，實驗結果會將負向新聞的讚數總數加總，正向新聞的讚數加總，然後做一個平均的動作，可以發現負向新聞的平均讚數是

高於正向新聞的讚數，所以可以從方法二得知現代人是偏好負向新聞，對於負向新聞會特別的關注。平均讚數公式如下：

負向: $86945/40= 2173.625$ 負向全部讚數加總/高吸引度個數 (1)

正向: $84600/40= 2115$ 正向全部讚數加總/高吸引度個數 (2)



第五章 結論與未來展望

本論文利用情緒分析的技術，分析現在網路新聞內文經常使用哪些詞彙或語句，可激發閱聽人的情緒反應以增加其閱讀興趣。從方法一可以得知高吸引度及低吸引度新聞所使用的詞彙有哪些，以及顯示高吸引度及低吸引度新聞所使用的詞性為 VH 以及 VC，其中使用 VH 為最高使用的詞性，其次為 VC。而從東莞台商子弟學校的老師吳建華他教授的課程：文宣編輯與行銷課程中可以發現，吳老師所說的新聞寫作技巧中說到，新聞報導使用的動詞、名詞較多，形容詞、副詞較少，此實驗亦與本研究得到的結論一致。

方法二的結果顯示，現代人對於負面新聞會比較有興趣，本論文利用 TF-IDF 的方法，並且配合粉絲專頁的讚數，得知一家報導通常會比較喜歡撰寫負面新聞報導居多。在方法二的研究數據當中，發現負向新聞所平均出來的值為 2173.625，高吸引度所平均出來的值為 2115，所以可以看出，負向新聞的讚數是多於正向新聞的。而從研究者圖斯勒(Marc Trussler)和索羅卡(Stuart Soroka)在加拿大麥基爾大學(McGill University)進行實驗中發現，其實在心理學上，人類有「負面偏好」(negativity bias)的證據，負面偏好是心理學上的一個詞，意思是人類都特別喜歡聽到並記下壞消息，此實驗亦與本研究得到的結論一致。

本研究因為是抓取蘋果日報一個禮拜的新聞量，一家報導，所以可能有很多類別的主題所組成，例如有：社會版、影音版、體育版等等。但是本研究抓取的資

料面向，是所有組成的實驗結果。所以在未來的研究上，可以分別去做每一家報導的主題類別，或許出來的結果會不同。本研究做出來的結果現代人是偏好於負面新聞，但是本研究是所有主題所組成得知的實驗結果，如果今天只做社會版的結果，或許出來的結果不一定現代人喜歡負面新聞居多。所以在未來，希望能夠達到依照本研究的方法，去擴充每一個類別主題的新聞，去得知不一樣的結果。而本研究是抓取蘋果新聞，在未來，也希望能夠有其他報家的研究與實驗結果，來看每一家報導所撰寫的不一樣之處。



參考文獻

Bollier, D., & Firestone, C. M. (2010). The promise and peril of big data.

Washington, DC:Aspen Institute, Communications and Society Program.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, 15(5), 662-679.

Li, S., He, H., Xu, W. R., & Guo, J. (2009, July). Automatic Chinese sentiment word extraction based on Aximum Entropy. In *Wavelet Analysis and Pattern Recognition*, 2009. ICWAPR 2009. International Conference on (pp. 437-441). IEEE.

Lu, B., & Tsou, B. K. (2010, July). Combining a large sentiment lexicon and machine learning for subjectivity classification. In *Machine Learning and Cybernetics (ICMLC)*, 2010 International Conference on (Vol. 6, pp. 3311-3316). IEEE.

Russim, P. (2011). Big data analytics. TDWI Best Practices Report, Fourth Quarter, 1-35.

Sui, H., Jianping, Y., Hongxian, Z., & Wei, Z. (2012, December). Sentiment analysis of Chinese micro-blog using semantic sentiment space model. In Computer Science and Network Technology (ICCSNT), 2012 2nd International Conference on (pp. 1443-1447). IEEE.

Wang, J. H., & Lee, C. C. (2011, October). Unsupervised opinion phrase extraction and rating in Chinese blog posts. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on (pp. 820-823). IEEE.

Wang, B., Min, Y., Huang, Y., Liu, Y., Li, X., Sun, Y., & Sun, C. (2013, March). Chinese reviews sentiment classification based on quantified sentiment lexicon and fuzzy set. In Information Science and Technology (ICIST), 2013 International Conference on (pp. 677-680). IEEE.

Yang, Y., & Zhou, Y. (2011, December). Chinese sentiment classification based on semantic structure of sentences. In Computer Science and Network Technology (ICCSNT), 2011 International Conference on (Vol. 3, pp. 1745-1749). IEEE.

Zhai, Z., Liu, B., Wang, J., Xu, H., & Jia, P. (2012). Product feature grouping for opinion mining. *IEEE Intelligent Systems*, 27(4), 37-44.

Zhai, Z., Xu, H., & Jia, P. (2010). An empirical study of unsupervised sentiment classification of chinese reviews. *Tsinghua Science & Technology*, 15(6), 702-708.

Zhang, H., Yu, Z., Xu, M., & Shi, Y. (2012, October). An Improved Method to Building a Score Lexicon for Chinese Sentiment Analysis. In *Semantics, Knowledge and Grids (SKG), 2012 Eighth International Conference on* (pp. 241-244). IEEE.

Zhuo, S., Wu, X., & Luo, X. (2014, August). Chinese text sentiment analysis based on fuzzy semantic model. In *Cognitive Informatics & Cognitive Computing (ICCI* CC), 2014 IEEE 13th International Conference on* (pp. 535-540). IEEE.

一次搞懂大數據（上）

<https://www.bnext.com.tw/article/35807/BN-2015-03-31-151014-36>

中研院斷詞系統

<http://ckipsvr.iis.sinica.edu.tw/>

呂瑞男. (2015). 大數據分析於行銷策略應用之研究. 中山大學高階經營碩士班學位論文, 1-39.

李凱斌(2011). 使用 Google Adwords 之效益分析－以某工業電腦廠商為例。碩士論文，元智大學。

結巴斷詞

<https://github.com/fxsjy/jieba>

葉乃嘉(2013). 大數據運用與企業競爭力之探索性研究。碩士論文，國立台灣大學。

趙麗慧(2014). 教育心理學－三化取向。台北：東華。

游琇雯(2015). 使用情緒分析於社群論壇消費者評論滿意度評估之研究－以 TripAdvisor 旅遊網站為例。碩士論文，國立中興大學。

維基百科:TF-IDF 介紹

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