106000000AU130006

挖掘網路世界的文字寶藏

─文字探勘與民意調查結合應用

內政部統計處自行研究報告

中華民國 107 年 8 月

本報告內容及建議，純屬研究人員意見，不代表本機關意見

目次

表次

圖次

摘要

關鍵字：CATI、文字探勘、中文斷詞、情緒分數、支持向量機(SVM)、影響度評估模型、輿情滿意度、沉默螺旋理論。

目前坊間的電話調查(CATI)準確度屢受質疑，原因除問項過多、題意艱澀、拒訪嚴重等問題外，尚有涵蓋面不足的問題。雖然如此，CATI 仍為調查 50 歲以上族群的最佳工具，不宜輕言廢棄；惟當今智慧型手機普及，年輕族群使用市話比例逐漸降低，造成調查結果不夠準確。

近年來，網路留言、討論與分享等方式構築了不同於以往的民意形式，而這些民意表達卻是電話調查無法蒐集到的，需透過網路輿情探勘方式予以蒐集、釐清、歸納、判讀及分析。

本篇分析之目的在於利用內政統計名詞、內政概要、1996 客服紀錄等相關資料建立內政領域詞庫，搭配網路上各種正負向詞庫及斷詞系統，針對國內 1 萬 5,000 多個網站進行文本資料撈取，將其文本去雜訊後再進行分類，計算其情緒分數。

另為精煉文本，提高其準確度，採用支持向量機法(SVM)建構文本對內政部的影響度評估模型，排除對內政部沒有影響的文章後，透過網路輿情分析及文字探勘評估政府的施政滿意度，以 106 年輿情資料為分析主軸，搭配已辦理的 106 年內政部施政滿意度調查，初探「地政」及「營建」業務，建立一套媒體與網民為主架構的輿情滿意度計算方法，從而分析兩項業務的施政滿意及不滿意情況，提供決策者更務實且貼近民意的施政參據。

Abstract

key words：CATI, Text Mining, Thesaurus, Chinese word segmentation, emotional score, SVM, Impact Assessment Model, Public satisfaction, Silent spiral theory.

Recently, Computer Assisted Telephone Interview (CATI) has been repeatedly questioned because of numerous questions, complicated intentions, seriously refusals, and lacks of coverage. Though, CATI is still the optimal tool to survey the population of over 50 years old, and it shouldn’t be abandoned. Nowadays, the smart phone is widespread, and the proportion of young group using tradional telephone is declined gradually. And this causes the result of survey inaccurate.

Comments, discussions and shares construct the different forms of public opinion, which could be collected by Text Mining rather than CATI.

The purpose of this analysis is to obtain the text information on 15,000 websites in Taiwan by using the vocabulary of the Interior Field and all kinds of Thesauruses and Chinese word segmentation system. Then classify the texts and calculate the emotional scores.

In order to refine the texts and improve the accuracy, we construct the Impact Assessment Model by Support Vector Machine（SVM）to eliminate the texts that have no influence to MOI. Then evaluate governance satisfaction of “Land administration” and “Construction and Planning” through Text Mining by social media information and MOI governance satisfaction survey in 2017. We also set up a method to calculate satisfaction based on media and netizens with public opinion data and governance satisfaction survey of MOI in 2017. It can analyze the satisfaction and dissatisfaction of two functions to provide decision makers with consultations which are more pragmatic and closer to public opinion.

挖掘網路世界的文字寶藏
-文字探勘與民意調查結合應用

第一章 前言

有鑑於2017年英國大選,民調顯示保守黨可輕鬆取得多數席次,但結果卻是未過半。2016年美國總統大選結果,主流媒體與傳統民調預測選情多顯示希拉蕊領先,但結果卻由川普當選,跌破各界眼鏡;同年英國6月脫歐公投,公投前傳統機構民調、市場預測、賭盤趨勢一面倒預測「留歐」勝出,但結果卻剛好相反。2015年3月以色列大選,總理成功二度連任;同年6月西班牙大選,極左派慘遭滑鐵盧;7月澳洲大選現任總理與執政黨贏得過半席次;9月希臘國會大選,因推翻紓困案公投結果而下台的前總理又班師回朝;10月加拿大自由黨取得歷史性大勝等等各國大選結果,都與民調預測大不相同,如此民調失準的情況使各界開始質疑電話民調(CATI)結果的可信度,學界及業界也紛紛投入研究造成民調偏差的原因,並尋找可行的解決方案。

第二章 研究目的

民調結果偏差的現象不只出現在選舉民調中，政府的滿意度民調亦有相同情況，除了電話民調問項過多、題意艱澀不易瞭解、拒訪情況嚴重等問題外，尚有市話電訪（CATI[1]）涵蓋面不足的問題。不可諱言的，CATI 目前仍為調查 50 歲以上族群的最佳工具，不宜輕言廢棄，但如今智慧型手機普及，年輕族群使用市話比例逐漸降低，電話訪問 39 歲以下的族群涵蓋率亦不到 6 成，這應該是造成調查結果不夠準確的重要原因。綜合以上的現象，不得不重新檢視調查工具的可信度，各種調整、取代的觀念及看法也紛紛被提出；加上近年來，網路的蓬勃發展造就許多網路社群平台百花齊放，網路留言、討論與分享架構出不同於以往的民意形式，而這些民意呈現卻是以電話民調無法蒐集得到的。

隨著網路科技的進步，資料儲存及讀取等技術的完善，過去幾年間數位資料的數量以倍數的方式累積，形成所謂的巨量資料（Big Data），而過去我們對某些既定的限制習以為常，以為別無他法，但其實只是受限於規模的不足及資料處理的能力，以往當我們面對大量的資料時，一直依賴著抽樣法，以為這是一種理所當然的方法，並且視為是一種限制，但是對於發展出「巨量資料分析方法」的現在而言，已經可以清楚觀察到所有的資料，並且使用這些資料。因此在網路蓬勃發展的現代，當網路民意漸漸被各界所重視，巨量資料分析方法能夠針對網路上的言論進行即時的輿情蒐集、儲存、讀取及分析，而巨量資料分析方法中，網路輿情分析正是其中一種，亦是本篇分析的主軸。且文字探勘相關實務上的應用亦趨多元，許多企業亦透過相關技

[1] CATI (Computer Assisted Telephone Interview)：電腦電話訪問輔助系統係結合電腦、電話設備及通訊科技於一體的電話訪問系統，以自動化的電腦輔助設備進行各種科學性抽樣、電話號碼分配、題目管理、監聽錄音、及資料處理和訪員管理等各項工作。CATI 最大的功能特色乃是擁有臺閩地區住宅電話化電腦資料庫，各鄉鎮市皆建置獨立抽樣資料庫，以確保抽樣的準確性。此外，CATI 系統更可透過電腦進行隨機抽樣，有效地降低抽樣誤差。

術挖掘商業價值性，如 IBM 收購深度學習技術業者 AlchemyAPI，將其整合進 Watson 核心平台；微軟收購以色列文本分析公司 Equivio，開發演算法進行歸類分析；AC 尼爾森(Nielsen)將應用社群大數據強化甚至逐步取代既有收視率調查方式，並稱之為社群調查(Social Rating)、寶僑家品(P&G)為關心消費者需求，亦將採用社群大數據的新方法來傾聽消費者的聲音等，由此可見文字探勘技術應用面之廣泛。

於是，本篇分析之目的在於利用內政統計名詞等相關資料建立內政領域詞庫，搭配網路上各種正負向詞庫及中文斷詞系統，針對國內 1 萬 5,000 多個網站進行文本資料撈取，將其文本去雜訊再進行分類，計算其情緒分數。另採用支持向量機法(SVM)建構文本對內政部的影響度評估模型，排除對內政部沒有影響的文章後，透過網路輿情分析及文字探勘評估政府的網路輿情滿意度，以 106 年輿情資料為分析主軸，搭配原本已辦理的 106 年內政部施政滿意度調查，以地政及營建業務為例，建立一套可行的綜合滿意度分析方法，從而分析兩項業務的施政滿意程度及改善方向，提供決策者更具體可信的參考依據。

第三章 相關文獻介紹

第一節 斷詞系統

1. 結合長詞優先與序列標記之中文斷詞研究（中央大學資訊工程學系林千翔、張嘉惠、陳貞伶）

近年來的斷詞系統傾向於機器學習式演算法來解決中文斷詞的問題，研究者利用訓練資料中已斷詞的文件，建立一個辭典，再利用長詞優先比對提供正向及反向標記資訊，讓學習模組得以學習最佳參數；實際斷詞時，將未斷詞之文章，同樣利用長詞優先比對，產生與訓練資料相同的測試資料，藉由以訓練好的模型，標記文件並得到斷詞結果。

(1) 以機器學習式演算法來解決中文斷詞的問題時，最常使用的方法就是轉換成字元分類，將每個字元都給予對應的類別，透過字元類別來分類，以出現在中文詞當中的特定位置來決定其類別，可以分為位於詞的開始（B）、詞的中間（I）、詞的結尾（E）以及由單一字元組成的詞（S）等四種類別，因此也稱為「BIES 分類與序列標記問題」。

(2) 長詞優先法是最簡單也最廣泛使用的辭典比對式斷詞法，是由句子一端開始，試著比對出在辭典中最長的詞，當作斷詞結果，接著剩下的部分繼續做，直到句子另一端結束為止。如果使用的辭典夠大，長詞優先法斷詞可達90%以上的準確率。長詞優先法依照比對方向的不同又可分為「正向長詞優先法」（FMM）及「反向長詞優先法」（BMM），正向長詞優先法由句子開頭的第一個字元開始，由左而右逐一比對出最長的詞，直到句尾而結束，反向長詞優先法則反之。

2. 以遺傳演算法為基礎的中文斷詞研究（中央大學資訊管理系陳稼興、謝佳倫，真理大學資訊管理系許芳誠）

遺傳演算法(GA)是 John H. Holland 教授受達爾文天擇說:「物競天擇，適者生存」的啟發而發展的演算法則(Adaptation in Natural and Artificial System)。遺傳演算法採用一組特別的字串模擬各種生物的染色體,並計算所有染色體對環境的適應度,在每個世代之間讓各個染色體以隨機的方式進行交配與突變來產生下一代，再根據該染色體的適應度選擇是否讓其生存。這個演化交替的動作會一直持續到達成最終目標為止。該研究運作程序基本上可分為三個步驟，分別概述如下：

(1) 斷句：首先取得一些文章，除去所有的標點符號、阿拉伯數字,只留下純中文字。接著在連續兩個刪除點之間,切出一串串短句。將文章切成純中文的短句，主要目的是為下階段製作詞庫作準備。

(2) 製作詞庫：將步驟(1)斷句產生的短句，分別以 2 字詞、3 字詞、4 字詞、5 字詞等，加上累計每個 N 字詞(N＞1)出現的次數製作出詞庫。即詞庫中每個紀錄中的資料，包含了所有詞的內碼、在文章中出現的累積次數及下一個儲存格的位置。

(3) 用 GA 斷詞：假設欲接受斷詞的中文短句有 n 個中文字，對應於短句的染色體長度為 n-1,並讓這些基因分別對應 n 個字的 n-1 個間隔。在設計上，當基因值為 1，表示此間隔是斷詞處；當基因值為 0，表示不應打斷。藉由 GA 的三個基本操作：複製、交配、突變，以及適應函數的導引，適應度高的染色體被保留，即較正確的斷詞解得以保留；而適應度低的染色體被淘汰。如此反覆演化,GA 將可找到最佳的斷詞方式。

3. 斷詞系統對於 Queried keywords 的影響（亞洲大學資訊多媒體應用系陳宜惠，中興大學資訊管理學系呂瑞麟、黃政傑）

中文斷詞系統為一專精之研究領域，不僅需先經過長期及系統性的蒐集文件才能累積足夠的詞庫或語料庫以進行文件的分析及比對，對於歧義性及未知詞的比對更需利用不同的演算法，如長詞優先、法則式、統計方法等，才能提高斷詞的正確率，因此研究者採用中央研究院資訊科學小組所開發的中文斷詞系統(CKIP)以及採用 MMSEG4j 進行斷詞，以下分別介紹：

(1) 中研院所開發之斷詞系統(CKIP)

中研院於 1986 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。中文斷詞系統為其研究之一，其特色包含一個約 10 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。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。

(2) MMSEG4j 斷詞系統

MMSEG4j 乃為基於詞典之斷詞方式。演算法主要區分為簡單與複雜兩種方式進行解析，此兩種方式都是使用最大匹配演算法進行處理，其簡單的方式準確率達 95%，而複雜方式準確率達 98%，由於詞庫可以自行擴充，因此可以根據新的詞(未知詞)自行添加，並依此來進行斷詞，規則如下：

- 規則一：簡單最大匹配係以找出資料庫中最長的詞彙為原則。複雜最大匹配則為當分割詞彙時，若有歧義的分詞，再往前分析兩個詞彙，以分析三個詞彙為最長的長度為原則，並且取第一個詞彙為最終選擇。
- 規則二：最大平均單詞長度以最大平均單詞長度從 chunk(語

塊)中取得第一個單詞。

- 規則三：單詞長度的最小方差取 chunk 中擁有單詞長度最小方差的作為單詞。
- 規則四：單字單詞的語素自由度的最大和選取 chunk 中擁有最大頻率的第一個詞。

第二節 情緒分類方法

1. 基於社群網路情緒分析之民意預測研究（臺北科技大學資訊工程學系陳冠廷）

研究指出，人們會對於負向回應比較積極，而在正向回應中容易有負向的內容出現，因此須利用情緒分析的方法進行內容的偵測與判斷，並將評價進行調整，得出一評論分類：

$$category(d)\begin{cases} 1, User(d) > 0 \; and \; Sentiment(d) > 0 \\ 0, User(d) = 0 \\ -1, else \end{cases}$$

d為文字資料內容，$User(d)$為使用者評價，$Sentiment(d)$為內容情緒分析評價，$category(d)$為評論d的分類，如表 1。

表 1、評論分類表

$Sentiment(d)$ / $User(d)$	正面評價 (Sentiment(d)＞0)	負面評價 (Sentiment(d)＜0)
正面評價(User(d)＞0)	正向評論 (1)	負向評論 (-1)
中立評價(User(d)＝0)	中立評論 (0)	中立評論 (0)
負面評價(User(d)＜0)	負向評論 (-1)	負向評論 (-1)

針對文章內容(Post)及使用者(User)分別擬定意向指標(Score)及民意指標(Poll)，並進行整合分析：

(1) 各面向(文章面向、使用者面向)民意分數：

$$PollScore(x) = Poll(x) \times weight(x)$$

x為面向，包括文章面向(T_i)及使用者面向(P_i)，$Poll(x)$為各面向民意指標，$weight(x)$為各面向權重：

文章面向權重：$weight(T_i) = \frac{|T_i|}{\sum_{j=1}^{k}|T_j|}$

使用者面向權重：$weight(P_i) = \frac{|P_i|}{\sum_{j=1}^{k}|P_j|}$

$PollScore(x)$為各面向整合後之民意分數，分數越高代表民意越正向，分數越小則越負面。

(2) 整合文章面向及使用者面向之民意分數：

$$FinalPoll(Q_i) = \alpha PollScore(P_i) + (1-\alpha)PollScore(T_i)$$

Q_i為指定概念關鍵字，P_i為指定概念使用者面向，T_i為指定概念文章面向，α為使用者面向權重，$(1-\alpha)$為文章面向權重，$FinalPoll(Q_i)$為概念(Q_i)之民意預測指標，指標數值越高，代表特定議題支持度、滿意度越高，反之則越低。

2. 以 Google App 評論為字詞權重調整之情緒分析系統（靜宜大學資訊管理學系林彩雯）

意見是由目標個體、個體屬性、情緒表示、意見持有者及發表時間所組合而成，研究者蒐集合併知網(HowNet)與臺灣大學意見詞庫(NTUSD)所建立的意見詞庫、程度詞庫及自行編撰的否定詞庫做交互比對，篩選出一個有效詞集，詞集中包含能反映正反情緒的意見詞(Term；T)、能加強情緒強度的程度詞(Degree；D)以及能翻轉情緒指向的否定詞(Negation；N)，再套用至本論文提出的 TDN 演算法，最後得出整體評論的分數。

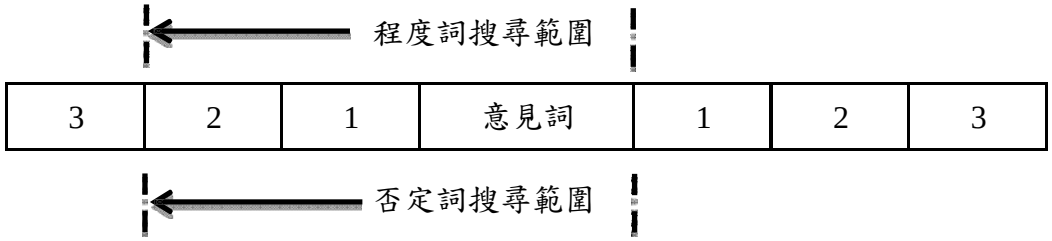
情緒配分由原先意見詞之分數，配合意見詞、程度詞、否定詞三種詞的距離給予不同的計算方式，並只針對意見詞的前兩個位置進行程度詞與否定詞的搜尋，利用知網的程度級別字詞建立程度詞庫，依照強度大小分成超、最、很、較、稍、欠等

6 個級別，程度詞權重定義依強度強弱定義權重(詳表 2)。

表 2、程度詞庫權重

程度	超	最	很	較	稍	欠
權重	5	4.5	3.5	2.5	1.8	1.5

圖 1、程度詞與否定詞搜尋範圍



依圖 1 所示，位置的排列組合有下列 5 種定義：

定義 1：意見詞前 1 位置為否定詞，前 2 位置為程度詞。

定義 2：意見詞前 1 位置為否定詞，前 2 位置非程度詞。

定義 3：意見詞前 1 位置為程度詞，前 2 位置為否定詞。

定義 4：意見詞前 1 位置為程度詞，前 2 位置非否定詞。

定義 5：意見詞前 1 及前 2 位置皆非程度詞或非否定詞。

以意見詞「喜歡」、程度詞「很」、否定詞「不」為例，其權重值分別為 1、3.5、-1，並將三種不同詞別的位置排列組合套用到五種定義計算情緒值，如表 3 範例所示。

表 3、詞彙位置排列組合範例

定義	情緒值計算公式	範例	情緒值
1	程度詞權重* (−1) *意見詞權重	很不喜歡	3.5*(−1)*1= −3.5
2	否定詞權重*意見詞權重	不喜歡	(−1)* 1= −1
3	程度詞權重*意見詞權重/2	不很喜歡	3.5*1／2= 1.75
4	程度詞權重*意見詞權重	很喜歡	3.5*1= 3.5
5	意見詞權重	喜歡	1

3. 使用情緒分析於圖書館使用者滿意度評估之研究（中興大學圖書資訊學系張育蓉）

研究者以人工方式將語料分成館員、館藏、服務、設備，以及空間與環境 5 大類，情緒詞以（S，正+/負−）標記；程度詞以（D，強 3/中 2/弱 1）標記；否定詞以（N）標記；情緒搭配詞以(C,情緒詞)標記。整個句子也標記帶有的情緒與強度，以（S+/−，D3/2/1）表示之，單一句子中可能包含多個程度詞，整句情緒之程度取決於中庸、保守及多數規則。最後，將 4 種詞彙匯總建立情緒詞辭典、程度詞辭典、否定詞辭典與情緒搭配詞辭典。情緒分析的方法有兩種：

(1) 導入程度詞、否定詞與情緒詞權重之情緒分析

測試語料分類後，計算 5 大類句子中情緒詞分數，正向情緒詞為 1 分，負向情緒詞為 -1 分，此外利用程度詞和否定詞給予加權計算句子情緒分數，弱程度詞「稍微」乘以 1.5、中程度詞「很」乘以 2，強程度詞「超級」乘以 3，否定詞「不」則乘上 -1，接著將句中各情緒詞分數加總，即可得到句子情緒分數。之後，將各類中句子情緒分數加總且平均，此為類別情緒分數，類別情緒分數加總代表圖書館的總評分。

如：圖書館資料種類（C，多）沒有（N）很（D+）多（S，+）。

→情緒標記為（S−，N，D2），句子情緒分數為 1x(-1)x2= -2。

(2) 情緒極性與程度類別之情緒分析

測試語料分類後，分別計算 5 大類句子中情緒詞分數。以情緒極性與程度類別之情緒分數，將句子中情緒詞分數加總並求平均值，所得分數為句子的情緒分數，再將各類別中所有句子的情緒分數加總後求平均，得出每類別情緒分數，各類情緒分數相加後所得即為圖書館的總評分。

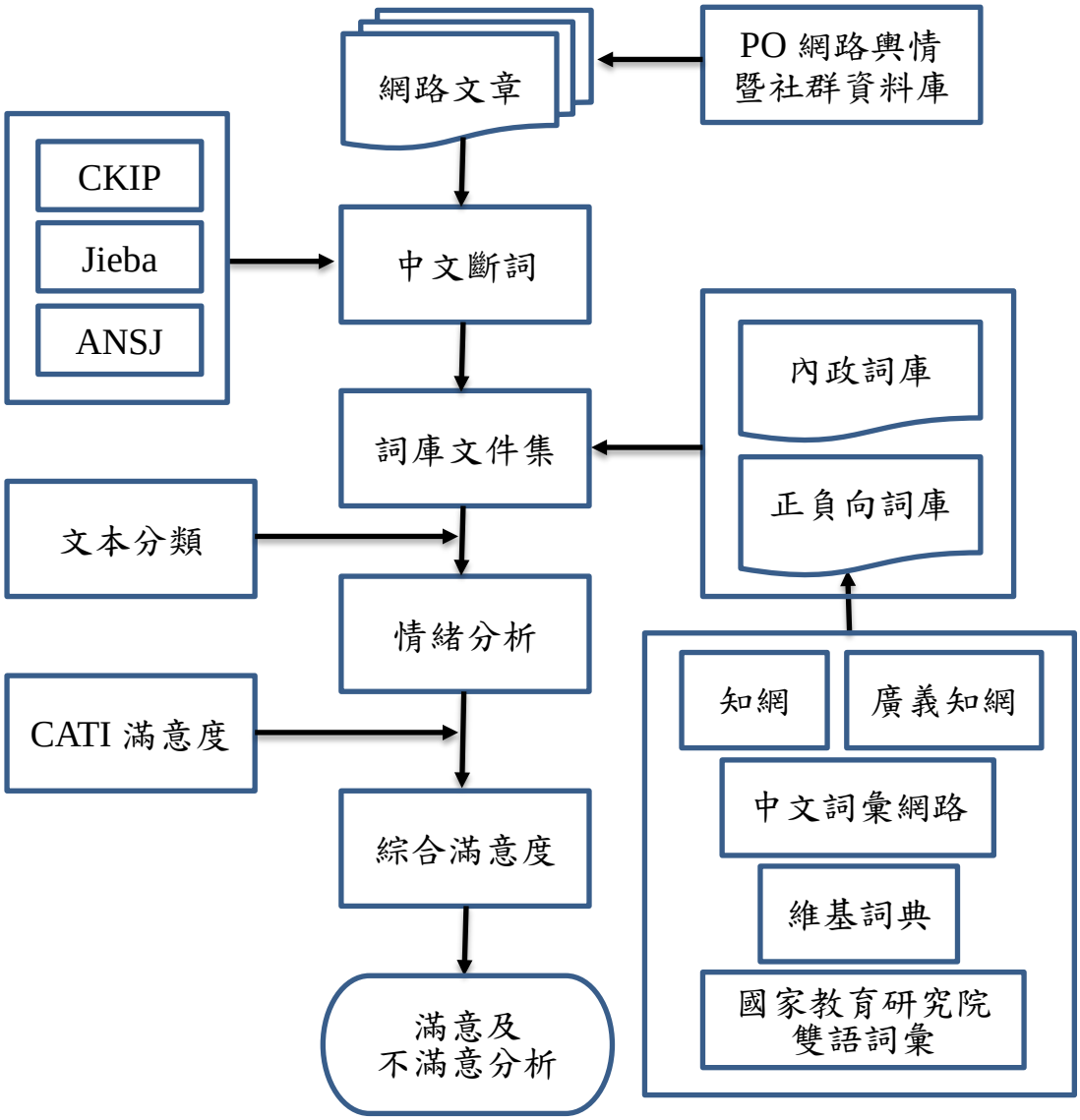
第四章 研究方法概述

文字探勘（Text Mining），又稱文本挖掘、文字採礦或文字知識發現等，係指從非結構化的文字資料中，有效率地萃取出有用的重要資訊或知識。文字探勘與資料探勘（Data Mining）間的差異在於它的原始輸入資料是沒有特定結構的文字資料，這些文字內容都是以人類的自然語言所寫成，所以無法直接套用資料探勘的演算法來計算出結果。且它是一個剛起步的學科領域，是透過資訊擷取、資料探勘、機器學習、統計學、電腦語言學來達成。生活中，除了具固定的欄位、格式，以利程式進行後續取用及分析之結構性資料外，也有相當大量的文字資料，如新聞報導、 社群貼文、網路回文、部落格文章、專利文件、醫學資料等，在這些自然語言文字型的資料中，同樣蘊藏可觀、極具潛力的「礦產」，即有價值的資訊，等待我們用資訊技術去開採。這就是文字探勘技術及應用所希望達成的目標。

文字探勘技術結合文字分析技術與資料探勘技術，目的是希望能夠透過電腦的運算處理能力，過濾並轉化大量文字內容，讓人能夠更有效率的運用。目前文字探勘技術雖然才剛起步，但在現今網路盛行及普及的情況下，相關技術的理論與實務已逐漸成形，除應用於選舉民調、輿情觀察外，其他如公眾行為預測、醫療疾病分類、犯罪行為分析、店家口碑評論、股市漲跌趨勢、專利文件檢索、決策輔助領域等等，都是文字探勘技術運用的成功案例，顯示文字探勘的應用價值及未來潛力。

本分析報告所提之研究方法如圖 2，首先透過浚鴻數據開發股份有限公司之「PO 網路輿情暨社群資料庫[2]」，在網路上定時搜尋相關文章、意見及評論，經過彙整、去雜訊、去重複，並經中文斷詞系統斷詞後，結合網路上既有正負向詞庫，以及內政部地政、營建業務之相關詞庫，構成分析所需之詞庫文件集，藉由網路上相關輿情資料之情緒分數計算，配合電話訪問系統 CATI 調查結果，得出綜合滿意度，並深入探討滿意及不滿意原因，作為施政決策之參考依據。

圖 2、研究架構圖



[2] PO 網路輿情暨社群資料庫網址：http://www.dataa.com.tw/applications/po

第一節 中文斷詞系統簡介

世界上任何語言都是以「字」為基本單位，由字到詞，由詞再到句，再由句建構出文章，但中文語意處理的基本單位為「詞」而非「字」，且在語言學中，詞是能獨立運用並含有語義內容的，即具有表面含義或實際含義的最小單位。中文自然語言處理，例如：文件檢索、中文輸入、字體辨識、語音辨識、機器翻譯、語言分析等，亦都需先對中文句子進行「斷詞」，才能進行下一步的處理，因此如何將正確的詞切分出來，就成為自然語言處理的最基礎工作。以下就針對幾種常用的分詞系統做簡略概述：

1. 中研院開發之中文斷詞系統(CKIP)[3]

中研院於民國 75 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（詳圖 3）。中文斷詞系統為其研究之一，其特點如下：

(1) 包含一個約 10 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。除了基本詞彙庫外，使用者可依需要附加領域專屬詞庫。詞類標記為選擇性功能，可附加文本中切分詞的詞類解決詞類歧義並猜測新詞之詞類。分詞系統採用之詞典俱可擴充性，使用者可依據不同領域文件，補充以領域詞典做為分詞之用。

(2) 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。

(3) 由中研院研發的一款斷詞器，不過並未對外公布技術細節。

[3] 中研院中文斷詞系統(CKIP)網址：http://ckipsvr.iis.sinica.edu.tw/

圖 3、中研院中文斷詞系統(CKIP)

2. Jieba

Jieba 中文斷詞程式是由中國百度的程式工程師所開發的 Python 中文分詞器，為簡體中文的版本，不過因為它是開放原始碼，因此目前已經可以支援繁體中文，也因為它有開放原始碼，預期未來的功能將變得更加完善。Jieba 特點如下：

(1) 使用 Trie 生成句子中所有可能成詞的情況，然後使用動態規劃依詞頻來找出最大機率的路徑；在辨識新詞方面則使用隱式馬可夫模型（HMM）及 Viterbi 算法來辨識。

(2) Jieba 自帶一個 2 萬多條詞的詞典，名叫 dict.txt，包含了詞條出現的次數和詞性，具有查找速度快速的優勢。

3. ANSJ

ANSJ 中文斷詞程式為 JAVA 的中文分詞器，是基於中國科學院的 ictclas 中文分詞演算法而開發出來的，分詞速度可達每秒約 200 萬字左右，準確率能達到 96%以上。目前可完成中文分詞、中文姓名識別、用戶自定義詞典、關鍵字提取、自動摘要以及關鍵字標記等功能。可以應用到自然語言處理等方面，適用於對分詞效果要求高的各種項目。（詳圖 4）

圖 4、ANSJ 下載頁面

4. 小結：

雖然目前市面上有不少的中文斷詞系統，功能、效能及方法各有所長，但在處理不同領域的文件時，相關的特殊詞彙或專有名詞常常造成分詞系統會因為參考詞彙的不足而產生切分上的誤判，故有效解決的方法是補充相關領域詞典，加強詞彙的搜集，並精進詞庫的精確性。因此詞庫彙整或關鍵詞的自動抽取將成為中文斷詞的先期準備步驟，也是文字探勘技術最重要的基礎工程。

第二節 詞庫文件集

1. 正負向詞庫：

(1) 知網(HowNet)[4]

中國科學院計算機語言信息中心語言於 1988 年建立了一個漢語語意辭典，稱為知網(HowNet)，包含了動詞，名詞、屬性、副詞、並列詞、連接詞、助詞、單位詞等詞性，以及一個約 6 萬個詞彙的辭典。不過，它包含的字詞庫為簡體中文版，且檔案的形式化和規範化程度都不高，因此於 2003 年與我國

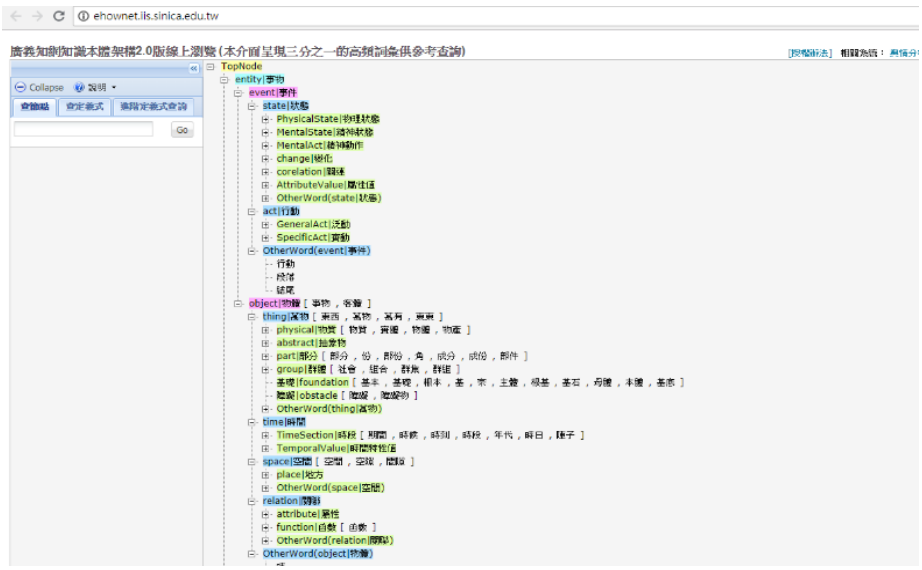
[4] 知網(HowNet)網址：http://www.keenage.com/html/c_index.html

中央研究院資訊所詞庫小組合作，將詞庫小組詞典的詞條與知網連結，架構出廣義知網(E-HowNet)。

(2) 廣義知網(E-HowNet)[5]

中央研究院資訊所詞庫小組於 2003 年起開始建構繁廣義體知網(E-HowNet)，基於知網的本體架構與語義機制，並加入繁體中文常用的詞彙並修改部分定義方式，使得每個詞彙的語義定義更符合語義表徵，也讓功能詞和實詞在定義上有一致的區分，同時符合基本語義的分類方式，包含事物、動作、屬性、屬性值等 4 類，盼能在語法和語義的整體考量後做到適切的語義結合。繼中文詞庫內的 8 萬目詞語義定義初步完成後，詞庫小組更開發詞彙自動分類器，依據詞彙的語義定義式將語義自動分類於廣義知網知識本體架構下，讓詞彙有系統地依據語義來呈現。此套系統於 2011 年上線，使用者能夠簡易地瀏覽詞彙語義定義，並提供檢索功能方便使用者查詢，惟需備妥申請文件，且為付費系統，在使用上的進入門檻較高。(詳圖 5)

圖 5、廣義知網(E-HowNet)



[5] 廣義知網(E-HowNet)網址：http://ehownet.iis.sinica.edu.tw/

(3) 中文詞彙網路(Chinese WordNet)[6]

中文詞彙網路是國科會計畫的工作成果，目的是在提供完整的中文詞義區分與詞彙語意關係知識庫。詞義的區分與表達，必須建立在完善的詞彙語意學理論與知識本體架構基礎上。在實際應用上，這個資料庫可望成為中文語言處理與知識工程不可或缺的基底架構。自 2003 年起已累積了十餘年的研究成果，對詞義區分定義，與詞義知識表達方式，漸次做了修正，並在 2006 年於中央研究院語言學研究所正式啟用，目前計畫網站轉由國立臺灣大學語言學研究所維護。(詳圖 6)

圖 6、中文詞彙網路(Chinese WordNet)

(4) 維基詞典[7]

維基詞典是維基百科的姊妹工程，目標是建立一個基於所有語言的自由詞典，並於 2002 年啟動，係由志願者共同編纂的多語言詞典計劃，旨在囊括各種語言詞彙的語源、讀音和解釋，其中文版始於 2004 年 5 月，目前已具備 83 萬餘個中文詞條。維基詞典既是詞典，也是社群，任何人都可以編輯，亦可至維基資料庫中下載中文語料庫，以擴充詞庫內容，目標是提供每一個人都可以自由使用。(詳圖 7)

[6] 中文詞彙網路(Chinese WordNet)網址：http://lope.linguistics.ntu.edu.tw/cwn/
[7] 維基詞典網址：https://zh.wiktionary.org/zh-hant/

圖 7、維基詞典

(5) 國家教育研究院雙語詞彙[8]

國家教育研究院係我國最高教育學術研究機構，其研究以教育政策、課程、教學、測驗與評量、教科書、圖書編譯等議題為主要範圍，偏重應用性、發展性之研究。「雙語詞彙、學術名詞暨辭書資訊網」原為三個獨立資訊網，於 101 年 12 月完成資料庫整合，目前已建置有 70 多領域 120 餘萬則學術名詞，9 部辭書 6 萬餘則辭條及雙語詞彙 10 萬餘則詞彙，總計超過 140 萬筆資料的專業資訊查詢網站。（詳圖 8）

圖 8、國家教育研究院雙語詞彙

[8] 國家教育研究院雙語詞彙網址：http://terms.naer.edu.tw/

2. 內政相關詞庫

在詞庫的蒐集方面，除了透過中文斷詞系統所切分出的字詞，以及網路上既有之正負向詞庫外，尚須針對內政部地政、營建業務，以人工方式蒐集 106 年度相關詞彙，共同構成分析所需之詞庫文件集，包括內政統計名詞、內政概要、施政滿意度調查、106 月 7 月起新政策措施、第 9 屆立法委員名單及 105 年 1 月至 106 年 7 月與本部相關之輿情文本(計 5,479 篇)，以及 101 年至 106 年 8 月「1996 客服紀錄」(計 6 萬 7,033 則)等，經過斷詞之後提取高詞頻字詞納入詞庫。經過彙整後，地政業務相關詞彙如地政司、仲介經紀商、土地代銷、地政士、不動產估價師、不動產交易、土地所有權登記、建物所有權登記、公告土地現值、公告地價、地價指數、土地增值稅、地價稅、不動產實價登錄等；營建業務相關詞彙如營建署、違章建築、房地合一稅、住宅價格指數、房價指數、社會住宅、住宅補貼、都市更新、國家公園、污水下水道、包租代管、綠建築等。

第三節 情緒分數計算方式

情緒分數的計算是將「意見詞庫」內的意見詞所設定的分數，搭配「程度詞庫」以及「否定詞庫」，將這三種詞庫，設定字串距離，依照排列組合來計算。目的是要從大量的文字中，辨識出正面與負面等情緒，好讓這些結果產生出商業價值，但事實上還有許多問題仍待解決，如特殊的中、日文文法、反諷句、網路流行語等，因此如何做的更加精準，尚需投入大量的人力、資源與成本，更重要的是要「跨領域方面」，因為在大數據時代講究的是跨領域的整合，單就文字上解析還不夠，還需要更多元的資料，才能做得更好。目前情緒分類技術可以分成以下兩種：一種為機器學習法，另一種為建立情緒字詞辭典法。

1. 機器學習法

機器學習法是運用機器學習演算法及語言學特徵。機器學習法需要一個已經被標記好的訓練集來建立模型並學習文件中不同的特徵來區分正面、負面、中性訊息，支持向量機（Support Vector Machine；SVM）便是機器學習中非常知名的算法，它廣泛地應用於分類、迴歸、異常點檢測等問題中。其技術係以統計學為基礎以及二分法的分類器，對於一多維空間中之資料找出一個超平面可將資料分為兩類，而超平面與其最接近的支持向量之間的距離稱為邊際，支持向量機目標為於空間中之資料找出具有最大邊際之超平面以作為分類邊界。SVM 的優點為可解決特徵樣本小及高維度、高稀疏化的特徵，而對非線性的問題尚未有真正通用之解決方案為其缺點。

2. 建立情緒字詞辭典法

建立情緒字詞辭典法是利用統計或語意的方式判斷情緒極性。又分成以字典法(Dictionary-Based approach)及語料庫為基礎(Corpus-Based Approach)的方式來收集和編輯情緒字詞。

(1) 字典法是將一部份已知的情緒方向的字詞先收集起來作為種子字詞，再藉由辭典如 E-HowNet 或 WordNet 等，找出種子字詞的同義字或反義字，若與種子字詞為同義詞則擁有相同極性，反之則為相反極性，利用此方式以達到擴充詞庫效果。字典法是情感分析議題中較常使用的方法，許多學者會利用已建立好的字典集進行情感分類、情感詞擷取或情感偵測，良好的字典集可以提高進行情感分析時的正確。

(2) 語料庫為基礎法則透過演算法蒐集大量文件集，由大量的語料庫中自動學習詞彙、語句、文章間與意見傾向的關係，利用統計方法觀察出一些規律與法則，以挖掘文件裡的概念詞庫並判斷極性(正負向情緒)。但此種方法較依賴種子詞彙的個數及質量，且需要大量的語料來做訓練學習，可能造成一些詞彙具有多義性而造成詞彙的極性判斷錯誤。

第四節 輿情滿意度計算方式(結合 CATI)

在調查民眾對政府的施政滿意度時，市話電訪(CATI)雖存在誤差，但如果僅利用文字探勘來推論，本質上存在嚴重的抽樣母體代表性的問題。首先，網路上的輿情資料無法呈現不上網者的意見；其次，即使有上網，不發言者多於發言者也是網路世界的另一個特色，顯示網路輿情資料反應的是具有高度公民參與意願者的意見；最後，對於事物具有負面態度的民眾在網路上發言的意願高於具有正面態度的民眾，這也造成蒐集到的意見偏向負面意見，即所謂的「沉默螺旋理論[9]」。因此需將兩者互相結合，截長補短，才能較完整地陳述民眾滿意度，並期待能高度貼近民意，以下介紹網路輿情滿意度與市話民調滿意度結合之計算方法。

網路上民眾的意見與情緒是在閱讀某個媒體文本後所表示認同或不認同的態度，即統計網路上所有相關文本所得到的正負面文本占比，是一個民眾態度的條件機率值。媒體所發表的文本按其報導情緒可區分為正面、負面或中立報導，民眾閱讀完報導後所產生的態度可分為正面、負面態度或無態度。表 4 呈現出在不同報導情緒下，民眾產生相對應情緒的各種條件機率值。例如，當媒體為正面報導的情況下，民眾也表達了正面態度，在此情況下我們測得的正面文本比例，是在媒體正面報導的情況下民眾滿意的條件機率值 $P(民^{+}\mid 媒^{+})$。當媒體為負面報導的情況下，民眾表達了對主文的負面態度，所測得的負面文本比例，是在媒體負面報導的情況下民眾滿意的條件機率值 $P(民^{+}\mid 媒^{-})$。透過全機率法則[10]求得民眾正面態度的機率值，即網路輿情滿意度。

[9] 「沉默螺旋理論」是由德國學者諾爾紐曼（Noelle-Neumann）於 1970 年提出，主要概念為當人們覺得自己的觀點屬大眾中的少數時，他們將不願意表達自己的看法；而如果他們覺得自己看法與多數人一致，則會勇敢的說出來。而且媒體通常會關注多數的觀點，輕視少數的觀點。於是少數的聲音越來越小，多數的聲音越來越大，形成一種螺旋式上升的模式。

[10] 全機率法則係指全機率公式將對一複雜事件 A 的機率求解問題轉化為了在不同情況或不同原因 B_n 下發生的簡單事件機率之求和問題。全機率法則：$P(A)=\sum P(B_n)\times P(A\mid B_n)$。

表 4、網路上民眾態度類型彙整表

	媒體正面報導	媒體中立報導	媒體負面報導
民眾正面態度	$P(民^{+}\mid 媒^{+})$	$P(民^{+}\mid 媒^{0})$	$P(民^{-}\mid 媒^{-})$
民眾無態度	$P(民^{0}\mid 媒^{+})$	$P(民^{0}\mid 媒^{0})$	$P(民^{0}\mid 媒^{-})$
民眾負面態度	$P(民^{-}\mid 媒^{+})$	$P(民^{-}\mid 媒^{0})$	$P(民^{+}\mid 媒^{-})$

民眾的網路輿情滿意度 $P(民^{+})$計算公式如下：

$$\mathbf{P(民^{+}) = P(媒^{+})\times P(民^{+}\mid 媒^{+}) + P(媒^{-})\times P(民^{+}\mid 媒^{-}) + P(媒^{0})\times P(民^{+}\mid 媒^{0})} \quad \cdots\cdots ①$$

由於網路輿情資料偏向有上網者、會表態者、負面態度者的意見，易忽略其他特性民眾的意見，導致公式 ①之 $P(民^{+}\mid 媒^{+})$被低估。為解決此問題，本分析以電話訪問法(CATI)所獲得的滿意度取代公式 ①中的 $P(民^{+}\mid 媒^{+})$，以解決 $P(民^{+}\mid 媒^{+})$低估的問題。

電話訪問的施政滿意度調查問卷大多為正面表述，且可以同時涵蓋未上網者、未表態者及正面態度者的意見，故以電話訪問所得到的滿意度 P(CATI)取代 $P(民^{+}\mid 媒^{+})$，可以達到修正網路輿情資料 $P(民^{+}\mid 媒^{+})$低估的問題。但另一方面，由於電話訪問在態度良好的訪員引導下很容易讓受訪者對評估的服務項目產生正向態度，若受訪者原持負向態度則容易隱蔽自己真實的態度，造成民眾真實滿意度高估的情形，本研究所提出結合網路輿情資料及電話訪問資料的分析方法可適度修正電話訪問法的訪員引導效應，亦可修正電話訪問法中年輕人受訪比例偏低所造成的偏差，調整後之綜合滿意度如公式 ②。

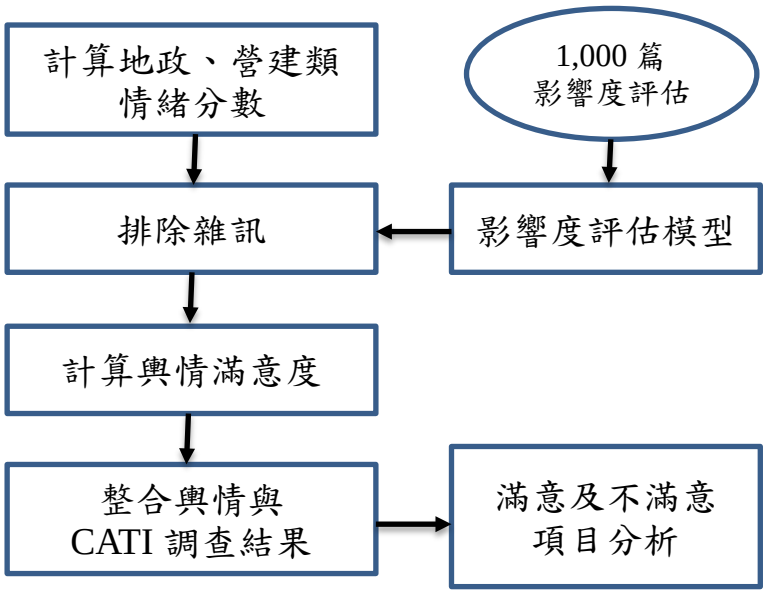
調整後的綜合滿意度 $P^{*}(民^{+})$公式如下：

$$\mathbf{P^{*}(民^{+}) = P(媒^{+})\times P(CATI) + P(媒^{-})\times P(民^{+}\mid 媒^{-}) + P(媒^{0})\times P(民^{+}\mid 媒^{0})} \quad \cdots\cdots ②$$

第五章 實證分析

本研究將文本按業務別分類後初步篩選出營建、地政類相關文本，分別獲得 78 萬 9,949 篇及 21 萬 584 篇文本，計算出情緒分數，並採用支持向量機(SVM)建立「內政部影響度評估模型」，以剔除與本部施政相關性較低的文本。研究中將 1,000 篇經過人工標記的本部相關文本切分為 900 份訓練資料以及 100 份測試資料，然後以 900 份訓練資料進行模型訓練，最終獲得內政部影響度的預測模型，再針對 100 筆測試樣本進行預測。接著透過「PO 網路輿情暨社群資料庫」計算輿情滿意度，並結合電話訪問系統 CATI 調查結果，得出調整後綜合滿意度，進而深入探討不滿意原因，作為施政決策之參考依據，實證分析研究流程如圖 9。

圖 9、輿情滿意度與 CATI 調查結合流程圖



第一節 影響度評估模型

由本部同仁針對隨機挑選的 1,000 篇內政相關文本進行對本部的影響度標記，對本部有高度正面影響的文本標記為 5 分，對本部有高度負面影響的文本標記為-5 分，對本部沒有影響的文本標記為 0 分，每篇文本分別由三位同仁獨立評分，作為影響度評估模型建置之訓練資料。本研究採用支持向量機法(SVM)建構

文本對內政部的影響度評估模型，將 1,000 篇經過人工標記的內政部相關文本切分為 900 份訓練資料以及 100 份測試資料，以 900 份訓練資料進行模型訓練，最終獲得內政部影響度的預測模型，透過 100 份測試資料進行模型測試，並比較人工標記的影響度(實際影響度)以及模型預測的結果(預測影響度)，大致可將預測結果區分為下列四種情形(詳表 5)：

TP(真正類)：真實類別是正類，模型預測成正類
FP(假正類)：真實類別是負類，模型預測成正類
FN(假負類)：真實類別是正類，模型預測成負類
TN(真負類)：真實類別是負類，模型預測成負類

表 5、模型預測結果分類

		實際	
		+	−
預測	+	TP	FP
	−	FN	TN

其中，資料轉換邏輯如下：

1. 可容許範圍係指容許正負 1 單位的誤差。例如資料實際為-2，預測為-1 或-3 為可容許範圍。
2. TP：
   (1) 完全預測正確。例如資料實際為 2，預測為 2。
   (2) 資料預測結果於可容許範圍內，但預測值需高於實際值。例如資料實際為 2，預測為 3，或資料實際為-2，預測為-1。
3. FP：當資料被高估。例如資料實際為 1，預測為 3 以上，或資料實際為-3，預測為-1 以上，或資料實際為-1，預測為 1 以上。
4. FN：當資料被低估。例如資料實際為 3，預測為 1 以下，或資料實際為-1，預測為-3 以下，或資料實際為 1，預測為-1 以下。
5. TN：資料預測結果於可容許範圍內，但預測值需低於實際值。例如資料實際為 2，預測為 1，或資料實際為-2，預測為-3。

表 6、內政部影響度 SVM 模型之混淆矩陣

		實際值							
		-3	**-2**	**-1**	**0**	**1**	**2**	**3**	**4**
預測值	**-3**	0	0	0	0	0	0	0	0
	-2	0	0	0	0	0	0	0	0
	-1	0	2	9	2	0	0	0	0
	0	1	1	2	5	3	2	0	0
	1	0	2	1	7	8	0	0	0
	2	0	1	3	1	6	20	5	0
	3	0	0	0	0	1	3	14	0
	4	0	0	0	0	0	0	0	1

TP = 0+0+0+2+9+2+5+7+8+6+20+3+14+0+1 = 77

FP = 0+1+0+0+0+0+1+2+1+0+0+1+3+0+0+1+0+0+1+0+0 = 11

FN = 0+0+0+0+0+0+0+0+0+2+0+0+0+0+0+0+0+0+0+0 = 2

TN = 0+0+2+3+0+5+0 = 10

將預測結果統計如表 7，並計算各類模型精準度指標如下所示：

表 7、100 筆測試樣本的預測結果

		實際	
		+	−
預測	+	77	11
	−	2	10

$$\text{Accuracy(準確度)}：\frac{\text{TP}+\text{TN}}{\text{TP}+\text{FP}+\text{FN}+\text{TN}}=\frac{77+10}{77+11+2+10}=0.87$$

$$\text{Precision(精確度)}：\frac{\text{TP}}{\text{TP}+\text{FP}}=\frac{77}{77+11}=0.875$$

$$\text{Recall(召回率)}：\frac{\text{TP}}{\text{TP}+\text{FN}}=\frac{77}{77+2}=0.9745$$

$$\text{F1}-\text{measure}：\frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}}=\frac{2\times 77}{2\times 77+11+2}=0.9222$$

從各類精準度指標來看，準確度(Accuracy)為 87.00%，精確度(Precision)為 87.50%，召回率(Recall)更高達 97.45%，F1-measure(又稱 F-score)亦達 92.22%，顯示內政部影響度評估模型具有一定的預測能力，故進一步利用此模型評估每篇文本對內政部的影響度，排除影響度為 0 的文本(屬於雜訊)，最終留下營建類文本 74 萬 6,822 篇、地政類文本 12 萬 9,000 篇，作為本研究最終分析資料，並透過浚鴻數據開發股份有限公司之「PO 網路輿情暨社群資料庫」計算輿情滿意度。

第二節 計算情緒分數

計算情緒分數時，會先以自然語言處理(NLP)的方式將每個提及的文字(媒體報導、推文、FB 貼文等)分類，範圍由-10 到 10。分數 10 最正面，0 為中性，而-10 最負面。接著，對應的機器學習模式會針對正面及負面類別來指派強度層級。在 NLP 為基礎的方式中，會使用意見措詞的字典集(包括字詞、片語及表情符號)來進行情緒分類，並根據那些措詞來分析。為了判斷提及的情緒分數，模型會先對輸入文字執行前置處理(正規化及詞性標記等)，其次識別意見措詞及其輸入文字的對應情緒分數，最後使用演算法來獲得輸入文字的彙總情緒分數。針對情緒分析說明以下幾種可能狀況：

1. 傳達正面情緒的詞語

   如提及數個正面情緒的字詞出現，模型會根據那些字詞的彙總來計算情緒分數，但情緒分數不可能大於 10。例如：快樂，感動，安全，自由，勇敢，積極等。

2. 傳達負面情緒的詞語

   如提及數個負面情緒的字詞出現，模型會根據那些字詞的彙總來計算情緒分數，但情緒分數不可能小於-10。例如：恐懼，憤怒，害怕，驚訝，失望，絕望、不安，擔心等。

3. 加強正面情緒的修飾詞語

有些單字是做為其他單字的正面修飾語，模型會在計算期間評估修飾詞數目。若具有多個修飾詞，情緒分數不可能大於 10。例如：很、非常、超級等。

4. 抵消正面情緒的修飾詞語

有些修飾詞會否定單字的正面情緒。例如：不是、沒有等。

5. 加強或抵消正面情緒的修飾詞語

包含一個強化正面情緒的修飾詞，同時又包含一個否定正面情緒的修飾詞。例如: 這項措施很便民，但是太過美好反而令人難以相信。系統判斷「便民、美好」皆為正面情緒，惟「但是」否定了「美好」的正面情緒。最終的情緒分數應該是略微負面。

6. 被上下文抵銷的正面情緒詞語

有些單字會被視為傳達正面情緒，但在模型檢查文字時，判斷該單字與其他單字組成的上下文中，並未傳達正面情緒。例如：如果這項政策能多加宣導，就能達到成效。系統判斷「達到成效」傳達正面情緒，但同時判斷「如果」否定了正面情緒，因此情緒分數應該是負面的。

7. 抵消正面情緒的詞語

有些單字與傳達正面情緒的單字連用時，實際上是否定正面情緒。例如：這政策原本應該是用意良善。系統判斷「用意良善」傳達正面情緒，但同時判斷「原本應該」否定了正面情緒。因此這個情緒分數應該是負面的。

8. 傳達情緒的常用片語

使用包含常用片語及慣用語的字典來協助判斷情緒。例如：酸葡萄、潑冷水等。

9. 圖釋

剖析表情符號來識別其情緒。例如：「:-)」判斷這個表情符號傳達非常正面的情緒。

10. 不規則的意見措詞正規化及拼字錯誤更正

民眾有時會使用不規則的意見措詞來表達情緒。偵測出那些措詞並予以正規化，藉此偵測並更正某些拼字錯誤。例如：這對情侶是「ㄈㄈ尺」，系統將其更正為「CCR」，意指異國戀愛(Cross Cultural Romance)。

11. 名稱實體辨識

識別出文章中提及的具名實體。例如：活動、人、組織等。

第三節 計算輿情滿意度

計算內政部的網路輿情滿意度前，需先針對媒體報導情緒(簡稱媒體情緒) 及網民對媒體報導文本的情緒(簡稱網民情緒)給定正負面態度的操作型定義。

1. 媒體情緒的操作型定義

「PO 網路輿情暨社群資料庫」中所提供的媒體情緒分數為-10~10 的等級數值，其中情緒分數為-6~6 的文本約占所有文本的 68%(大致符合常態分配)。為了解不同情緒切點對滿意度估算結果所造成的影響，訂定 2 種對正向態度評估有不同標準的媒體情緒操作型定義(詳表 8)，有媒 I 類及媒 II 類。

表 8、兩種媒體情緒的操作型定義

態度 / 類型	媒體情緒 正面	媒體情緒 中立	媒體情緒 負面
媒 I 類	≥ 6	−5~5	≤ −6
媒 II 類	＞ 0	0	＜ 0

2. 網民情緒的操作型定義

為了解不同情緒切點對滿意度估算結果所造成的影響，訂定 4 種對正向態度評估有不同標準的網民情緒操作型定義(詳表 9)，從民 I 類到民 IV 類，對正向態度的評估標準越趨寬鬆。

表 9、四種網民情緒的操作型定義

態度 / 類型	網民情緒 正面	網民情緒 中立	網民情緒 負面
民 I 類	≥ 6	$-5\sim5$	≤ -6
民 II 類	> 0	0	< 0
民 III 類	≥ 0		< 0
民 IV 類	≥ -6		< -6

3. 地政類網路輿情滿意度

依據公式 ①計算內政部地政業務的網路輿情滿意度結果如下表 10，彙整後如表 11。若網民正向情緒評估標準為大於等於 6 時(民 I 類)，網路輿情滿意度在媒 I 類及媒 II 類皆低於 1 成；當網民正向情緒評估標準放寬為大於 0 時(民 II 類)，網路輿情滿意度皆提升至近 2 成；當具中立情緒的網民文本(網民情緒分數為 0)也視為網民具正面態度時(民 III 類)，網路輿情滿意度分別為 80.0%及 66.9%；若把網民情緒分數大於等於-6 皆視為網民具正面態度(民 IV 類)，網路輿情滿意度分別為 89.0%及 73.3%。

表 10、不同操作型定義下的地政業務輿情滿意度

媒	民	$P(媒^{+})$	$P(民^{+}\|媒^{+})$	$P(媒^{-})$	$P(民^{+}\|媒^{-})$	$P(媒^{0})$	$P(民^{+}\|媒^{0})$	$P(民^{+})$
I	I	33.87%	7.69%	9.26%	4.44%	56.87%	8.91%	8.1%
I	II	33.87%	16.24%	9.26%	31.11%	56.87%	20.30%	19.9%
I	III	33.87%	86.32%	9.26%	31.11%	56.87%	84.16%	80.0%
I	IV	33.87%	97.44%	9.26%	2.22%	56.87%	98.02%	89.0%
II	I	47.44%	7.47%	25.45%	3.67%	27.11%	7.41%	6.5%
II	II	47.44%	16.67%	25.45%	18.35%	27.11%	18.52%	17.6%
II	III	47.44%	83.91%	25.45%	18.35%	27.11%	82.72%	66.9%
II	IV	47.44%	97.70%	25.45%	1.83%	27.11%	97.53%	73.3%

表 11、地政業務輿情滿意度彙整表

地政類	媒 I (媒體正向情緒分數≧6)	媒 II (媒體正向情緒分數＞0)
民 I(網民正向情緒分數≧6)	8.1%	6.5%
民 II(網民正向情緒分數＞0)	19.9%	17.6%
民 III(網民正向情緒分數≧0)	80.0%	66.9%
民 IV(網民正向情緒分數≧-6)	89.0%	73.3%

4. 營建類網路輿情滿意度

本篇分析依據公式 ①計算內政部營建業務的網路輿情滿意度結果如表 12，彙整後如表 13。若網民正向情緒分數大於等於 6 時(民 I 類)，網路輿情滿意度在媒 I 類及媒 II 類皆僅 1 成左右；當網民正向情緒分數放寬為大於 0 時(民 II 類)，網路輿情滿意度皆提升至 2 成左右；當具中立情緒的網民文本(網民情緒分數為 0)也視為網民具正面態度時(民 III 類)，網路輿情滿意度分別為 74.3%及 64.5%；若把網民情緒分數大於等於-6 皆視為網民具正面態度(民 IV 類)，網路輿情滿意度分別為 87.0%及 71.0%。

表 12、不同操作型定義下的營建業務輿情滿意度

媒	民	$P(媒^{+})$	$P(民^{+}\|媒^{+})$	$P(媒^{-})$	$P(民^{+}\|媒^{-})$	$P(媒^{0})$	$P(民^{+}\|媒^{0})$	$P(民^{+})$
I	I	33.47%	11.78%	9.78%	3.23%	56.75%	10.21%	10.1%
I	II	33.47%	19.94%	9.78%	18.28%	56.75%	20.60%	20.2%
I	III	33.47%	81.87%	9.78%	18.28%	56.75%	79.58%	74.3%
I	IV	33.47%	96.37%	9.78%	2.15%	56.75%	96.13%	87.0%
II	I	47.04%	11.21%	26.36%	3.33%	26.60%	11.42%	9.2%
II	II	47.04%	19.86%	26.36%	19.17%	26.60%	19.75%	19.6%
II	III	47.04%	82.24%	26.36%	19.17%	26.60%	78.09%	64.5%
II	IV	47.04%	96.50%	26.36%	1.67%	26.60%	94.75%	71.0%

表 13、營建業務輿情滿意度彙整表

營建類	媒 I (媒體正向情緒分數≧6)	媒 II (媒體正向情緒分數＞0)
民 I(網民正向情緒分數≧6)	10.1%	9.2%
民 II(網民正向情緒分數＞0)	20.2%	19.6%
民 III(網民正向情緒分數≧0)	74.3%	64.5%
民 IV(網民正向情緒分數≧−6)	87.0%	71.0%

第四節 結合 CATI 調查之綜合滿意度

1. 調整後地政類綜合滿意度

依據公式 ②，以電話訪問 CATI 所得到的滿意度 P(CATI)= 94.19%(滿意+不知道)取代公式 ①之 $P(民^{+}\mid 媒^{+})$，計算內政部地政業務的綜合滿意度如表 14，彙整後如表 15。若網民正向情緒評估標準為大於等於 6 時，調整後綜合滿意度在媒 I 類及媒 II 類分別為 37.4%及 47.6%；當評估標準放寬為大於 0 時，調整後綜合滿意度分別為 46.3%及 54.4%；當把具中立情緒的網民文本(評估標準等於 0)也視為網民具正面態度時，調整後綜合滿意度分別為 82.6%及 71.8%；若把評估標準大於等於-6 皆視為網民具正面態度，調整後綜合滿意度分別為 87.9%及 71.6%。

表 14、調整後不同操作型定義下的地政業務綜合滿意度

媒	民	$P(媒^{+})$	P(CATI)	$P(媒^{-})$	$P(民^{+}\mid 媒^{-})$	$P(媒^{0})$	$P(民^{+}\mid 媒^{0})$	$P^{*}(民^{+})$
I	I	33.87%	94.19%	9.26%	4.44%	56.87%	8.91%	37.4%
I	II	33.87%	94.19%	9.26%	31.11%	56.87%	20.30%	46.3%
I	III	33.87%	94.19%	9.26%	31.11%	56.87%	84.16%	82.6%
I	IV	33.87%	94.19%	9.26%	2.22%	56.87%	98.02%	87.9%
II	I	47.44%	94.19%	25.45%	3.67%	27.11%	7.41%	47.6%
II	II	47.44%	94.19%	25.45%	18.35%	27.11%	18.52%	54.4%
II	III	47.44%	94.19%	25.45%	18.35%	27.11%	82.72%	71.8%
II	IV	47.44%	94.19%	25.45%	1.83%	27.11%	97.53%	71.6%

表 15、調整後的地政業務綜合滿意度彙整表

地政類	媒 I (媒體正向情緒分數≧6)	媒 II (媒體正向情緒分數＞0)
民 I(網民正向情緒分數≧6)	37.4%	47.6%
民 II(網民正向情緒分數＞0)	46.3%	54.4%
民 III(網民正向情緒分數≧0)	82.6%	71.8%
民 IV(網民正向情緒分數≧−6)	87.9%	71.6%

2. 調整後營建類綜合滿意度

依據公式 ②，以電話訪問 CATI 所得到的滿意度 P(CATI)= 67.43%(滿意+不知道)取代公式 ①之 $P(民^{+} \mid 媒^{+})$，計算內政部營建業務的綜合滿意度如表 16，彙整後如表 17。若網民正向情緒評估標準為大於等於 6 時，調整後綜合滿意度在媒 I 類及媒 II 類分別為 28.7%及 35.6%；當評估標準放寬為大於 0 時，調整後綜合滿意度分別為 36.0%及 42.0%；當把具中立情緒的網民文本(評估標準等於 0)也視為網民具正面態度時，調整後綜合滿意度分別為 69.5%及 57.5%；若把評估標準大於等於-6 皆視為網民具正面態度，調整後綜合滿意度分別為 77.3%及 57.4%。

表 16、調整後不同操作型定義下的營建業務綜合滿意度

媒	民	$P(媒^{+})$	P(CATI)	$P(媒^{-})$	$P(民^{+}\mid 媒^{-})$	$P(媒^{0})$	$P(民^{+}\mid 媒^{0})$	$P^{*}(民^{+})$
I	I	33.47%	67.43%	9.78%	3.23%	56.75%	10.21%	28.7%
I	II	33.47%	67.43%	9.78%	18.28%	56.75%	20.60%	36.0%
I	III	33.47%	67.43%	9.78%	18.28%	56.75%	79.58%	69.5%
I	IV	33.47%	67.43%	9.78%	2.15%	56.75%	96.13%	77.3%
II	I	47.04%	67.43%	26.36%	3.33%	26.60%	11.42%	35.6%
II	II	47.04%	67.43%	26.36%	19.17%	26.60%	19.75%	42.0%
II	III	47.04%	67.43%	26.36%	19.17%	26.60%	78.09%	57.5%
II	IV	47.04%	67.43%	26.36%	1.67%	26.60%	94.75%	57.4%

表 17、調整後的營建業務綜合滿意度彙整表

營建類	媒 I (媒體正向情緒分數≧6)	媒 II (媒體正向情緒分數＞0)
民 I(網民正向情緒分數≧6)	28.7%	35.6%
民 II(網民正向情緒分數＞0)	36.0%	42.0%
民 III(網民正向情緒分數≧0)	69.5%	57.5%
民 IV(網民正向情緒分數≧-6)	77.3%	57.4%

3. 調整後綜合滿意度差異

考量到媒體報導的內容將高度影響民意的走向，民眾意見較易受外在輿論影響，因此操作型定義針對媒體採較嚴格定義之媒 I，針對民眾意見採較寬鬆之民 III，因此調整後的地政業務綜合滿意度為 82.6%，營建業務綜合滿意度為 69.5%，與 106 年民眾對內政部施政滿意度 CATI 調查所得到的營建類滿意度 67.4%(滿意+不知道)差異不大，而與地政類滿意度 94.2%(滿意+不知道)卻有著 11.6 個百分點的差異，說明透過文字探勘計算出輿情滿意度，並結合 CATI 調查結果之操作方式，得出調整後輿情滿意度較符合一般民眾對於本部施政滿意度的看法。

第五節 滿意及不滿意原因分析

以本研究建立之內政領域詞庫中與地政、營建施政或政策相關的關鍵詞，分別對其文本進行細項標記，並篩選出與各關鍵詞相關的文本計算其綜合滿意度。標記結果形成一個 617x875,822 的矩陣，呈現出每個地政、營建業務相關的關鍵詞在每篇文本中出現的次數。採用媒 I 及民 III 兩個操作型定義，分別計算各地政、營建業務相關關鍵詞在公式 ②的綜合滿意度，滿意度大於 70%之關鍵字視為民眾對地政或營建業務的滿意項目，滿意度小於 30%之關鍵字視為民眾對營建或地政業務的不滿意項目，藉此明確點出施政優勢及盲點，持續發揮施政績效。

1. 調整後地政類項目輿情滿意及不滿意情況

(1) 地政類滿意項目

分析結果顯示，地政業務相關關鍵字中滿意度最高的項目為「土地增值稅」，滿意度達 92.1%，其次依序為「市地重劃」（89.1%）、「區段徵收」(83.7%)等。（詳表 18）

表 18、對地政業務滿意的項目

項目	調整後滿意度	網民正面態度事件摘錄 (摘列網民情緒大於等於 9 之文本)
土地增值稅	**92.1%**	■ 增加政府財源： ♦ 花蓮公告 106 年土地現值，全縣平均調幅 0.26%(新浪網、Hinet) ■ 討回大戶欠稅： ♦ 長安高爾夫球場巨額欠稅及土地增值稅拍出逾 2 億元，為新竹縣庫注活水(新浪網) ♦ 欠稅又 4 度流標，土地直接抵繳(Yahoo) ■ 退稅： ♦ 高市稅處，土增稅重購退稅清查作業起跑(Yahoo 房地產) ■ 稅改： ♦ 土地稅「精準」實價課稅，財政部：會打折(Yahoo)
市地重劃	**89.1%**	■ 地方建設： ♦ 打通大新營經絡，南市積極推轉運中心、聯絡道路工程(Hinet、Yahoo 房地產) ♦ 高雄第 71 期市地重劃工程動土典禮(民報) ♦ 台中市政府運用都市計畫及土地開發策略，引導城市健全發展(新浪網) ♦ 台大醫院擴建新竹湳雅院區，府院攜手趕進度(PChome 新聞、Yahoo 新聞) ♦ 南港跨區市地重劃，潛力無限(Hinet、PChome 新聞) ♦ 市地重劃，苗市袍澤變綠地(Yahoo) ♦ 促濱海地區發展，台東縣府 11 月發新土地權狀(Hinet、Yahoo) ♦ 中職冠軍賽，中市府重劃區紓解停車需求(中央社) ■ 捷運用地開發： ♦ 司法院對徵收之捷運用地得否用於聯合開發案解釋，建構行政、監察與司法三權新關係(PChome 新聞) ■ 自辦市地重劃： ♦ 自辦市地重劃違憲條文，內政部修法強化公開聽證(蘋果日報) ♦ 鄭文燦：草漯新市鎮若無法自辦，市府將接手公辦(Yahoo 房地產) ■ 公民參與： ♦ 健全公民參與，聽證制 11 月上路(Yahoo)

項目	調整後滿意度	網民正面態度事件摘錄 (摘列網民情緒大於等於 9 之文本)
區段徵收	**83.7%**	■ 資訊透明： ♦ 烏日九德區段徵收，中市府辦說明會廣納民眾建言(Hinet、新浪網) ■ 爭議處理： ♦ [新聞]遲來的正義！內政部：大埔張藥房確定重建(Gossiping) ■ 執行成效： ♦ 中市地政局標售區段徵收區配餘地，標脫率逾 8 成(新浪網) ♦ 湖子內區段徵收土地標售，一、二標住宅及商業區被搶光(新浪網) ■ 地方建設： ♦ 亞洲新灣區開發，打造高雄國際門戶(PChome 新聞網) ♦ 台中水湳經貿路通車，紓解中清路車流(新浪網) ♦ [新聞]高雄果菜市場鄰近里民挺市府：速處理北側(PTT Kaohsiung) ♦ 龜山捷運棕線，鄭文燦：目標 109 年動工(Yahoo) ■ 生態保護： ♦ 保育茄萣濕地，市府重視生態與地方發展(PChome 新聞網)
土地徵收	**70.0%**	

(2) 地政類不滿意項目

地政業務相關關鍵字中滿意度最低的項目為「抵費地」，滿意度僅 6.7%，其次依序為「土地登記」,「地價稅」、「公告地價」、「平均地權」、「地籍」、「規定地價」、「農地重劃」、「實價登錄」等。（詳表 19）

表 19、對地政業務不滿意的項目

項目	調整後滿意度	網民負面態度事件摘錄 (摘列網民情緒小於等於-9 之文本)
抵費地	**6.7%**	■ 弊案： ♦ 長億集團楊天生父子，涉台中黎明自辦市地重劃弊案
土地登記	**21.6%**	■ 個人權益損害： ♦ [爆卦]我爸土地被政府登記給他人(PPT Gossiping) ♦ 活在騙局 60 年，劉明：只想捍衛一輩子的家(PTT PublicIssue) ♦ 【社會】轉型正義何時來？原墾戶抗爭七十年(New7) ♦ 原墾民權益與國土保安難兩全？環團：適度限制發展權減少中下游居民致災(Hinet) ■ 原住民權益： ♦ 原住民的端午節悲歌：兩艘滑不動的舟(民報)

項目	調整後滿意度	網民負面態度事件摘錄 (摘列網民情緒小於等於-9 之文本)
		■ 弊案： ♦ 遠雄行賄官員，包庇海山礦災舊址造鎮蓋樓(PChome 新聞) ■ 土地登記的時效抗辯： ♦ 劉昌坪專欄：政府不應該在土地登記錯誤案件中主張時效抗辯(Yahoo)
地價稅	**25.8%**	■ 超收稅： ♦ [新聞]稅收超徵 4,000 億元，藍營雙北議員疾呼發消費卷(PTT HatePolitics)
公告地價	**26.0%**	
平均地權	**26.4%**	■ 呼籲打房： ♦ [新聞]高房價背後元凶，專家：囤地成本宜加重(PTT Gossiping) ♦ 巢運：地價稅妥協就是居住正義跳票(聯合新聞網、新浪網) ■ 打擊房市： ♦ [新聞]持有稅調，高調查：逾 5 成購屋者將止步(PTT home-sale) ■ 抗議漲稅： ♦ 地租、地價稅雙漲，企業壯士斷腕(Hinet) ♦ 稅制雙漲，上千人嗆臺南市府(Yahoo) ♦ [新聞]地價稅下月開徵，平均漲 30%(PTT Gossiping) ♦ [新聞]嘆要賣血繳稅，自宅地價稅 16 萬漲到 133 萬(PTT Gossiping) ■ 帶動物價上漲： ♦ 柯文哲：地價稅和房屋稅雙漲的民怨甚於油電雙漲(Hinet) ■ 農地問題： ♦ 宜蘭擬重評公告地價，鄉鎮長砲轟(Hinet) ■ 弊案： ♦ Re：[新聞]房產稅恐逾 300 億，遠雄圖謀「棄蛋」(PTT home-sale) ■ 公平性： ♦ [新聞]自用宅僅限 1 戶，綠委：對擁多房夫妻不公平(PTT Gossiping)
地籍	**26.4%**	■ 民間土地糾紛： ♦ 恆春「好嬸」命案偵破，兇嫌是 8 旬鄰居(蘋果日報) ♦ 八煙聚落封村，居民找回平靜(自由時報) ■ 人民與財團抗爭(亞泥事件)： ♦ [新聞]力挺傳領納私有地，朱天衣：財團搶土地(PTT Gossiping) ♦ [新聞]亞泥佔地 40+20，原住民保留地上的流浪記(DCard)
規定地價	**26.6%**	■ 地價稅上漲： ♦ 地價稅惹議，中市通過延期或分期繳納辦法(蕃新聞)
農地重劃	**26.7%**	■ 個人權益損害： ♦ 漂浪島嶼／郭志榮：黑暗重劃與哭泣園長(蘋果日報)
實價登錄	**28.2%**	■ 加速房市漲跌： ♦ 實價登錄加速漲跌，多頭空頭大不同(Yahoo 房地產) ♦ 實價登錄新增功能，歷次交易全都露(Yahoo 房地產)

2. 調整後營建類項目輿情滿意及不滿意情況

(1) 營建類滿意項目

分析結果顯示，營建業務相關關鍵字中滿意度最高的項目為「違章建築」及「農舍」，滿意度達 9 成以上，其次依序為「公共設施用地」、「青年住宅」、「包租代管」、「綠建築」、「都市更新」等。（詳表 20）

表 20、對營建業務滿意的項目

項目	調整後滿意度	網民正面態度事件摘錄 (摘列網民情緒大於等於 9 之文本)
違章建築	**91.0%**	■ 公權力聲張： ♦ 違建鐵皮屋成賭場，市府今強制拆除(蘋果日報) ♦ 工務局不手軟，鋼骨違建遭強拆(PChome 新聞網) ♦ 騎樓路歹「行」，工務局強制拆除，還路於民(PChome 新聞網) ♦ [新聞]議員施壓讓北市違建 10 年拆不掉，柯 P 霸氣(Gossiping) ■ 政策推動： ♦ 106 年度房屋稅籍清查作業，開始囉！ ■ 居住正義： ♦ 新竹縣創居住正義先例，「全國第一」尖石鄉鳥嘴部落完成 8 戶合法證明執照(PChome 新聞網)
農舍	**90.7%**	■ 爭議處理： ♦ [時事]農地違規加倍課稅引爭議，宜縣府：停用(PTT I-Lan) ■ 違規農舍處理： ♦ 宜縣將強拆 3 棟農舍，屋主靜悄悄(蕃新聞、Yahoo) ■ 農舍農用政策： ♦ [新聞]農舍豪華，宜蘭要徵豪宅稅(PTT、Hinet、MSN、Yahoo) ♦ 林聰賢：支持地方政府落實農地農用政策(MSN)
公共設施用地	**83.7%**	■ 地方建設： ♦ 西屯區龍潭里活動中心動土，林佳龍盼提供市民優質活動空間(新浪網、PChome 新聞網) ♦ [討論]台北越來越進步了(PTT HatePolitics) ♦ 體二用地解套再造，中市府盼重現「台中水源地」光彩(Hinet、Yahoo) ♦ 倡議 10 年！永康創意園區動土，預計 107 年 6 月完工(Yahoo、PChome 新聞網) ♦ 龜山區里基層建設座談會，鄭文燦：加緊腳步建設龜山(Yahoo) ■ 區域計畫： ♦ 內政部通過區域計畫，加速大台中 123(中央社)

項目	調整後滿意度	網民正面態度事件摘錄 (摘列網民情緒大於等於 9 之文本)
青年住宅	**83.0%**	■ 地方推動： ♦ 今天「構宅」了嗎？32 萬獎金等你拿(蕃新聞) ♦ 新北社會住宅動土典禮，朱立倫親主持(民視新聞、Hinet) ♦ 三重青年住宅招租，中籤率 4%搶破頭(民視新聞、Youtube、PChome、Hinet) ♦ 新北三峽青年宅 731 首次開放線上申請(Yahoo 奇摩新聞) ♦ 中和豪華青年宅落成，7 張小朋友入住(蘋果日報、蕃新聞、Yahoo 房地產)
包租代管	**82.0%**	■ 廣納民意： ♦ 內政部最新民調：近 8 成 4 接受住家附近蓋社會住宅(民報) ■ 鼓勵措施： ♦ 鼓勵空屋出租，租金 6,000 內享免稅優惠(PChome 新聞網、中央社、PTT Gossiping、東森新聞) ♦ 活絡租屋市場，委託包租代管可享租稅優惠(中廣新聞網、台視新聞) ♦ 政院拍板：租賃專法，祭 3 大租稅優惠(中時電子報、工商時報、PChome 新聞網) ■ 居住正義： ♦ 落實居住正義，政院宣示 8 年 20 萬戶只租不賣社會住宅(PChome 新聞網、Yahoo)
綠建築	**81.0%**	■ 前瞻藍圖： ♦ 智慧綠建築國際論壇，規劃未來城市藍圖(PChome 新聞網) ♦ 蔡總統：前瞻建設理念是與環境共生(中央社、TVBS) ■ 中央及地方推動： ♦ 三義鄉公所綠建築啟用，新穎大樓桐花意象造福鄉民(PChome 新聞網) ♦ 屏東縣工程大獲 2017 年建築園冶獎讚賞，8 件公共建築景觀與社區景觀營造類作品獲獎(蕃新聞、PChome 新聞網、Hinet) ♦ 桃市總圖新建競圖發表，綠建築「生命樹」獲第一(Hinet) ♦ 2017 建築園冶獎雲縣拿六項，打造宜居的舒適環境(Hinet) ♦ 總統接見 2017 建築園冶獎得獎企業及機關代表(Yahoo 房地產)
都市更新	**80.7%**	■ 都更政策： ♦ 內政部推都市針灸，小規模老宅重建(聯合財經網、PChome 新聞網、Hinet) ♦ 落實居住正義，內政部將設住宅都更中心(聯合財經網) ■ 地方推動： ♦ 北市府 1 月 12 日舉辦都市更新論壇(Hinet、PChome 新聞網) ♦ 閒置 17 年，原台中遠東百貨及綜合大樓都更啟動(Hinet、PChome 新聞網) ♦ 正義國宅都更聯貸案定案，政府可分 1,400 坪(東森新聞網)

(2) 營建類不滿意項目

營建業務相關關鍵字中滿意度最低的項目為「河川區」，滿意度僅 12.3%，其次依序為「濕地保育法」、「拆除執照」、「房地合一稅」、「使用執照」等。（詳表 21）

表 21、對營建業務不滿意的項目

項目	調整後滿意度	網民負面態度事件摘錄 (摘列網民情緒小於等於-9 之文本)
河川區	**12.3%**	■ 水環境建設： ♦ 鍾振坤：前瞻水環境建設有戰術沒戰略(蘋果日報)
濕地保育法	**18.5%**	■ 環評子法： ♦ 環評子法公聽，環團籲礦業展限納環評脫鉤先走，詹順貴：無法(Yahoo 奇摩新聞) ■ 環評案： ♦ 無視《濕地法》及黑琵保育，茄萣 1-4 道路再闖環評過關(Hinet) ♦ 新豐濕地座談會，鄉民翻桌尷尬散會(自由時報)
拆除執照	**18.9%**	■ 拆除爭議： ♦ [新聞]又現都更爭議，民眾協商完回家「房子被拆光」(PTT Gossiping) ■ 居民抗爭： ♦ 北市辦居住正義論壇，都更自救會場外抗議(蕃新聞、PTT Gossiping) ♦ 北市最難搞都更案動工，最快 2020 年完工(PTT Gossiping、民報)
房地合一稅	**19.5%**	■ 影響房市： ♦ 理財周刊／房價尚未落底，重稅政策、供需變化理性看待(東森新聞雲) ♦ 好房直播／五月繳稅加房貸，張金鶚：房地合一是關鍵(Hinet)
使用執照	**20.1%**	■ 執照申請困難： ♦ 現代建物多元需求，老建物改造大挑戰(Hinet、公視新聞) ♦ 台南民宿申請合法難，無照、增建頻卡關(Hinet、TVBS) ■ 違規使用案件： ♦ [新聞]獲高雄厝獎勵取照後變更，依法即拆(PTT Kaohsiung) ♦ 〈南部〉無使照，雙城地區逾九成民宿難合法(自由時報)

第六章 結論

(一) 大數據文字探勘技術為網路輿情分析之重要工具。

行政院國家發展委員會 106 年調查研究[11]顯示，唯手機族已占 24.3%比率，觀察唯手機族年齡結構，其中又以 20 至 39 歲年輕人最多，比例逾 6 成(20~29 歲占 30.7%，30~39 歲占 30.7%)，代表電話訪問(CATI)對合格受訪者的涵蓋率已愈來愈低，其代表性面臨極大挑戰。近年幾次重要的選舉都顯示電話民調似乎逐漸失靈。引起各界對於「電話訪問這個方法是否仍有效、準確」的廣泛討論。因此統計調查單位應研究造成失準的原因並評估其嚴重性，同時開發新的調查工具及資料整合分析工具來解決問題，故利用大數據文字探勘結合電話訪問調查作為輿情分析工具將是未來的趨勢。

(二) 電話訪問調查與網路輿情爬蒐兩者互補不足，以獲得具有代表性的調查結果。

市內電話訪問調查有隨機性、機率均等、調查結果可進行較嚴謹的統計推估等符合統計理論之優點，但其實也有涵蓋性不足、受訪者短時間無法充分瞭解問項內容以及成本過高的缺點，網路輿情爬蒐可適度彌補電話調查的缺點，但網路輿情爬蒐隨機性、代表性不足、無法進行統計推估等問題，又是目前無法克服的缺點，短期內較難用網路社群分析、網路輿情分析的結果直接取代電話調查的量化結果，需透過整合分析得到一個預測結果，並加以解讀再融會貫通，兩者互補不足，才能獲得具有代表性的調查結果。

[11] 國家發展委員會 106 年持有手機民眾數位機會調查報告網址：https://www.ndc.gov.tw/cp.aspx?n=55c8164714dfd9e9

(三) 整體結果可適度修正電訪員引導效應外，亦可修正電訪中年輕人受訪比例偏低所造成的偏差。

由於網路輿情資料具未上網偏差、未表態偏差及正向情緒不表態的偏差，導致網路輿情媒體正面或中立報導下民眾滿意機率容易被低估；而電話訪問的施政滿意度調查問卷大多為正向表述，又可以同時涵蓋未上網者、未表態者及正面態度者的意見，故本研究提出以電話訪問所得到的滿意度取代網路輿情媒體正面報導下民眾滿意機率，可適度修正電訪員引導效應，也可以修正電話訪問法中年輕人受訪比例偏低所造成的偏差。

(四) 地政業務最滿意為「土地增值稅」，最不滿意為「抵費地」；營建業務最滿意為「違章建築」及「農舍」，最不滿意為「河川區」。

調整後的地政業務綜合滿意度為 82.6%，營建業務綜合滿意度為 69.5%。詳細探討各業務滿意及不滿意關鍵字，地政業務最滿意為「土地增值稅」(細節摘錄如：可增加政府財源、討回大戶欠稅、退稅、稅改等)，其次依序為市地重劃、區段徵收、土地徵收，最不滿意為「抵費地」(細節摘錄如：長億集團涉市地重劃弊案)，其次依序為土地登記，地價稅、公告地價、平均地權、地籍、規定地價、農地重劃、實價登錄；營建業務最滿意為「違章建築」及「農舍」(細節摘錄如：公權力聲張、政策推動、居住正義、爭議處理、違規農舍處理、農舍農用政策等)，其次依序為公共設施用地、青年住宅、包租代管、綠建築、都市更新，最不滿意為「河川區」(細節摘錄如：前瞻水環境建設有戰術沒戰略)，其次依序為濕地保育法、拆除執照、房地合一稅、使用執照等。以上各業務滿意及不滿意關鍵字及細節摘錄可作為相關單位施政措施之改善方向，並提供決策者更具體可信的參考依據。

第七章 未來建議

(一) 文字探勘結合 CATI 電訪擴增至本部各項業務

由於需處理之資料量太大，106 年度先試行地政與營建 2 類資料，未來將再持續擴增至各項業務範疇。相信隨結構化與非結構化數據整合分析的技術日趨成熟，未來電話調查與大數據輿情探勘結合應用，必能獲得更真實的民意。

(二) 納入聲量考慮因素

1. 將關鍵字的聲量納入考量，以更明確瞭解民眾的感受。

目前已藉由內政領域詞庫中地政及營建的關鍵字作為文章細分類的依據，每個關鍵字用它相關的文章計算滿意度，可以解讀成：民眾討論到這個關鍵字的文章平均滿意度如何，滿意度越高應該就表示民眾目前對於這個關鍵字領域越滿意，越低代表越不滿意。由這些關鍵字去探究民眾對地政或營建業務滿意或不滿意的地方，因每個關鍵字的文章數量及留言數量(聲量)不太一樣，討論的文章縱使滿意度低，但因聲量小，造成的負面能量也相對較小，因此建議將關鍵字的聲量納入考量，以更明確瞭解民眾的感受，作為本部施政措施的參據。

2. 聲量探討能否考量到結構代表性，仍是目前網路輿情探勘、情緒分析及聲量指標尚待解決的關鍵疑義。

聲量指標部份，「總聲量」為網友提及該主題的次數，代表網友關注程度，但總聲量並非愈多愈好，需進一步區分正、負情緒聲量，亦非一昧追求零負評，而是維持正情緒聲量大於負情緒聲量(相除之後大於 1)；關於負評論，並非將它視為洪水猛獸，而是當作改進施政作為的動力或強化機關形象的契機。另外「集中度」係討論可能僅限於某特定族群，影響度並未擴散，例如討

論集中在自家粉絲團，傳遞對象有限，也無法觸及潛在目標對象，這項指標是指該主題在聲量數前 10 名中的量值，愈高表示愈集中，愈低表示分散愈廣。然媒體或社群中很多意見都是重複的，在沒有辦法檢視樣本代表性的情況下，最終結果有可能過度誇飾，正面可能低估，負面亦可能高估，故上述聲量探討能否考量到結構代表性，如常上網留言的民眾是否多為臺北人、男性多還是女性多、是否多為年輕人、如何校正等問題，仍是目前網路輿情探勘、情緒分析及聲量指標尚待解決的關鍵疑義。

(三) 完善詞庫周延性

1. 在關鍵字的萃取上需再聚焦，建議將關鍵字以集合方式建置，如 A 加 B 加 C 才是目標關鍵字。

研究中發現滿意項目與不滿意項目可能互為因果關係，或不滿意項目其實為滿意項目之子項目，如地政業務中，市地重劃是滿意項目，抵費地則最不滿意，但是抵費地只在市地重劃才有(子項目)，其原因為抵費地中，對重劃的弊案非常不滿意，可是對於市地重劃促進地方建設卻很滿意。因此在關鍵字的萃取上需再聚焦，於關鍵字建置時，建議將關鍵字以集合方式建置，如 A 加 B 加 C 才是目標關鍵字等，未來在關鍵字詞庫的建置過程中，將與業務單位討論，互相合作配合，將關鍵字詞庫建置地更加精確，畢竟文字探勘技術最重要且最精髓的部分就是詞庫的建置。

2. 建議文本由業務單位及統計處同仁共同辦理評分作業，作為影響度模型建置之訓練資料，藉以剔除與本部施政相關性較低的文本。

除了詞庫的建置外，需與業務單位配合的部分尚有採用支持向量機(SVM)建立「內政部影響度評估模型」過程中，將針對隨機挑選的 1,000 篇經過人工標記，對本部有高度正面影響的文本標記為 5 分，有高度負面影響的標記為-5 分，沒有影響的標記為 0 分，每篇文本建議分別由業務單位及統計處同仁共同合作辦理

評分作業，作為影響度評估模型建置之訓練資料，藉以剔除與本部施政相關性較低的文本。而為求模型建置之嚴謹，亦建議進行評分同仁間評分結果的信度檢定，以提高模型之可信度。

(四) 改良精進滿意度結果以貼近真實民意

在計算調整後輿情滿意度時，是以電話訪問 CATI 所得到的滿意度 P(CATI)取代 $P(民^{+}\mid 媒^{+})$，但 P(CATI)其實是民眾對施政感覺滿意(非常滿意+還算滿意)之比例加上不知道、很難說、沒意見、未回答之比例，建議以扣除不知道、很難說、沒意見、未回答後之重新計算之滿意比例 P^{*}(CATI)代替，經過再次調整後，地政業務的網路輿情滿意度如表 22，彙整後如表 23，營建業務的網路輿情滿意度如表 24，彙整後如表 25；雖然與前次調整的輿情滿意度差異不大，但藉此再次檢視滿意及不滿意之關鍵字及施政項目，配合關鍵字詞庫更精確的建置，以修正並釐清民意的動向。

表 22、再次調整後不同操作型定義下的地政業務綜合滿意度

媒	民	P(媒$^{+}$)	P^{*}(CATI)	P(媒$^{-}$)	P(民$^{+}$|媒$^{-}$)	P(媒^{0})	P(民$^{+}$|媒^{0})	P^{**}(民$^{+}$)
I	I	33.87%	93.91%	9.26%	4.44%	56.87%	8.91%	37.29%
I	II	33.87%	93.91%	9.26%	31.11%	56.87%	20.30%	46.23%
I	III	33.87%	93.91%	9.26%	31.11%	56.87%	84.16%	82.55%
I	IV	33.87%	93.91%	9.26%	2.22%	56.87%	98.02%	87.76%
II	I	47.44%	93.91%	25.45%	3.67%	27.11%	7.41%	47.49%
II	II	47.44%	93.91%	25.45%	18.35%	27.11%	18.52%	54.24%
II	III	47.44%	93.91%	25.45%	18.35%	27.11%	82.72%	71.64%
II	IV	47.44%	93.91%	25.45%	1.83%	27.11%	97.53%	71.46%

表 23、再次調整後的地政業務綜合滿意度彙整表

地政類	媒 I (媒體正向情緒分數≧6)	媒 II (媒體正向情緒分數＞0)
民 I(網民正向情緒分數≧6)	37.3%	47.5%
民 II(網民正向情緒分數＞0)	46.2%	54.2%
民 III(網民正向情緒分數≧0)	82.6%	71.6%
民 IV(網民正向情緒分數≧−6)	87.8%	71.5%

表 24、再次調整後不同操作型定義下的營建業務綜合滿意度

媒	民	P(媒$^{+}$)	P*(CATI)	P(媒$^{-}$)	P(民$^{+}$|媒$^{-}$)	P(媒 ^{0})	P(民$^{+}$|媒 ^{0})	P**(民$^{+}$)
I	I	33.47%	66.36%	9.78%	3.23%	56.75%	10.21%	28.32%
I	II	33.47%	66.36%	9.78%	18.28%	56.75%	20.60%	35.69%
I	III	33.47%	66.36%	9.78%	18.28%	56.75%	79.58%	69.16%
I	IV	33.47%	66.36%	9.78%	2.15%	56.75%	96.13%	76.97%
II	I	47.04%	66.36%	26.36%	3.33%	26.60%	11.42%	35.13%
II	II	47.04%	66.36%	26.36%	19.17%	26.60%	19.75%	41.52%
II	III	47.04%	66.36%	26.36%	19.17%	26.60%	78.09%	57.04%
II	IV	47.04%	66.36%	26.36%	1.67%	26.60%	94.75%	56.86%

表 25、再次調整後的營建業務綜合滿意度彙整表

營建類	媒 I (媒體正向情緒分數≧6)	媒 II (媒體正向情緒分數＞0)
民 I(網民正向情緒分數≧6)	28.3%	35.1%
民 II(網民正向情緒分數＞0)	35.7%	41.5%
民 III(網民正向情緒分數≧0)	69.2%	57.0%
民 IV(網民正向情緒分數≧−6)	77.0%	56.9%

(五) 現有技術基礎下深化辦理成果

1. 除了關鍵字逐項分析外，建議可新增關鍵事件推薦，或是關鍵事件之聚類，但聚類模型牽涉到較大的技術門檻，目前仍待評估。

文字探勘承襲資料探勘技術，擁有包含聚類、分群及關聯等

演算法，因此在聚類方面，除了滿意及不滿意關鍵字逐項分析外，建議可新增關鍵事件推薦，或是關鍵事件之聚類，因為透過模型所篩選出的不滿意關鍵字有可能同屬一個原因，或不滿意關鍵字與滿意關鍵字互為因果關係，甚至為其子項目，透過聚類可以向上彙整成關鍵事件並進行剖析，亦可在某種程度上解決滿意與不滿意之間的矛盾；但事件聚類模型牽涉到較大的技術門檻，目前仍待評估。

2. 針對民眾有感的業務項目，可透過網路輿情聲量的變化預測趨勢，但需先選定範圍以聚焦標的主題。

針對民眾有感的業務項目，透過網路輿情聲量的變化預測趨勢，以作為政策擬定方向之參考，例如針對營建業務可以計算民眾房價感受溫度計、民眾買房需求指數、民眾售屋意願指數、買賣供需平衡指標等，但不同業務有不同的指標可以做，且業務範圍太大，需先選定範圍以聚焦標的主題。

3. 以演算法進行斷詞，並自動進行情緒判斷，然而內容分析方法的客觀性易遭受抨擊，如何克服仍是目前有待精進的方向與挑戰。

將文字中的情緒與意圖進行量化分析會遭遇到一些困難及障礙，如俚語、諷刺、誇張、疊字、注音文等詞彙可能會阻礙數據的分析。情緒分析可以被歸類為一個分類上的問題，目的是將訊息依不同情感區分為正向、負向兩類，然而情緒表達是會隨著時間、地域、國家、身份的不同而改變，因此分析的演算法與判斷依據也會因此而不同，情緒判斷的準確性也會因此降低。文字探勘是以一套演算法進行斷詞，並將正負面語意的詞庫寫入程式中訓練演算法以自動進行判斷；然而內容分析方法的客觀性易遭受抨擊，如詞庫編碼人員的訓練、分析單位、信效度，甚至分析是否可重製等都備受質疑，因此文字探勘如何克服上述缺點，仍是目前有待精進的方向與挑戰。

第八章 參考文獻

1. 林千翔、張嘉惠、陳貞伶。2010。結合長詞優先與序列標記之中文斷詞研究。碩士論文。中央大學資訊工程研究所。
2. 林彩雯。2015。以 Google App 評論為字詞權重調整之情緒分析系統。碩士論文。靜宜大學資訊管理研究所。
3. 周智倫、林廷宇。2016。基於文本分類之社群網路內容分析。碩士論文。銘傳大學電腦與通訊工程研究所。
4. 陳稼興、謝佳倫、許芳誠。2000。以遺傳演算法為基礎的中文斷詞研究。碩士論文。中央大學資訊管理研究所、真理大學資訊管理研究所。
5. 陳宜惠、呂瑞麟、黃政傑。2013。斷詞系統對於 Queried keywords 的影響。碩士論文。亞洲大學資訊多媒體應用研究所、中興大學資訊管理研究所。
6. 陳冠廷。2016。基於社群網路情緒分析之民意預測研究。碩士論文。臺北科技大學資訊工程研究所。
7. 張育蓉。2013。使用情緒分析於圖書館使用者滿意度評估之研究。碩士論文。中興大學圖書資訊研究所。
8. 張曉珍。2013。運用文字探勘技術在社群行為上之人格預測。碩士論文。交通大學資訊管理學程碩士班。
9. 陳岳群。2014。使用情緒分析於公眾行為預測之研究。碩士論文。樹德科技大學資訊工程研究所。
10. 楊正銘。2004。以文字探勘技術應用於疾病分類之輔助系統−以出入院病歷摘要為例。碩士論文。台北醫學大學醫學資訊所。
11. 李淑惠。2014。運用文字探勘技術於口碑分析之研究。碩士論文。東吳大學資訊管理研究所。
12. 簡智宏。2015。應用文字探勘技術於概念股輿情與股價共同移動之研究−以蘋果供應鏈為例。碩士論文。政治大學資訊管理研究所。

13. 羅郁仁。2011。中文專利指標及文字探勘之研究。碩士論文。臺北科技大學工業工程與管理研究所。
14. 國家發展委員會。2017。106 年持有手機民眾數位機會調查報告。網址：https://www.ndc.gov.tw/cp.aspx?n=55c8164714dfd9e9。
15. 楊立偉、邵功新。2016。社群大數據：網路口碑及輿情分析。新北市：前程文化事業有限公司。
16. 余清祥、顏貝珊。2016。大數據：知識經濟與實務應用。臺中市：滄海書局。
17. 謝邦昌、鄭宇庭、謝邦彥。2017。玩轉社群：文字大數據實作。臺北市：五南圖書出版股份有限公司。
18. 謝邦昌。2015。Text Mining 文本探勘。新北市：中華資料採礦協會。
19. Viktor Mayer-Schönberger、Kenneth Cukier。2013。大數據。臺北市：遠見天下文化出版股份有限公司。