圖書資訊學研究17：2 (June 2023)：71-107
DOI:10.30177/JLISR.202306_17(2).0003

臺灣網路論壇關注之檔案事件主題及其情感分析

Public Opinion Mining and Sentiment Analysis for Archive Issues on Internet Forums in Taiwan

林巧敏*

Chiao-Min Lin

國立政治大學圖書資訊與檔案學研究所教授
Professor
Graduate Institute of Library, Information and Archival Studies
National Chengchi University

李育賢

Yu-Hsien Li

國立政治大學圖書資訊與檔案學研究所研究生
Graduate Student
Graduate Institute of Library, Information and Archival Studies
National Chengchi University

【摘要 Abstract】

大眾的意見經常是公務機關提供服務的參考依據，本研究借助數位工具，探討網路論壇中有關檔案議題的貼文，以巨觀及微觀的視角對其內容呈現的主題、情感、脈絡、關聯加以剖析。本研究以四家網路論壇為資料來源，蒐集2012 ~ 2021年間，共586篇檔案主題貼文，進行文本預處理後，使用WEKA、CORPRO、CVAW4.0及Gephi數位工具，依序進行主題分析、語料庫分析、情感分析及社會網絡分析。結果顯示，網路論壇討論的檔案議題以

*通訊作者：林巧敏 cmlin@nccu.edu.tw
投稿日期：2022年9月16日；接受日期：2022年12月27日

「檔案解密公開」為最多；從檔案相關關鍵字的高共現詞彙可觀察出近十年關注的檔案主題，與政治事件高度相關；網路論壇的檔案貼文內容偏負面情感，但在「檔案推廣應用」主題，則是正面高於負面，加入時間分析亦可觀察情感的轉折變化。

The opinions of the public are often the reference for public authorities to provide services. This study uses digital tools to explore the posts on archives issues in online forums, and analyzes and discusses the trend of public opinion in online forums from the perspectives of macroscopic and microscopic on themes, emotions, contexts, and associations presented in their content. Four online forums were taken as data sources to collect a total of 586 posts from 2012 to 2021. After text preprocessing, WEKA, CORPRO, CVAW4.0 and Gephi were conducted in sequence for theme analysis, corpus analysis, sentiment analysis and social network analysis. The results indicated that “archive decryption and disclosure” is the most discussed topic in online forums. According to the high co-occurrence keywords in archives online forum posts, the public opinion of archives in the past ten years can be observed, which is highly correlated with political events. The content of archive posts on online forums is slightly negative, and public opinion is more positive than negative for “archive promotion and access”. Adding time analysis can also detect changes in public opinion.

【關鍵詞 Keywords】

檔案社會議題；社群媒體；文本探勘；文本情感分析；主題探勘

Archives of Social Issues; Social Media; Text Mining; Text Sentiment Analysis; Topic Mining

壹、前言

社會大眾的情感認知是專業領域尋求社會支持的基礎，民眾往往是藉由報章與網路媒體取得時下重要資訊，尤其網路資訊隨手可得，網路輿論對於民眾觀點和認知的影響尤為嚴重，網路論壇已成為民眾抒發公共議題的主要場域。近年來，已有不少對於網路輿論的分析和研究，多數是對於政治、社會現象的討論（Dwianto, Nurmandi & Salahudin, 2021; Sabatovych, 2019; Yaqub, Chun, Atluri, & Vaidya, 2021），不僅能及時瞭解

民眾的看法和思想動態，亦可從中提取人們的訴求，有助於制定更適切的公共政策。

社會大眾的意見經常是公務機關提供服務的參考依據，檔案典藏機構亦是如此；相較於傳統的新聞報紙，在網路及智慧型手機的普及下，網路論壇及社群媒體成為了公民發揮「公共性」的場所（王毓莉，2007）。當民意趨向和檔案工作產生衝突，民眾會因缺乏對於檔案議題的認識，而對檔案事業產生負面印象（張奕萱、林巧敏，2022）。可知在網路議論事件的擴散過程中，網路文本的情感傾向對於事件往往發生關鍵引導作用，若能分析掌握這種情感擴散規律，制定應對策略，就能避免公眾情感偏極化（陳育正、孫懋嘉、顧志文、林立偉，2020；劉嘉薇，2017）。

臺灣檔案事業發展於《檔案法》通過及檔案管理局設置後，我國檔案管理工作以及檔案應用的觀念開始受到社會重視，而近幾年也不乏社會議題因牽涉檔案還原史實而受到媒體關注，使得社會大眾對於檔案的認識受到時事輿論的影響，尤其在網路資訊傳播快速的時代，在網路社群造成議論焦點後，經常會影響社會大眾對於檔案專業的認知與評價（張奕萱、林巧敏，2022；Huang & Lin, 2020）。若能針對媒體報導討論的內容，進行內容與情感分析，將有助於檔案管理機關瞭解社會關注的檔案議題，提供更好的公共決策，甚至可藉由情感分析的瞭解，預測多數民眾在閱聽後，可能產生的情緒傾向，對於媒體公關和服務推廣過程，能以公眾理解的角度編整內容，促進社會大眾對於檔案事業有更多的關注和瞭解。

有鑑於網路論壇所帶引的議題風向，可能牽動檔案機構以及檔案管理工作之發展，影響力不容小歔。但國內對於網路社群的內容分析，即使存在網路媒體推出每日更新的輿論分析，計算網路聲量及正負面聲量比例，提供時下網路世界中對於公眾人物或特定事件的討論熱度及排序分析，例如：網路溫度計（https://dailyview.tw/）；或是已有商業公司開發網路輿論探測工具，如OPview網路口碑資料庫（https://www.opview.com.tw/）提供資料分析服務。然而針對「檔案」此一非熱門的特殊議題，在一般社群及論壇受到的關注和討論聲量不高，無法在一般網路大數據分析平臺取得正確數據。如果希望探究在網路輿論中有關檔案的文本內容和情感傾向，需要研究者藉助數位工具進行論壇內容之篩選及分析。

因此，本研究以實作擷取網路論壇內容，運用數位工具剖析檔案議題內容，試圖以文本探勘技術，找出受到社會大眾關注的檔案議題與核心概念，瞭解社會大眾對於檔案時事的觀點及其情感意涵。因社群論

壇內容往往比新聞報導擁有更自由論述的空間，故而分析論壇內容，可更貼近民眾觀點，有助於洞悉社會大眾對於檔案的看法與情感認知（Huang & Lin, 2020）。本研究期許可提供我國檔案機構未來推廣檔案素養教育、改善公共關係、提升社會形象等相關政策研議之參考。研究目的在於：

一、分析國內議論時事的網路論壇中，討論與檔案相關的內容主題以及高頻人物、事件或機構團體，掌握網路議論檔案的主題面向，以策進檔案工作發展。

二、探討網路議論檔案的情感傾向，分析對於重要主題、人物、事件或機構團體的情感認知，作為改善檔案社會形象之參考。

貳、文獻探討

一、文字探勘技術及相關研究分析

文字探勘（text mining）又稱為「文本探勘」，是相對於發展較早的「資料探勘」（data mining），由結構化資料的分析，轉向半結構或非結構化的文字資料研究；藉由對文字的進階運算、統計整理及量化文字詞頻等方式，可快速分析或找出文本中隱含重要的資訊或有意義的規則（Feldman & Sanger, 2007）。相較於結構化資料，由文字組成的非結構化資料屬於自然語言處理的範疇，但因數位工具的進展，領域學者得以藉助已開發工具的操作，挖掘隱藏在文字中的意涵，甚至有學者提出「文字即資料」（text as data）的概念，突顯出探勘文字資料的研究趨勢（Grimmer & Stewart, 2013）。

文字探勘是多項專業結合的研究技術，對巨量文字進行編輯、組織與分析的過程，可發掘文字間隱含的關聯特徵或是新穎模式（Blake, 2011）。常見的文字探勘技術包含：詞頻統計（term frequency）、關鍵詞提取（keyword extraction）、共詞分析（co-occurrence analysis）、集群分析（clustering）、情感分析（sentiment analysis）、主題模型（topic model）、詞嵌入模型（word embedding）。在電腦運算速度飛速進展以及演算法不斷開發之下，亦出現各種深度學習（deep learning）模型（王貿，2020）。文字探勘的技術與方法已發展出各種數位工具，可供使用者學習運用，但推陳出新的工具也造成使用者面臨「選擇」適用工具的判斷困難。目前顯然沒有最佳、最好的方法可用於分析所有類型的文本，需要根據文本內容特性以及需要探究的問題，尋求適切的方法與工

具（邵軒磊、曾元顯，2018；Grimmer & Stewart, 2013）。因此，以數位工具進行文字探勘，雖是以自動化方式大量降低人工時間與成本，但仍需要內容專家參與驗證，比較能確保結果的正確性。

運用網路論壇進行文字探勘之研究，國外常以臉書（Facebook）及推特（Twitter）為研究對象。其中與圖書資訊領域有關之研究，有Al-Daihani與Abrahams（2018）分析學術圖書館在社群媒體的行銷策略，以2014年《世界大學學術排名》英語系前100名的大學圖書館為對象，匯集各大學臉書2013 ~ 2014年間張貼總數18,333篇的文章，結合敘述統計、內容分析、語意分析，找出最高與最低參與的貼文內容，分析此兩類文章中，分布最廣泛的詞彙概念，藉此瞭解讀者對於學術圖書館感興趣的主題及其服務，可提供圖書館調整改善在社群媒體的行銷策略。

Patra（2019）使用Twitter API於2019年6 ~ 7月間下載18所印度圖書館的推文，使用R軟體分析推文內容高頻詞以及對於主題情感的認知。在擷取分析的5,540篇推文中，可發現印度圖書館在Twitter上的活躍度並不高，推文出現頻率最高的詞，取決於各圖書館熱中推廣的服務類型。對於推文的情感分析，則顯示多數圖書館推文屬於正面積極的表現。

陳育正等人（2020）為瞭解中國大陸在網路上對臺灣進行的統一戰線工作，使用「庫博中文獨立語料庫分析工具」，以中國大陸對臺灣進行統戰的「對臺31項措施」進行語料庫分析。研究對象分為中國大陸報導及國內報導，前者使用「hibuzzer輿情匯流平臺」針對中國大陸、港、澳地區的網路新聞爬文；後者透過「OpView社群口碑資料庫」進行資料蒐集，最後搭配語料庫分析工具，呈現詞頻統計、文字雲、正負情感比值、時間趨勢圖和詞語共現網絡圖，找出網路輿論中「誰在帶風向」以及「風向的內容為何」。

郭文平（2020）以「語料庫輔助媒體論述分析」（corpus-assisted discourse analysis, CADS），分析臺灣《聯合報》、《中國時報》、《自由時報》、《蘋果日報》等四家平面媒體，關於「中國夢」的報導，共計1,100篇文本，嘗試透過CADS的詞頻統計、共現詞分析以及語境脈絡分析，從微觀字詞分析到巨觀論述分析的模式，闡述該模式導入人文研究的可能性及特點。

綜觀以上研究，可以發現進行文字探勘是以資料蒐集最大化為目標，採用與研究問題相關的論壇為分析標的；而隨著專業領域、研究對象、研究問題的差異，採用的研究工具也大相逕庭。參考相關文獻對於網路論壇內容分析的啟發，本研究可採用詞頻統計、主題模型、情感分析以及共現分析，作為本研究進行文字探勘的方法。

二、文本情感分析技術和相關研究探討

情感分析領域的研究起源於1990年代末至2000年代初，彼時自然語言處理技術上有極大的突破，開始可透過斷詞區分情緒，最早的分析標的是股票市場的留言板（stock board），透過統計式或簡單機器學習模型判斷使用者的評論為正面（positive）或負面（negative），以此達到二元分類的功效（陸澤凱、謝穎，2021；Tang, Tan, & Cheng, 2009）。Dave、Lawrence與Pennock（2003）採用「意見探勘」（opinion mining）一詞，針對產品的屬性，挖掘文本中的意見，分為好、中、壞三類。文本情感分析雖然是由商學領域發跡，但運用的面向廣泛，是一個跨領域研究的主題，涉及自然語言處理、資料庫、資訊檢索、資料探勘、人工智慧等領域知識（周立柱、賀宇凱、王建勇，2008；Lopatovska & Arapakis, 2011）。

Lee、Kim與Song（2020）將情感分析的技術，分為監督式或非監督式兩大類，前者如「字典法」（lexicon-based），後者常見的是「機器學習法」（machine learning-based），近年由於類神經網絡研究的突破性發展，也嘗試加入「深度學習法」的探勘方式。「字典法」是建立於情感詞庫的基礎工程，分析結果的精確度，取決於詞庫的規模大小、更新與維護工作，優勢在於對特定領域的情感分析表現優異，然而無法統合上下文的關聯進行綜合判斷（李建華、劉功申、林祥，2017；張奕萱、林巧敏，2022；Palomino & Aider, 2022）。

「機器學習法」是透過演算法，先行建立已經標記完成的訓練集，對於未經標記的文本進行預測，當資料量足夠龐大，可以達到比字典法高出許多的正確率，是屬於較為客觀的情感分析方法，然而容易受到訓練集本身的領域屬性與數量影響分類準確性（Huang & Lin, 2020）。陳迪等人（2021）認為一般的機器學習方法，在網路論壇的情感挖掘上不盡如人意，原因有三：（一）模型嚴重依賴特徵工程，需要人工提取訓練，工作繁瑣而龐大；（二）由於特徵被提前設定，模型難以充分利用訓練數據；（三）基於領域知識提取的特徵，無法有效地應用到新的工作項目。

Thelwall（2017）指出社會學者對於機器學習的情感分析寄予厚望，網路文本是容易蒐集大眾意見的實證資料平臺，透過意見探勘技術可以幫助社會學者偵測到大眾對公共議題的情感傾向，特別有助於大規模的意見蒐集時，此類巨量資料分析有助於建立某種層次的社會網絡，情感分析技術的研究領域，包括：社群媒體的使用、人機互動、行銷、經濟、健保、政治、教育等議題。

在實證分析的研究中，Na、Thet、Nasution與Hassan（2011）使用Fedora資料庫工具，蒐集876篇電影評論，建置出基於情感的數位影評圖書館，可選擇以電影、編劇、角色三者為評價，分別呈現正面、負面及中性情感，最後加入使用者對於此一情感分類的數位影評圖書館的認同度投票。

LIBER年會是歐洲研究型圖書館人員專業交流的重要場域，Papachristopoulos與Tsakonas（2020）藉由分析第47屆LIBER年會的5,500條Twitter貼文，瞭解LIBER年會對於學術交流生態系統所發揮的功能，分析與會者表達的熱門話題、對於組織和會議的看法，並分析貼文表達的正面與負面情感事件，可提供會議籌辦以及會員關係改善的參考。

Stewart與Walker（2018）希望瞭解與非裔相關的學院和大學（Historically Black Colleges and University）在Twitter社交媒體中，討論圖書館的意見分析，利用IBM Watson的分析引擎，收集為期18個月超過23,000條推文，包括轉發和情緒的傳播，以評估追隨者的參與度。分析發現此類發表圖書館貼文鮮少有追隨者參與，也觀察到大量圖書館推文是圍繞著機構宣傳主義，而不是圖書館現象的討論。研究發現非追隨者傳播的貼文，反而有比較多的轉發，而圖書館宣傳貼文雖呈現正面情感，但轉發傳播量低，以致無法擴大影響力。

Lamba與Madhusudhan（2018）提供了一項運用社群討論支援圖書館學科服務的可行方向，圖書館運用情感分析工具進行在Twitter討論與生產力經濟學相關主題及其情感態度，使用RapidMiner為資料探勘工具，從Twitter檢索與生產力相關的貼文，採用AYLIEN文本分析軟體進行情緒分析。從13天的Twitter討論中，挖掘出6,416條相關推文。發現討論生產力經濟問題，經常被議論的地區是英國，其次是印度、美國、中國，大多數推文屬於中性情感，此一研究是基於分析讀者於網路論壇的輿論焦點與情感傾向，協助圖書館進行學科服務的調整方向。

Park、Bier與Park（2021）除了使用機器學習的情感分析方法之外，也結合共現分析、社會網絡分析，調查2018年7月1日 ~ 2019年5月17日間，因「南北韓高峰會」、「川金會」舉行，在網路聲量大增的北韓相關影片的留言評論，共計23,774篇。分析影片留言評論，探討南韓民眾對於北韓的負面印象，是否會因影片內容大幅增加，而有所轉變。情感分析結果顯示，正面情感的文本占59.55%，負面占31.41%，中性則有9.03%；在共現分析中，以「北韓」此詞彙為例，經常共同出現的負面詞彙有「共產黨」、「左翼」，在正面情緒詞彙中則有「統一」、「和平」；利用社會網絡分析，觀察留言者之間的交流互動，發現留言者的

來源很平均，並且留言者之間的互動很少，只有0.1%的使用者會給予他者回覆。但節目形式的軟性新聞受到群眾的歡迎，能有助於改善新聞中對於北韓的負面形象。

Dwianto等人（2021）觀察到美國與印尼在新冠肺炎（Coronavirus disease 2019, COVID-19）爆發期間，兩國國家元首皆透過Twitter公開宣傳防疫相關政令及應對措施，故以2019年9月21日～10月21日間，採用佐柯威（Jokowi）、川普（Trump）與新冠肺炎等關鍵字為搜尋字詞，探討民眾對於兩位元首政策推文的情感傾向，利用Brand 24自動化軟體，蒐集38,813篇推文，並進行情感分析，結果顯示民眾對於兩位元首推文的情感是負面大於正面。

可知已有越來越多的研究將視線關注於網路文本的情感分析，也逐步建立起實施文本情感分析的研究流程。原本基於英語文本研究所發展出來的斷詞工具和情感詞典不免有語言環境的差異，即使已有中文斷詞及情感詞庫之運用，但仍然需要有不同領域的實證應用研究，才能瞭解現有工具和技術在中文環境操作的適用性，進而始可發展在中文語境的情感分析方法和工具使用的建議。

褚乃慈（2020）針對臺灣獨立樂團的歌詞，以及樂迷留言進行情感分析，根據Russell（1980）提出的模型，採用字典法的情感分析模式，將斷詞結果與CVAW4.0情感詞典進行對照；在樂迷留言部分，以Youtube平臺上的留言為研究對象，分析三個部分：（一）估計單篇留言的情感分數；（二）從單篇留言的情感分數估計歌曲的整體留言情感分數以及各留言的一致程度；（三）從單篇留言的情感分數估計各年度的情感分數。最後以情感分數趨勢圖，觀察各年代的情感變化。

Huang與Lin（2020）蒐集「知識贏家資料庫」與「自由時報電子資料庫」中，與檔案事件相關的新聞主題文本，使用文本探勘方法，利用CKIP中文斷詞和詞頻統計程式，排序各詞性中的高頻率出現詞，找出文本中核心的人名、地名等概念；在探討檔案新聞文本情感意涵上，使用「機器學習法」和「字典法」兩種模式，前者利用WEKA的演算法製作訓練集，並調整TF-IDF特徵值擷取，後者使用CVAW4.0中文維度型情感詞典搭配Excel進行情感分數的計算，研究結果得出434篇檔案新聞主題文本中，約有61%的文本呈現負面情感。

張奕萱與林巧敏（2022）以中國大陸專注時事的「風聞社區」論壇為研究對象，分析近兩年（2019～2020）關於檔案的論壇貼文呈現的主題概念及其內容情感傾向，運用NLPIR大數據語義分析平臺進行簡體

中文斷詞、詞性標註及詞頻統計等文本預處理，進而採用中文情感詞彙本體庫進行文本情感分析。研究結果顯示，民眾最關注的檔案議題歸為兩類，一是以「毛澤東」、「蔣介石」、「斯大林」及「中蘇」、「國共」和「抗美援朝」為核心的特定人物與事件議題；二是以「檔案工作實務」和「檔案公開」為代表的社會議題。在情感方面，檔案主題貼文對於特定人物與事件討論時，負面情緒明顯；而在談及檔案工作實務時，則展現出對於檔案開放應用的正面態度。

前述Huang與Lin（2020）、張奕萱與林巧敏（2022）皆曾使用字典法與機器學習法，驗證兩個方法所得出結果，並無太大的差異，採行字典法，可利用經多次實驗驗證的CVAW4.0中文維度型情感詞典，計算情感總分數，並回推各文本，以瞭解每篇文本中呈現「正面」、「負面」或「中性」的情感傾向，進而以WEKA機器學習法輔助驗證字典法的正確性。綜合前述文獻，可知字典法是多數採用的方式，在比對文本詞彙，計算情感分數後，賦予每篇文本正面、負面或中性的情感傾向分類。本研究參照Park等人（2021）將情感分數計算結果，與主題分析、共現分析加以結合，並加入時間趨勢圖結合情感分析，探討特定主題情感的時間變化。

參、研究設計與實施

一、研究設計

本研究進行過程分為文本預處理與內容分析兩部分，整體研究設計如圖1所示。

第一部分透過關鍵詞檢索、日期排序篩選，蒐集四個網路論壇（想想論壇、風傳媒、關鍵評論網、鳴人堂）中，與檔案議題相關之文本（貼文），使用CKIP Tagger進行斷詞與詞性標註，並建立權威控制檔，反饋清洗文本。第二部分使用數位工具進行四項分析：

（一）主題分析：先透過WEKA進行自動化主題分析，找出檔案主題下的子分類，根據斷詞後的詞頻統計，排序列出高頻詞彙，分析網路論壇中對於各項檔案主題之討論聲浪。

（二）語料庫分析：利用CORPRO語料庫分析工具的共現分析，找出高頻詞彙之間的關係。

（三）社會網絡分析：透過詞頻統計篩選出各詞性的高頻詞彙，可觀察

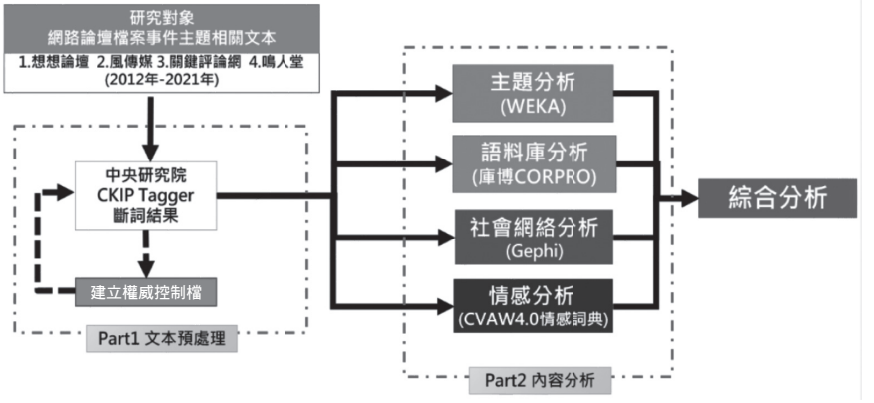

圖1　研究設計示意圖

出文本中較常出現的人名、機構團體與事件，使用Gephi工具繪製出這些主題詞彙之間的關係遠近，和群集之間的關聯性。

（四）情感分析：採用CVAW 4.0中文維度型情感詞典，搭配Excel與Access軟體，將文本斷詞與情感詞典相互比對，並回推計算出各篇文本的情感分數總和，以呈現各文本內容文字呈現的正面或負面情感傾向。

最後將四項工具分析結果，進行交叉比對，可巨觀地看見文本整體的脈絡傾向，也可微觀地詳細剖析文本的內容，再輔以文獻探討進行質性解釋（表1）。

表1

本研究採用之分析工具說明

研究項目	分析工具	工具說明
文本預處理	CKIP Tagger	由中央研究院中文詞知識庫小組所開發，為2011年上線之舊版斷詞系統（CKIPWS [classic]）之改良版本，根據CKIP Lab經由ASBC 4.0 測試集（50,000句），斷詞結果優於舊版斷詞系統及中國大陸開發的結巴斷詞系統（繁體中文版），平均斷詞正確率可高達97%。

表1

本研究採用之分析工具說明（續）

研究項目	分析工具	工具說明
主題分析	WEKA	由紐西蘭懷卡托大學（University of Waikato）所開發，全名為懷卡托智慧型分析環境（Waikato Environment for Knowledge Analysis, WEKA），是一款用於資料探勘與文字探勘的軟體，主要有分類、分群、預測及關聯式規則預測功能，使用者不必自行編寫程式，即可實現機器學習，可以藉由下載擴充套件，執行不同的演算法（Hall et al., 2009）。
情感分析	CVAW4.0情感詞典	由元智大學資管系禹良治的實驗室團隊開發，4.0版有5,512個中文單詞可供使用。在此情感詞典中，參考情感心理學理論，包含效價（valence）及喚起（arousal）兩個維度，前者代表情緒的正負面程度，後者表示情緒從平靜到興奮的極距，分數範圍由小到大，分為1至9分（Yu et al., 2016）。
語料庫分析	庫博CORPRO	「庫博中文獨立語料庫分析工具」（簡稱庫博CORPRO）是由陳光華與闕河嘉共同主持開發，適用於中文特殊語境的數位工具。本研究主要使用其中的「共現分析」及「詞語趨勢」功能（闕河嘉、陳光華，2016）。
社會網絡分析	Gephi	Gephi為一基於NetBeans平臺和Java開發的網路分析和視覺化軟體，用於進行社會網絡分析，透過電腦計算出節點與節點、群集與群集之間，以及群集內的節點關係，並以點線之間的密集關聯程度，呈現視覺化效果（劉勇、杜一，2017）。

二、研究對象

本研究選取之論壇以有公開姓名或筆名，對於貼文內容負責之網路評論為對象，若屬提問和回應的簡短內容，不適合分析。經瀏覽各家論壇後，以與檔案概念相關之關鍵字進行檢索，採用「檔案」、「檔案管理」、「檔案館」，以及在中文環境常與檔案關連之「文書」或「紀錄」等詞彙併同查詢，並以搜尋結果篇數達50筆以上的論壇為符合資料蒐集對象，符合條件的論壇包括：「想想論壇」、「風傳媒」、「The News Lens 關鍵評論網」（簡稱關鍵評論網）、「鳴人堂」為本研究對象（表2），經檢索及人工判斷後，共獲得586篇與檔案相關的貼文。

表2

網路論壇內容及相關文本篇數說明

來源名稱	論壇內容說明	相關篇數
想想論壇	想想論壇（https://www.thinkingtaiwan.com/）於2012年上線，主要關注臺灣的教育、歷史、經濟、政策、法治及文化等社會議題，鼓勵民眾時常「想想」公眾議題。於徵稿上，開放民眾以真實姓名及筆名投稿，提供公民公共議題討論的網路空間。	172
風傳媒	風傳媒於2014年開站（https://www. storm.mg/），為網路科技公司「風傳媒集團」設立，發揮媒體扮演的「第四權」角色。網站內容接受讀者投書，也有長期約稿，投稿者須附上姓名、單位及聯絡方式。	219
關鍵評論網	關鍵評論網（https://www.thenewslens.com/）2013年正式成立，2015年中英雙語國際版網站上線，以「獨立評論媒體」為自我定位，兼具新媒體與新創企業雙重身分，該論壇開放讀者以真實姓名進行讀者投書，也有特約記者與作者群持續筆耕。	119
鳴人堂	鳴人堂（https://opinion.udn.com/opinion/index）原為聯合新聞網體系下的「udn專欄」，於2000年隨聯合線上服務上線，改版後獨立成站，該論壇雖未開放一般民眾投書，但可透過商談成為特約作者，內容涵蓋時事觀察、社會運動、法律知識推廣、文藝評論、國際時事等層面。	80

三、資料處理與分析

（一）文本預處理

瀏覽檢視國內收錄與檔案主題相關的網路論壇，決定適合分析之論壇對象後，以網路爬蟲並輔以人工判斷，擷取文字資料，匯集成Excel資料集，接著利用CKIP Tagger進行文本內容斷詞與詞性標註，觀察結果後，進行詞彙權威控制修正，避免同義異名問題，接著運用CKIP詞頻統計程式，分析各類詞性並統計高頻詞彙。

（二）內容分析

採用四個數位工具，進行內容主題分析、語料庫分析、情感分析及社會網絡分析，並將呈現結果進行交叉比對。

1. 主題分析

運用WEKA進行自動化主題分群，透過「string to word vector」套件執行預設的詞袋模型（bag of words），以無監督式的機器學習，將586篇文本，拆分成數個次主題，藉此歸納出網路論壇中常被議論的內容主題。本研究採用層疊式*K*平均法（Casades *K*-means），可以兼顧群內差異小，群外差異大的分群目標，不預設最佳的分群數量，而是透過反覆試驗，設定分群數目門檻值（閾值），以找出最佳分群數量。若為太少群集數，組內概念混雜，太多群集數也難以解釋結果，通常以3 ~ 7群最具解釋力。本研究將不同閾值群分出來的3 ~ 7群結果，考量各類高頻詞的的互斥性，以及單類主題的聚集性，並參考Pedregosa等人（2011）鑑別集群同質性的概念，輔以學科專家檢視不同閾值分群結果，選擇同質性概念涵蓋在相同分群主題，並兼顧同一集群詞彙概念完整性，決定閾值的分析結果。並將WEKA分群結果匯入「WEKA分群結果分析器」，以統計原理提取各群集的高詞頻，以及群集間相對顯著的特徵值。

2. 語料庫分析

為瞭解文本中的脈絡關係，使用CORPRO工具，匯入已經完成斷詞的文本，採用工具「共現分析」和「詞語趨勢」功能，挖掘文本與文本間、文本與詞句間，及詞句與詞句之間的關聯性。並以「編輯詮釋資料」功能，輸入需要觀察分析的關鍵詞，瞭解重要詞彙的時間分布趨勢。

3. 情感分析

採用情感分析中的字典法，利用Excel與Access輔助，將經過斷詞後的文本內容，與CVAW4.0中文維度型情感詞典比對，以共同出現的3,519個詞彙，回推各篇文本中出現的情感詞彙數量及詞頻統計，情感詞彙原始分數範圍介於1 ~ 9之間，將數值轉換成正、負面，較容易辨別其情感傾向。因此，本研究將原本情感的分數皆減5，轉換分數為-4 ~ 4之間，分數大於0為正面情感，小於0則為負面情感。然後將各情感詞彙分數乘上詞頻，將每篇文本的正、負面詞彙分數加總並除以詞彙數之後，得到每篇文本的情感數值，藉此回推統計每篇文本呈現的情感傾向為正面或負面。

4. 社會網絡分析

依照詞性列出普通名詞、地名、人名及專有名詞等可能具有特殊意涵的詞彙，使用Gephi工具，以節點的形式呈現，繪製成社會網絡圖，用以觀察人名、專有名詞、事件之間的關係遠近。斷詞後詞彙數量龐大，本研究採用高頻詞中的「專有名詞」，因其詞彙意涵專指性高，可為社會網路分析的標的。

肆、研究結果分析

一、網路議論檔案主題分析

（一）高頻詞統計

本研究完成586篇網路論壇貼文的收集、整理與匯總後，以CKIP Tagger程式，進行文本斷詞與詞性標註，得到共58,057個詞彙，總詞頻數為904,647，參照詞性分析，分為九大詞類。根據總詞頻統計，可知在九大詞類中，名詞占57%，動詞占33%，兩者涵蓋90%以上的詞彙。由於名詞與動詞具有比較明確的意涵，本研究將此兩大詞類，進一步找出核心概念。先去除無明確意義的單字詞（例如：「人」、「者」、「大」、「做」），並且將因異體字或慣用簡寫被拆分的詞彙合併（例如：「台灣」及「臺灣」、「中國國民黨」及「國民黨」、「共產黨」及「共黨」），得出前20名的高頻名詞（表3）及動詞（表4），普通名詞中，

表3

排序前20的高頻名詞統計表

	普通名詞（Na）		專有名詞（Nb）		地方詞（Nc）	
排序	詞彙	詞頻	詞彙	詞頻	詞彙	詞頻
1	檔案	2,705	國民黨	1,189	臺灣	4,589
2	政府	2,029	蔣介石	891	中國	1,850
3	歷史	1,580	中共	501	美國	1,513
4	國家	1,566	蔣經國	387	日本	1,298
5	政治	1,197	白色恐怖	361	大學	620
6	資料	987	民進黨	360	促轉會	619
7	社會	934	蔡英文	311	委員會	607
8	問題	856	二二八事件	271	國際	594
9	總統	817	陳儀	210	檔管局	504
10	轉型正義	789	共產黨	335	公司	488
11	時期	741	施明德	180	臺北	460
12	事件	739	朴槿惠	165	中華民國	442
13	機關	679	李登輝	153	香港	429
14	人民	578	彭孟緝	134	英國	390
15	工作	553	黃國書	128	德國	364
16	民主	535	媽祖	123	中央	352

表3

排序前20的高頻名詞統計表（續）

	普通名詞（Na）		專有名詞（Nb）		地方詞（Nc）	
排序	詞彙	詞頻	詞彙	詞頻	詞彙	詞頻
17	文化	529	陳文成	99	高雄	308
18	運動	529	美麗島	92	學校	299
19	政權	520	葉虹靈	88	東德	284
20	時代	508	張發奎	83	國史館	282

表4

動詞前20名高頻詞彙統計表

	動作及物動詞（Vc）		狀態不及物動詞（Vh）		動作不及物動詞（Va）	
排序	詞彙	詞頻	詞彙	詞頻	詞彙	詞頻
1	進行	714	相關	683	存在	217
2	發展	657	重要	583	殖民	159
3	統治	470	這樣	448	工作	154
4	處理	461	不同	428	立法	141
5	開放	443	出現	356	保密	139
6	成立	416	獨立	291	和解	138
7	監控	415	特別	284	執政	124
8	提出	395	樂活	266	革命	121
9	參與	347	受難	265	作戰	117
10	面對	302	自由	262	運作	115
11	出版	293	一樣	257	觀光	104
12	完成	291	直接	250	叛亂	100
13	接受	289	正式	232	抗戰	100
14	使用	279	安全	216	防疫	100
15	支持	273	如此	205	犯罪	99
16	參加	264	戒嚴	205	消失	90
17	通過	264	受害	204	出獄	76
18	推動	260	嚴重	186	示威	75
19	協助	251	積極	178	出生	73
20	審判	245	合作	171	集會	68

除了「檔案」為檢索詞彙成為高頻詞之外，其餘詞頻統計結果，則呈現政府、歷史、國家、政治、資料、社會等詞彙是在議論檔案時會經常出現的概念；在專有名詞部分，則以國民黨、蔣介石、中共、蔣經國、白色恐怖、民進黨等與政治相關人物及團體為主；地方詞統計，則顯示與臺灣關係密切之中國、美國、日本等地區，而與檔案管理相關的機構團體，常見檔管局、促轉會、國史館等高頻詞。

分析前20名動詞可以補充說明與名詞搭配的動作或狀態，在一般及物動詞（Vc）中，例如：檔案處理、檔案開放、檔案出版，狀態不及物動詞（Vh）中，受難、自由、戒嚴、受害，可能與白色恐怖有關；在動作不及物動詞（Va）中，立法、保密、革命、作戰、叛亂、抗戰等高頻詞，雖然概念較為龐雜，但印證文本內容，主要是討論與二二八事件、戰爭記憶有關的議題。

將前述排序前20之名詞與動詞，以文字雲呈現視覺化圖示如圖2所示。

圖2　網路議論檔案主題高頻詞文字雲

（二）WEKA主題分群結果

運用WEKA進行自動主題分群，發現當群分數量越少，每一類呈現不只單一的主題概念，以分為3類為例，同一類中可能隱含有多元龐雜的概念；但若採群分為7群的結果，類和類之間的高頻詞會有相同詞彙出現的情況。考量各類高頻詞的的互斥性，以及單類主題的聚集性，經學科專家檢視不同閾值分群結果，採用主題分為5大類的結果，最貼近檔案學科的主題概念，如表5所示。分群產生的五大主題為：

1. 檔案解密公開：出現解密、機密、保密、審查等高頻詞，圍繞著因機密檔案開放與否的相關問題討論。
2. 轉型正義與政治檔案：出現政治、轉型正義、威權、促轉會等高頻詞，主要討論因《促進轉型正義條例》引發對於政治檔案之認定與爭議的相關討論。
3. 歷史研究：出現歷史、社會、國家、文化、事件等高頻詞，圍繞著運用檔案支援史實研究的相關討論。
4. 檔案推廣應用：出現委員會、檔案管理、國家檔案、國發會、檔管局等檔案業務機關詞彙，以評論檔案機關推廣檔案意識的相關活動為主。
5. 檔案史料稽憑：出現臺灣、中國、蔣介石、日本、美國等與中國近代史發展相關的重要詞彙，圍繞著關於檔案提供還原歷史憑證的歷史事件討論。

各主題分群文本篇數是由WEKA根據詞彙分析計算結果，其中文本數最多的主題為「檔案解密公開」，其次為「轉型正義與政治檔案」，代表被社會大眾高度議論的檔案主題，多半是因社會事件或是政治議題引發的輿論，例如對於檔案解密公開的呼籲、政治檔案條例對於政黨檔案認定的爭議。

表5

WEKA自動分群結果及其高頻詞彙對照表

群集	分群篇數	前15個高頻詞（依詞頻高低排序）	主題概念
第1群	224	文件、英國、民進黨、解密、機密、證據、蔡英文、立委、香港、圖片、審查、公文、外交部、保密、法國	檔案解密公開
第2群	106	政治、檔案、轉型正義、國家、政府、社會、工作、歷史、臺灣、威權、時期、總統、民主、促轉會、國民黨	轉型正義與政治檔案
第3群	110	歷史、政府、檔案、政治、社會、國家、中國、美國、作為、文化、事件、時期、存在、大學、國際	歷史研究
第4群	62	臺灣、國家、委員會、檔案、檔案管理、樂活、國家檔案、情報、政府、國發會、檔管局、民國、日本、臺北、歷史	檔案推廣應用
第5群	84	臺灣、中國、政府、蔣介石、政權、政治、日本、人民、軍事、組織、工作、事件、委員會、統治、美國	檔案史料稽憑

（三）文本關鍵字及核心議題

2012 ~ 2021年間共586篇檔案議題相關的主題文本，經過斷詞及主題分類，可以歸納出在網路論壇中，社會大眾經常討論的高頻人名、機構團體，以及社會事件如表6。在高頻人名中，「蔣中正」、「蔣經國」、「陳儀」、「彭孟緝」、「張發奎」、「李登輝」、「蔡英文」等人名，若對照原貼文內容，可知多是在討論二二八事件、轉型正義等議題時出現。而「黃國書」、「朴槿惠」、「葉虹靈」等高頻人名，則是因議論政治檔案時被提及，例如「朴槿惠」的閨密門事件，「黃國書」在威權時期的線民身分被起底，「葉虹靈」則是因擔任促轉會代主任委員，對於政治檔案的發言獲得討論聲量。至於高頻詞的機構團體名稱，有「國民黨」、「民進黨」、「中國共產黨」等政黨，也有「檔管局」、「促轉會」、「國史館」等檔案業務機構，皆因討論政治檔案和檔案管理工作時被多次論及。在社會事件方面，有「轉型正義」、「白色恐怖」和「二二八事件」為高頻詞，主要是討論檔案公開還原事件的問題。

表6

檔案相關貼文出現之人名、機構團體以及社會事件高頻詞

詞彙類型	高頻詞（次數）
人名	蔣中正（891次）、蔣經國（387次）、蔡英文（311次）、陳儀（210次）、施明德（180次）、朴槿惠（165次）、李登輝（153次）、彭孟緝（134次）、黃國書（128次）、葉虹靈（88次）、張發奎（83次）
機構團體	國民黨（1189次）、民進黨（360次）、中國共產黨（501次）、檔管局（504次）、促轉會（619次）、國史館（282次）
社會事件	轉型正義（789次）、白色恐怖（361次）、二二八事件（271次）

（四）網路議論檔案主題共現分析

主題共現分析的強度以測量統計分數呈現，在CORPRO中內建包含*Z*值、*T*值、MI（mutual information）值、對數近似值檢定（log likelihood test）等指標，參考闕河嘉與陳光華（2016）採用庫博中文獨立語料庫的經驗，以「對數近似值檢定」為共現分析顯著性與關聯性的方案，採用此統計量作為評判標準，進行共現強度的排序。將前述文本斷詞所得

58,057個詞彙，透過詞頻統計及自動化主題分群，找出文本中的關鍵詞彙與主題概念，接著透過「共現分析」計算與其他詞彙的高共現關係，能使主題中所蘊含的概念更加聚焦。由於高頻詞中的「專有名詞」，帶有較高的專指性，故取專有名詞排序前五的「國民黨」、「蔣介石」、「中國共產黨」、「蔣經國」、「白色恐怖」進行共現分析，比較能聚焦重要概念，並以排序前10個共現詞彙及其對數近似值檢定分數呈現，如表7所示。

在與檔案相關的論壇貼文中，常與「國民黨」一詞共同出現的字詞包含「民進黨」、「立委」、「政府」、「黨國」、「黨史館」、「蔣經國」，映證文本標題及內容，國民黨和民進黨為臺灣兩大政黨，通常會併同出現討論，此外對於國民黨檔案去留問題，也讓黨史館和蔣經國與此產生關連。

與「蔣介石」高共現的詞彙，為「紀念堂」、「轉型」、「文化部」、「李永得」、「銅像」等，多圍繞在「中正紀念堂」存廢的問題上，轉型正義是社會熱中的議題，在轉型正義中不乏有論戰「中正紀念堂」名稱，以及帶有威權象徵的銅像拆除問題。

與「中共」相關的高共現詞彙多屬地理名稱，前幾項為「中國」、「美國」、「蘇聯」、「臺灣」、「北京」，映證文本多為針對國際情勢分析，經常探討的是中、美、蘇三國之間的政治互動。

「白色恐怖」的高共現詞彙包含「受難」、「案件」、「政治」、「轉型正義」、「二二八」、「檔案」、「家屬」，白色恐怖屬於政治迫害，且與二二八與轉型正義議題相關，其共現詞當中有「檔案」與「家屬」兩個詞彙被突顯出來，主要是因為白色恐怖案件進行轉型正義工程，有一部分需要透過檔案證據進行司法平反，然而多數受難者已逝，評判與賠償需要透過家屬申請調閱檔案，還予政治受難者真相。檔案相關貼文內容有相當比例是議論社會時事新聞，討論的議題不免牽涉政治議題，故而呈現在高頻詞與共現詞的分析結果，皆是與政治概念有高度連結的詞彙。

進一步將上述數值轉化為視覺化呈現，透過Gephi軟體繪製成社會網絡關係圖，如圖3所示，將表7的詞彙設置成「節點」（node），有共現關係者則加入「連結」（edge），不同顏色代表的是以上述五個專有名詞為核心的共現詞組，共現的分數強度則由線條粗細呈現。觀察社會網絡關係圖可以看出，雖然是依照不同關鍵詞的共現詞彙進行分組，然而詞組之間仍然有交集的詞彙，例如「蔣經國」本身為一個詞組的中心詞彙，同時也是「國民黨」共現詞組之一，呈現出重要詞彙之間的概念連結關係。

表7

專有名詞前10組共現詞彙及分數

	專有名詞									
	國民黨		蔣介石		中共		蔣經國		白色恐怖	
排序	高共現詞彙	分數	高共現詞彙	分數	高共現詞彙	分數	高共現詞彙	分數	高共現詞彙	分數
1	民進黨	136.5	紀念堂	219.2	中國	14.8	玉照	12.9	時期	86.3
2	立委	78.8	轉型	27.8	美國	66.7	先總統	11.7	受難	75.1
3	黨	74.7	立法院	25.1	蘇聯	51.9	退位	11.0	案件	72.2
4	政府	61.7	兩廳院	23.4	共產黨	41.9	遺像	9.7	政治	62.8
5	政黨	50.4	文化部	21.9	中華	41.5	肖像	9.5	統治	53.0
6	臺灣	49.5	李永得	21.1	臺灣	40.0	國旗	9.4	轉型正義	43.9
7	政治	47.5	園區	21.0	北京	37.8	救國團	9.1	二二八	34.0
8	黨國	46.3	用途	19.3	左派	34.8	蔣公	8.9	國民黨	32.2
9	黨史館	44.8	牌樓	18.8	兩岸	34.8	國父	8.8	檔案	30.6
10	蔣經國	44.7	銅像	14.2	香港	33.2	元首	8.7	家屬	30.6

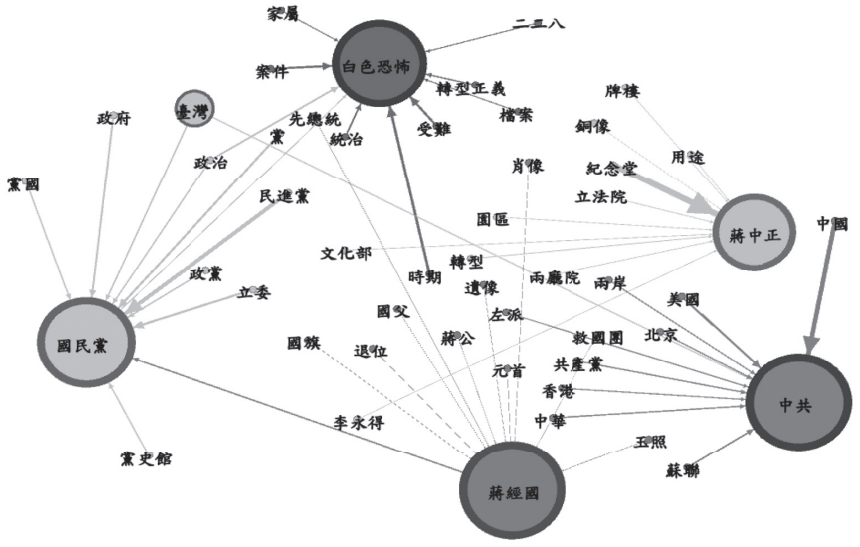

圖3　Gephi繪製的社會網絡關係圖

二、檔案貼文情感分析

（一）文本情感整體分布

586篇文本斷詞的所有詞彙，經與CVAW4.0的5,512個情感詞彙對比後，有3,519個共同出現的情感詞，在文本中共計出現59,502次，將所有情感詞彙及其極性數值，回推各篇文本中計算出現的詞頻與正負數值，計算出586篇文本的情感分數。情感詞彙分數介於±4之間，分數大於0為正面情感，等於0為中性情感，小於0則為負面情感。統計586篇文本所對應的情感傾向，得出正面情感貼文共263篇，占44.88%，而負面情感文本共有323篇，占55.12%，文本總數中是以負面情感文本略高於正面情感文本，如表8所示。

表8

網路論壇檔案文本情感傾向統計表

情感傾向	文本數	百分比（*N* = 586）
正面	263	44.88
負面	323	55.12
總計	586	100

統計586篇文本的情感平均分數為-0.035，可解釋為中立偏負面，而情感分數最高的文本情感數值為1.480，最低的文本情感數值為-1.571。多數文本情感分數介於±0.6中間，以直方圖觀察（圖表數值以四捨五入取至小數點第一位），呈現多數集中中間而向兩側漸次遞減分布的鐘型曲線分布圖，如圖4所示。

（二）情感傾向之時間變化趨勢

如果將情感分析結果加入時間軸，可以觀察出整體網路論壇討論檔案的情感時間變化。在2012 ~ 2018年間，正負面的文本數並無太大的差異，在2014與2018年出現兩次交叉，雖然偶有消長，但正負面情感大致呈現持平狀態，且2014 ~ 2017年之間，正面傾向的文本略多於負面；然而2018年是關鍵年分，負面文本上升的幅度較為劇烈，且遠高於正面文本。2018年5月促進轉型正義委員會掛牌成立，政治檔案議題受到討論，檔案與政治議題連結引發不少負面情感，如圖5所示。

三、檔案核心主題之情感分析

（一）自動分類主題之情感傾向

根據主題分析中統整出的五項檔案主題，分別計算不同主題各篇貼文的情感傾向，其中僅有「檔案推廣應用」為正面情感高於負面情感，

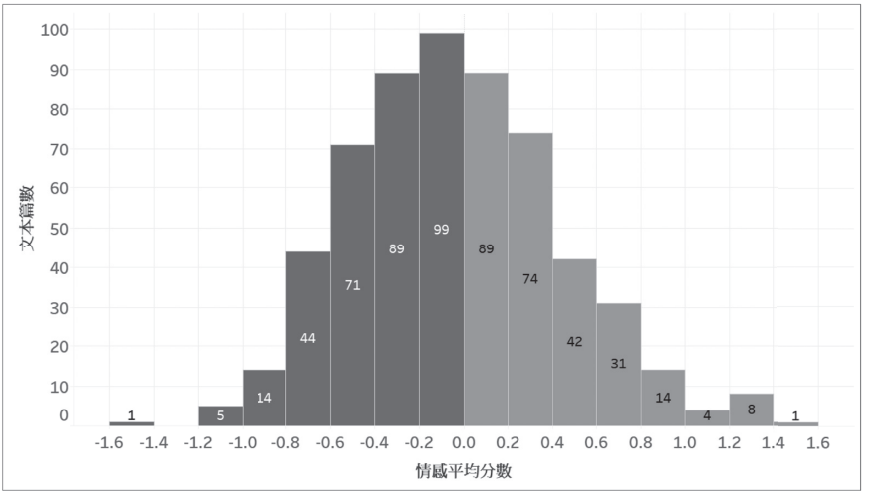

圖4　586篇文本呈現的情感分數分布圖

其餘主題皆是負面多於正面，尤其是「檔案史料稽憑功能」主題，負面聲浪遠高於正面，如圖6所示。

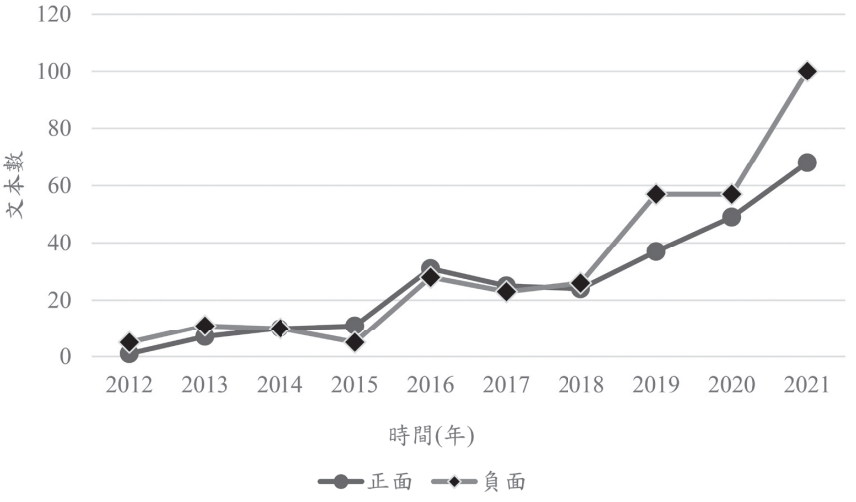

圖5　網路論壇檔案情感傾向趨勢變化圖（2012 ~ 2021年）

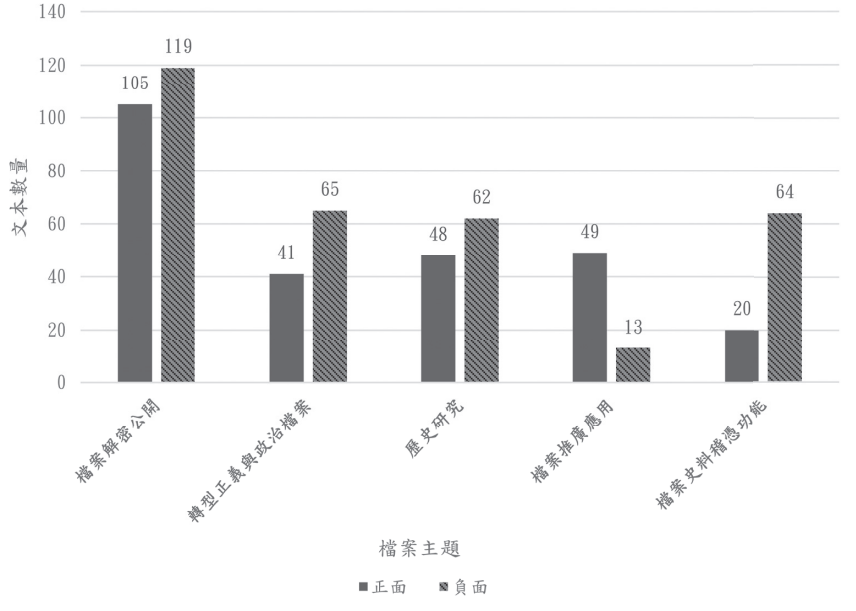

圖6　檔案主題情感統計圖

「檔案推廣應用」主題的正向文本共49篇，其內容多為解讀國家檔案內容的介紹，此類基於檔案內容回顧歷史的貼文，呈現之文本情感傾向於正面。檢視「檔案史料稽憑功能」的64篇負面文本中，有39篇來自於馬非白在想想論壇的【被遺忘的歷史】系列文章，該系列多講述1945年國民政府遷臺後，對於臺灣政治、經濟、言論的箝制，將過去不被公開卻影響社會發展的重要事件，佐以檔案和照片訴諸輿論，其內容情感分析偏向負面。

（二）檔案人物及事件情感分析

進一步探討在論壇討論檔案出現的高頻人名、機構團體以及社會事件的情感傾向（參閱前述表6），瞭解網路論壇貼文對於此類人物、機構或社會事件的情感認知。

1. 檔案關聯人物情感分析

在檔案貼文中經常被提及的人名，多是呈現負面情感多於正面，如圖7所示，其中以蔣中正相關文本數量最多，但其負面文本數量更是獨占鰲頭，貼文討論的是揭露在白色恐怖時期所扮演的「審判者」角色，常出現「匪諜」、「叛亂」、「逮捕」、「判刑」等負面字眼；陳儀、彭孟緝兩人因與二二八事件關連，多半也呈現負面情感，此外，施明德也是負面情感較多，議論的內容多是討論其欲透過申請檔案調閱，找出美麗島事件前後被特務、線民監控的證據，文本中常見負面詞彙，致使多數文本呈現負面情感。

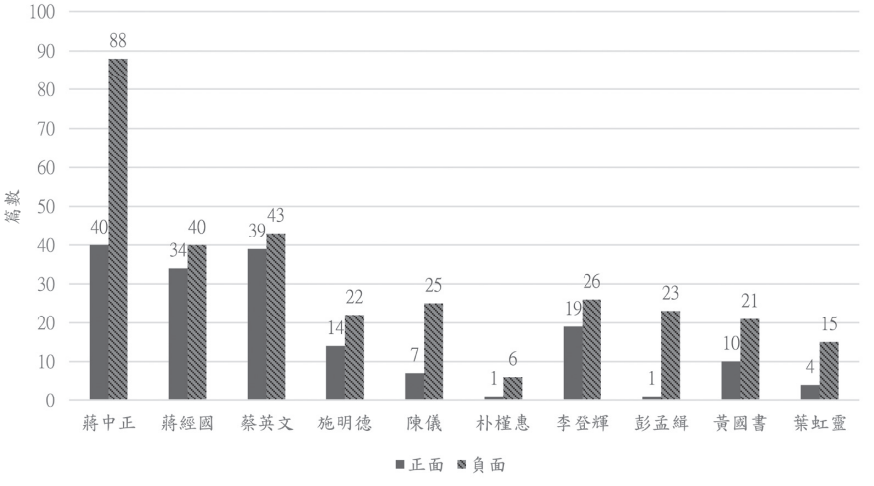

圖7　對於檔案相關人物的情感分析統計圖

2. 機關團體高頻詞的情感分析

檔案貼文所出現的機關團體，除了檔案典藏機構的「檔管局」與「國史館」為正面情感高於負面外，其餘皆為負面情感較多，如圖8所示。包含「國民黨」、「民進黨」、「中國共產黨」、「促轉會」都呈現負面情感居多，尤其促轉會有較多的負面情感，文本呈現的負面論述並不一定是針對機關本身，而是因為促轉會常與轉型正義同時討論，致使文本內容負面詞彙較多。

3. 社會事件高頻詞的情感分析

在與檔案相關的社會事件中，轉型正義、白色恐怖及二二八事件都呈現出負面情感多於正面情感，如圖9所示，且負面文本數量高出正面一倍之多，此三個社會事件皆與臺灣經歷威權統治時期的事件相關，無論是二二八事件的武力鎮壓，或是白色恐怖的高壓統治，甚至是為平反威權傷害的轉型正義，因為涉及對於不公義的討論，致使貼文內容呈現負面高於正面。

文本文字呈現的情感傾向，影響閱讀者對於議題的情感認知，本研究使用CVAW4.0情感詞典，進行情感詞彙的比對後，計算出各文本情感分數。在2012 ~ 2021年間，網路論壇共586篇與檔案相關的貼文，雖然有55%的貼文是負面情感，但進一步分析檔案主題以及檔案人物、機構、社會事件等高頻詞所在之文本呈現的情感差異，發現如果涉及檔案推廣應用是正面情感，但涉及批判歷史事件或是轉型正義議題，將檔案資料

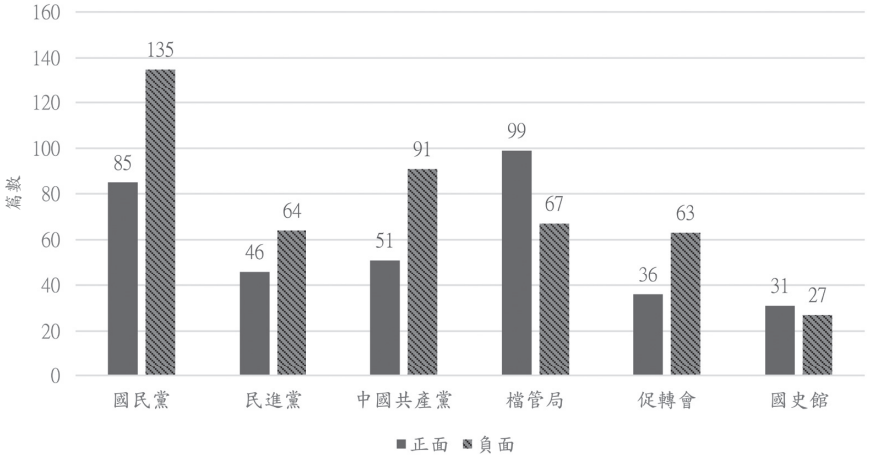

圖8　對於檔案機關團體的情感分析統計圖

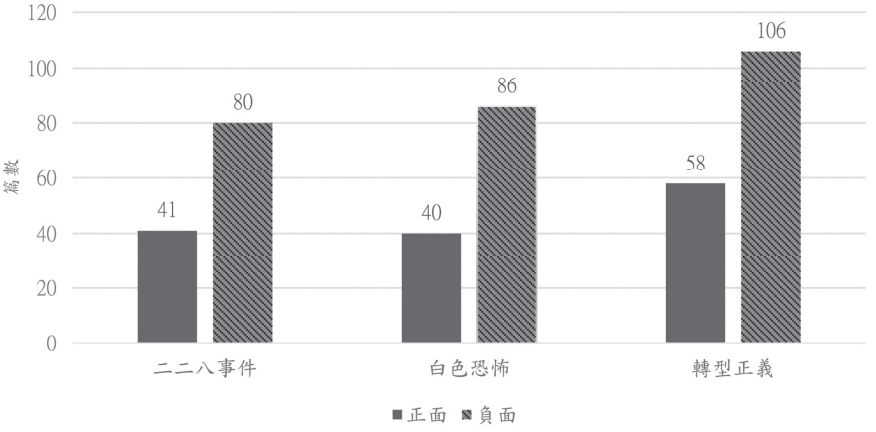

圖9　對於檔案相關事件的情感分析統計圖

當成史料憑證，批判是非、論述功過時，往往帶有批評色彩，致使文本呈現負面情感。

伍、結論與建議

基於前述網路論壇文本議論主題及情感分析結果，本研究歸納結論如下。

一、利用多種數位工具與分析方法，能全面性觀察社會大眾對於檔案議論的主題以及引發的情感傾向。

本研究使用多元的數位工具進行資料分析，以WEKA自動分群判讀檔案內容主題，利用語料庫分析挖掘詞彙概念的共現關係，以社會網絡分析視覺化呈現詞彙之間的關聯，以情感分析判斷群眾對於文本可能產生的情感認知。透過不同的資料分析方法，完成網路論壇檔案相關貼文的解讀，使得586篇檔案貼文內容變得更加立體，有助於檔案機構以及檔案研究者瞭解社會大眾對於檔案事業發展的評價，可提供我國檔案工作規劃及發展方向的參考。

二、分析網路論壇之檔案貼文詞彙及其共現詞，可知社會大眾關注議論的檔案主題多與時事及政治話題相關。

根據整體高頻詞萃取出的檔案關鍵字詞，如：「蔣中正」、「蔣經國」、「蔡英文」、「黃國書」都是政治人物，在團體機構中，除了

有檔案主管機關「檔管局」、「國史館」外，還有「國民黨」、「民進黨」、「中國共產黨」及「促轉會」；剖析檔案關鍵字的高共現詞彙，同樣會發現論壇意見內容呈現的檔案議題與政治事件環環相扣，例如：轉型正義、朴槿惠閨密門、黃國書線民案、蔡英文論文案。其中促轉會與政治檔案更是關注的熱點主題，議論的焦點多是針對政治檔案不夠公開透明、質疑公部門行政效率，並期許能解密檔案和開放應用等論點，意味著檔案資料在網路造成討論聲浪的議題，通常是反映當下受到輿論關注的社會和政治話題，突顯檔案具備與社會議題脈動連結的社會記憶價值。

三、網路議論檔案內容的情感整體為中立偏負面，若論及政治人物與機構團體多是呈現負面情感，若為檔案工作則呈現正面，正負面情感的時間轉折點在2018年開始轉為負面。

2012 ~ 2021年586篇議論檔案的網路文本，數值分析為55%為負面情感，負面略高於正面。若依照主題區分，「檔案解密公開」、「轉型正義與政治檔案」、「歷史研究」、「檔案稽憑功能」等主題皆為負面情感文本較多，而「檔案推廣應用」則呈現大量的正面情感。若根據文本內容常見的人名、機構以及社會事件區別情感傾向，發現若論及政治人物多屬負面大於正面，若論及檔案典藏機構（檔案管理局、國史館等），則呈現正面情感較強，若為政黨或是促轉會等政治團體是負面情感較強。可知社會大眾對於涉及政治相關之人物與團體，傾向負面情感，若是屬於檔案推廣活動則為正面情感。若以時間趨勢觀察，大抵從2018年始，正負面情感文本出現急遽變化，負面情感文本快速增加，顯示促轉會成立後，有大量輿論新聞和政治檔案相關，負面輿論容易造成社會大眾對於檔案社會功能的誤解，此一狀況應是檔案主管機關未來需要設法導正的社會形象。

根據前述研究結果，本研究提出對於檔案工作未來發展建議如下。

一、檢討檔案解降密作業並加速檔案開放應用

綜合網路論壇的檔案貼文主題分析及情感分析成果，臺灣近10年社會大眾關注的檔案議題，數量最多的是「檔案解密公開」，隨著網路資訊透明化、公民意識的崛起，民眾對於「知」的權利意識抬頭，民眾對於檔案解密公開不僅有高度期待，也帶著監督和批判的角度。因此，建議檔案典藏機構對於持有的機密檔案能遵循法治，屆臨解密年限應該履行公開程序，以促進檔案的開放應用。

二、檔案典藏機構可運用網路論壇主動推廣檔案資訊

前述研究分析來自「檔案樂活情報」轉載的貼文，呈現出檔案開放應用的正面情感，有助於社會大眾在負面的政黨攻訐的政治論戰中，看到較為持平的檔案工作介紹，也能一窺檔案內容的歷史價值，檔案的功能不在於提供政治攻防的工具，檔案蘊含的內容知識有助於歷史存真並喚起社會共同記憶。檔案機構可善用網路論壇主動發文，宣揚檔案價值。建議檔案典藏機構可以將檔案工作以及配合時事的重要檔案內容，轉載至擁有相當聲量的論壇平臺，進而可豐富論述的議題，例如：檔案的修護、個人檔案保存技術，應為一般民眾有興趣且有意義之議題，在網路論壇中積極發聲，能有助於社會大眾對於檔案的正確認識，並可平衡不實和偏頗的言論，減少民眾對於檔案隱諱和不正確的觀念。

參考文獻

王貿（2020）。公務人員關注議題之文字探勘：以PTT公職板為例。*調查研究——方法與應用*，*45*，119-154。【Wang, M. (2020). A computational text analysis on the core issues for public servants: Evidence from the PTT. *Survey Research—Method and Application*, *45*, 119-154. (in Chinese)】

王毓莉（2007）。網路論壇與國家機器的碰撞：從三個新聞事件看大陸網路論壇對公共性的實踐。*新聞學研究*，*92*，37-95。doi:10.30386/MCR.200707_(92).0002【Wang, Y.-L. (2007). The collision of the Internet forum and the state apparatus: The “publicity” practice of Mainland China’s Internet forum from three news event. *Mass Communication Research*, *92*, 37-95. doi:10.30386/MCR.200707_(92).0002 (in Chinese)】

李建華、劉功申、林祥（2017）。情感傾向性分析及應用研究綜述。*信息安全學報*，*2*(2)，48-62。doi:10.19363/j.cnki.cn10-1380/tn.2017.04.005【Li, J., Liu, G., & Lin, X. (2017). Survey on sentiment orientation analysis and its applications. *Journal of Cyber Security*, *2*(2), 48-62. doi:10.19363/j.cnki.cn10-1380/tn.2017.04.005 (in Chinese)】

周立柱、賀宇凱、王建勇（2008）。情感分析研究綜述。*計算機應用*，*28*(11)，2725-2728。【Zhou, L.-Z., He, Y.-K., & Wang, J.-Y. (2008). Survey on research of sentiment analysis. *Journal of Computer*

Applications, *28*(11), 2725-2728. (in Chinese)】

邵軒磊、曾元顯（2018）。文字探勘技術輔助主題分析——以「中國大陸研究」期刊為例。問題與研究，*57*(1)，29-62。doi:10.30390/ISC.201803_57(1).0002【Shao, H., & Tseng, Y.-H. (2018). Text mining-based topic analysis: A case study of the journal "Mainland China Studies". *Wenti Yu Yanjiu*, *57*(1)，29-62. doi:10.30390/ISC.201803_57(1).0002 (in Chinese)】

郭文平（2020）。語料庫輔助的媒體論述分析：以臺灣平面媒體中國夢報導為語料的實證研究。資訊社會研究，*38*，51-92。doi:10.29843/JCCIS.202001_(38).0005【Kuo, W.-P. (2020). Corpus-assisted media discourse analysis: Methods, theoretical framework and case study of the China dream corpora. *The Journal of Information Society*, *38*, 51-92. doi:10.29843/JCCIS.202001_(38).0005 (in Chinese)】

陳育正、孫懋嘉、顧志文、林立偉（2020）。由社群媒體的觀點論習近平主政後對臺統戰策略對我國民眾的影響：以對臺31項措施為例。中國大陸研究，*63*(2)，111-150。doi:10.30389/MCS.202006_63(2).0004【Chen, Y.-C., Sun, M.-C., Ku, C.-W., & Lin, L.-W. (2020). Examining the impact of the 31 Taiwan-related measures on public opinion in Taiwan: A corpus-based analysis. *Mainland China Studies*, *63*(2), 111-150. doi:10.30389/MCS.202006_63(2).0004 (in Chinese)】

陳迪、程朗、王志鋒、熊錦鵬、張玉茹、梨高贊（2021）。論壇情感挖掘研究綜述：現狀、挑戰與趨勢。計算機工程與應用，*57*(17)，17-28。doi:10.3778/j.issn.1002-8331.2012-0519【Chen, D., Cheng, L., Wang, Z., Xiong, J., Zhang, Y., & Li, G. (2021). Sentiment analysis for web forum: Status, challenges and trends. *Computer Engineering and Applications*, *57*(17), 17-28. doi:10.3778/j.issn.1002-8331.2012-0519 (in Chinese)】

張奕萱、林巧敏（2022）。運用數位人文工具探討檔案時事議論主題及其情感分類之實作。圖資與檔案學刊，*14*(2)，164-190。【Zhang, Y.-X., & Lin, C.-M. (2022). Using digital humanities tools to explore the implementation of topic and sentiment analysis in archival affairs commentary. *Journal of InfoLib & Archives*, *14*(2), 164-190. (in Chinese)】

陸澤凱、謝穎（2021）。微博中的「中美外交風波」輿情文本研究

——基於R語言的詞向量情感分析。傳媒觀察，*2021*(2)，54-61。doi:10.19480/j.cnki.cmgc.2021.02.007【Lu, Z., & Xie, Y. (2021). Research on public opinion texts of “Sino-U.S. diplomatic disturbance” in Weibo. *Media Observer*, *2021*(2)，54-61. doi:10.19480/j.cnki.cmgc.2021.02.007 (in Chinese)】

褚乃慈（2020）。臺灣獨立樂團歌詞與樂迷留言內容之情緒分析（未出版之碩士論文）。世新大學資訊傳播學系，臺北市。【Chu, N.-T. (2020). *Sentiment analysis of content in lyrics and comments of Taiwan independent bands* (Unpublished master’s thesis). Shih Hsin University, Taipei. (in Chinese)】

劉勇、杜一（2017）。網絡數據可視化與分析利器：*Gephi*中文教程。北京：電子工業出版社。【Liu, I., & Du, Y. (2017). *Wang luo shu ju ke shi hua yu fen xi li qi: Gephi zhong wen jiao cheng*. Beijing: Publishing House of Electronics Industry. (in Chinese)】

劉嘉薇（2017）。網路統獨的聲量研究：大數據的分析。政治科學論叢，*71*，113-165。doi:10.6166/TJPS.71(113-166)【Liu, J.-W. (2017). Public opinion of unification-independence on the web: A big-data analysis. *Taiwanese Journal of Political Science*, *71*, 113-165. doi:10.6166/TJPS.71(113-166) (in Chinese)】

闕河嘉、陳光華（2016）。庫博中文獨立語料庫分析工具之開發與應用。在項潔（編），數位人文：在過去、現在和未來之間（頁285-313）。臺北市：國立臺灣大學出版中心。doi:10.13140/RG.2.2.36253.97768【Chueh, H.-C., & Chen, K.-H. (2016). CORPRO: A Chinese language corpus tool and a case study of media representation of organic agriculture. In J. Hsiang (Ed.), *Digital humanities: Between past, present, and future* (pp. 285-313). Taipei: National Taiwan University Press. doi:10.13140/RG.2.2.36253.97768 (in Chinese)】

Al-Daihani, S. M., & Abrahams, A. (2018). Analysis of academic libraries’ Facebook posts: Text and data analytics. *The Journal of Academic Librarianship*, *44*(2), 216-225. doi:10.1016/j.acalib.2018.02.004

Blake, C. (2011). Text mining. *Annual Review of Information Science and Technology*, *45*(1), 121-155. doi:10.1002/aris.2011.1440450110

Dave, K., Lawrence, S., & Pennock, D. M. (2003). Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In B. White & G. Hencsey (Eds.), *Proceedings of the 12th International*

Conference on World Wide Web (pp. 519-528). New York, NY: Association for Computing Machinery. doi:10.1145/775152.775226

Dwianto, R., Nurmandi, A., & Salahudin, S. (2021). The sentiments analysis of Donald Trump and Jokowi's Twitters on COVID-19 policy dissemination. *Webology*, *18*(1), 389-405. doi:10.14704/WEB/V18I1/WEB18096

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge, UK: Cambridge University Press.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267-297. doi:10.1093/pan/mps028

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, *11*(1), 10-18. doi:10.1145/1656274.1656278

Huang, Y.-C., & Lin, C.-M. (2020, December). *The data mining and sentiment analysis of the archival news*. Paper presented at the 11th International Conference of Digital Achieves and Digital Humanities (DADH 2020), Taipei, Taiwan.

Lamba, M., & Madhusudhan, M. (2018). Application of sentiment analysis in libraries to provide temporal information service: A case study on various facets of productivity. *Social Network Analysis and Mining*, *8*(1), 63. doi:10.1007/s13278-018-0541-y

Lee, G. T., Kim, C. O., & Song, M. (2020). Semisupervised sentiment analysis method for online text reviews. *Journal of Information Science*, *47*(3), 387-403. doi:10.1177/0165551520910032

Lopatovska, I., & Arapakis, I. (2011). Theories, methods and current research on emotions in library and information science, information retrieval and human–computer interaction. *Information Processing & Management*, *47*(4), 575-592. doi:10.1016/j.ipm.2010.09.001

Na, J.-C., Thet, T. T., Nasution, A. H., & Hassan, F. M. (2011). A sentiment-based digital library of movie review documents using Fedora. *Canadian Journal of Information and Library Science*, *35*(3), 307-337. doi:10.1353/ils.2011.0018

Palomino, M. A., & Aider, F. (2022). Evaluating the effectiveness of text pre-processing in sentiment analysis. *Applied Sciences*, *12*(17), 8765.

doi:10.3390/app12178765

Papachristopoulos, L., & Tsakonas, G. (2020). More than a feeling: Insights and information from a sentiment analysis study. *Liber Quarterly: The Journal of the Association of European Research Libraries*, *30*(1), 1-12. doi:10.18352/lq.10325

Park, S., Bier, L. M., & Park, H. W. (2021). The effects of infotainment on public reaction to North Korea using hybrid text mining: Content analysis, machine learning-based sentiment analysis, and co-word analysis. *Profesional De La Información*, *30*(3), e300306. doi:10.3145/epi.2021.may.06

Patra, S. K. (2019). How Indian libraries tweet? Word frequency and sentiment analysis of library tweets. *Annals of Library and Information Studies*, *66*(4), 131-139.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825-2830. doi:10.48550/arXiv.1201.0490

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161-1178. doi:10.1037/h0077714

Sabatovych, I. (2019). Do social media create revolutions? Using Twitter sentiment analysis for predicting the Maidan Revolution in Ukraine. *Global Media and Communication*, *15*(3), 275-283. doi:10.1177/1742766519872780

Stewart, B., & Walker, J. (2018). Build it and they will come? Patron engagement via Twitter at Historically Black College and University libraries. *The Journal of Academic Librarianship*, *44*(1), 118-124. doi:10.1016/j.acalib.2017.09.016

Tang, H., Tan, S., & Cheng, X. (2009). A survey on sentiment detection of review. *Expert Systems With Applications*, *36*(7), 10760-10773. doi:10.1016/j.eswa.2009.02.063

Thelwall, M. (2017). Sentiment analysis. In L. Sloan & A. Quan-Haase (Eds.), *The SAGE handbook of social media research methods* (pp. 545-556). Thousand Oaks, CA: SAGE.

Yaqub, U., Chun, S. A., Atluri, V., & Vaidya, J. (2021). Analyzing social media messages of public sector organizations utilizing sentiment analysis and topic modeling. *Information Polity*, *26*(4), 375-390. doi:10.3233/IP-

210321

Yu, L.-C., Lee, L.-H., Hao, S., Wang, J., He, Y., Hu, J., ... Zhang, X. (2016). Building Chinese affective resources in valence-arousal dimensions. In K. Knight, A. Nenkova, & O. Rambow (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 540-545). San Diego, CA: Association for Computational Linguistics. doi:10.18653/v1/N16-1066

Public Opinion Mining and Sentiment Analysis for Archive Issues on Internet Forums in Taiwan

Chiao-Min Lin
Professor
Graduate Institute of Library, Information and Archival Studies
National Chengchi University

Yu-Hsien Li
Graduate Student
Graduate Institute of Library, Information and Archival Studies
National Chengchi University

Introduction

Public opinion is often the reference basis for the services provided by public authorities. In recent years, Taiwan has received media attention due to social issues involving the historical facts of document restoration, which has affected the public's understanding of archives. Especially in the era of rapid network communication, public perceptions and evaluations of archives are often affected by discussions created in the Internet community. If we can analyze the contents of community discussions, as well as the emotions regarding such content, it will help authorities related to archives understand the issues of social concern and make better public decisions. The authorities can even adjust the perspective of their publicity propaganda to promote increased attention to and a better understanding of the archives for the public.

Methodology

This study used keywords search, date sorting, and screening to collect data from four online forums (Thinking Taiwan, Storm Media, News Lens, and UDN Opinion) and posts related to archive issues. CKIP Tagger was used for word segmentation and speech tagging, and an authority control file

was established to feed back the clean text. Then, digital tools were used for analysis: (1) Theme analysis: WEKA was used to conduct automatic theme analysis, determine the classification of archive themes, and analyze the discussion waves of various archive themes in the network forum; (2) Corpus analysis: the CORPRO corpus tool was used to conduct co-occurrence analysis to identify the relationship between high-frequency words; and (3) Emotional analysis: the CVAW4.0 Chinese dimensional emotional dictionary was used to compare the posted words with the emotional dictionary, calculate the emotional score of each text, and present the emotional tendency of the text.

Results

Analysis of Archive Theme

WEKA was used to conduct automatic theme grouping, and the five themes generated were: archive decryption and disclosure, transitional justice and political archives, historical research, archive promotion and access, and archive historical data evidences. The archive themes with the largest numbers of posts by the public were "archive decryption and disclosure," followed by "transitional justice and political archives," most of which were public opinions triggered by social events or political issues.

Analysis of the Co-Occurrence of Discussion Themes

A considerable proportion of the posts were about social and current affairs, and the themes discussed inevitably involved political issues. According to the analysis results of high-frequency words and co-occurrence words, words highly connected with political concepts were presented. For example, the high co-occurrence words with "white terror" were "suffering," "case," "politics," "transitional justice," "February 28 Incident," "archives," and "family members," as white terror cases must be judicially rehabilitated through archival evidence to achieve transitional justice.

Emotion Distribution of Text

By comparing all segmented words with the emotion dictionary, the word frequency and positive and negative values of each text were deduced, and 263

positive emotion posts were obtained, accounting for 44.88%, while there were 323 negative emotion posts, accounting for 55.12%. If such emotion analysis results were added into the timeline, the emotional changes in online public opinion could be observed. Between 2012 and 2018, there was no significant difference between the number of positive and negative texts; however, since 2018, the growth rate of negative texts has been higher than that of positive texts.

Emotional Tendency of Archive Theme

This study calculated the emotional tendency of each theme according to the five archive themes. Only the "archive promotion and access" theme showed greater positive emotions than negative emotions, while the remaining themes were more negative than positive emotions. In particular, the negative voice raised in "archive historical data evidence" was far greater than the positive. In the social events related to archives, such as transitional justice, white terror, and the February 28 Incident, negative emotions are more prevalent than positive emotions, which are related to various events when Taiwan experienced authoritarian rule. Due to the discussion about unfair authoritarianism and high-pressure rules, the contents of the posts are more negative than positive.

Conclusions

Analyses of the archive-related posts in online forum themes show that the public is concerned about issues related to current affairs and political topics, meaning the archive issues presented in the forum's opinions are closely linked with political events. Among them, the Transitional Justice Commission and political archives are the hottest topics of concern. The focus of the discussion is primarily on the issue that political archives are not open and transparent enough, which calls into question the administrative efficiency of the public sector, and the public expectation of decrypting archives and disclosing them for public access. All such concerns imply that the issues causing archive data discussion waves on the Internet are usually the social and political topics currently of concern to public opinion, which highlights that archives have social memory value linked to the pulse of social issues.

Overall, while the emotions of online discussions about archive contents are neutral to slightly negative, when politicians and institutional groups are the concern of online discussions, the emotions tend to be negative. However, when the online discussion is about archive works, the emotions are positive. The turning point of positive and negative emotions was in 2018, when public opinion noticeably turned negative, which coincided with the establishment of the Transitional Justice Commission. This event initiated many discussions related to political archives, leading to public misunderstanding of the archives' functions.

In the past decade, Taiwan's public has been increasingly concerned about archive issues, and the articles posted are mostly about "archive decryption and disclosure." With the transparency of information and the rise of civic awareness, the public's awareness of the right to know has increased, meaning the public has high expectations for archive decryption and disclosure, as well as a perspective on supervision and criticism. Therefore, it is suggested that the archive institutions should follow the rule of law and perform open procedures for the confidential archives they hold in order to promote the open access of archives.