災難事件中社群媒體訊息之自動分類設計

施旭峰　政治大學資訊科學系
李蔡彥　政治大學資訊科學系
鄭宇君　玄奘大學大眾傳播學系
陳百齡　政治大學傳播學院

摘要

近年來，當重大災難發生時，人們經常透過網路通訊工具傳遞災情或求救訊息，大量的資訊人力已無法負荷，如何在第一時間進行有效分類，以即時傳遞到適當的救災、協尋或資源調度單位，一直是救援單位的重要課題。

本研究以台灣莫拉克風災期間的五個災難頻道的資料集為分析對象，包括地方救災中心報案紀錄、災情網站貼文、Twitter 等文字資料。經過文字處理與專家分類後，透過詞頻分布、分類結構組成、詞彙共現網絡等方式，探討不同頻道資料集之異同。進一步使用空間向量模型與機器學習的方法，建立社群媒體災難資料的自動化分類器。本研究的歸納與所發展出來的分類方式與資訊探索技術，將可用於開發更有效率的社群感知器。

關鍵詞：災難傳播、社群媒體、鉅量資料、機器學習、社群感知

Automated Classification Design for Social Media Information in a Disaster Event

Abstract

In recent years, when disaster events occur, people often transfer information or distress messages through communication tools. As huge amounts of disaster information flows in, processing the data without the assistance of computational technologies becomes an increasingly challenging task. Therefore, understanding how to effectively classify information from social media, provide reliable information to disaster reaction centers, and assist policy decision-making is an important topic of discussion.

In this study, we use the data collected during typhoon Morakot from five different channels, including the records of local disaster relief centers, the postings on disaster website, and tweets. After word processing and content classification by experts, we observe the difference between these datasets on the frequency distribution, classification structures, and word co-occurrence network. We further use the vector space model and machine learning method to train the classification model of social media information. We believe the techniques developed and results of the analysis can be used to design more efficient and accurate social sensors in the future.

Keywords: Disaster communication, Social media, Big data, Machine learning, Social sensors

壹、前言

災難是一種自然發生或是人為帶來的危害。近年來，各式重大災難在全球不斷發生，1999 年台灣中部埔里的 921 大地震；2001 年發生於美國的 911 恐怖攻擊事件；2003 年全球性的 SARS 疫情；2004 年南亞大海嘯、禽流感疫情爆發， 2009 年全球 H1N1 新流感疫情、台灣莫拉克颱風造成的八八水災；2010 墨西哥漏油事件、俄羅斯森林大火、海地大地震；2011 年日本東北近海 311 大海嘯事件等等。

上述這些人為或自然災難的災情規模與損害都十分龐大，加上全球環境變遷，新型態的災難可能不斷出現，因此災難發生時的資訊傳遞議題逐漸受到研究者重視。特別是現今網際網路技術的突發猛進和資訊基礎建設的高密集性，人們利用網際網路傳遞資訊的方式已漸多於信件、電話等傳統資訊傳遞方式，這使得網際在近年災難中扮演著訊息傳遞的重要角色。

網際網路在災難發生的時候，能夠以最快速度分享災情的文字訊息、照片或影片，不會因部分地區硬體建設毀損造成整體無法使用，亦無時間上的使用限制，隨時都能進行一對一、一對多、多對一或多對多的溝通。加上近年各種網站系統的成熟和社群網絡的高使用率，個人化媒體發展亦漸趨完備，網際網路已成為災難來臨的一個重要資訊傳遞途徑和媒介。

在災難的危機情境下，人們透過資訊科技協力進行緊急應變，使用網路上各種新興科技頻道，從事各種社會活動，包括：搜尋親友資訊、調度救災資源、管理頻道言論等。然而，這些新興媒體上的資料收集和分析方式和過去大不相同，存在於網路上的資料龐大且複雜，無法單純透過人力來擷取蒐集、過濾、管理、收藏及進一步的分析處理，每一個環節對於研究者來說都是極大的挑戰，透過鉅量資料（Bid Data）分析取徑，改變資料收集與分析處理的流程，有助於解決人力無法進行的大量資料分析(鄭宇君，2014)。

有鑑於人力能夠處理訊息之數量的限制性，本研究透過觀察 2009 年莫拉克風災期間於網路和救災應變中心收集的資料，試著建立不同來源災難訊息的處理流程與分析機制，包括臨時災情網站的貼文留言、Twitter、地方救災中心的報案記錄，針對各種資料集內容抽取文字特徵資訊，以機器學習的方式建立各項分類模組，探索自動化分類災難資訊的可能性，協助進行大量資訊的過濾和處理藉此輔助人力的不足，以瞭解大規模災難事件中，不同類型災難資訊所呈現的公眾需求與意見。

貳、文獻回顧

隨著網路內容供應平台及行動通訊設備的普及，災難資訊提供的角色已不僅是傳統專業媒體提供者，更加入廣大的網路公眾媒體。現今網路基礎建設普及，而公眾媒體內容平台眾多，所以資訊內容與流通的速度迅速增加，傳統廣播式媒介已經無法負荷，訊息數量已經遠超過媒介載具所能負載的頻寬。

一、 災難發生時資訊傳遞問題

災難發生後社會內部通常會發生重大失序，人們需要相關資訊以確認自己或是親友的安危，此時卻也是資訊最匱乏的狀態。由於能源中斷、設施毀損或指揮結構失靈等因素，使得相關訊息無法正常的被傳送，人們因而設法透過其他管道來發佈訊息，訊息經過重複的傳遞在短時間內造成龐大的訊息數量，使得原本負載過重的系統因此而癱瘓

(Quarantelli, 1998)。

災難時期訊息的瞬間巨量成長、資訊匱乏和資訊過載，讓平時主要的資訊流通傳播媒介面臨極大的挑戰，當巨量的災情訊息瞬間湧入救災單位、新聞機構和網路新媒體平台，造成人力一時無法處理的資訊數量，形成另一種這使得災難資訊的危機議題，因此如何有效在短時間內解決災難訊息的瞬間巨量問題便成為傳播與資訊專家希望協力共同解決的問題。

二、 災難期間的傳播活動

具體而言，人們在災難期間的傳播活動可以分為四種類型(陳百齡，鄭宇君, 2011)：

（一） **資訊蒐集和傳播**：人們透過媒體尋求與災區相關的親友近況，特別在通訊管道中斷之際，行動電話和網際網路等個人媒體發揮了聯繫作用。

（二） **物資徵集和流向**：人們利用新媒體向社會大眾徵集物資，或透過線上討論平台調度救災物資流向和數量，確保物資能用在正確的地方，不致於產生救災資源的浪費與落差。

（三） **組織人力和任務派遣**：重大災難發生的時候，由於需要大量人力參與救災或復原，往往會產生許多臨時性的新團體或由組織參與救災行動，這些因災難聚集的龐大人力透過媒體協調和指揮救災志工，進行策略編組和調度派遣。

（四） **抒發心情與表達支持**：在災難危機事件中，人們需要新媒體平台抒發慰問之情或表達救災看法和見解，以達到情感撫慰的目的，並消弭因災難而起的焦慮或恐慌。

三、 網際網路在災難中的角色

近年來許多重大災難事件，網際網路都扮演過去傳統媒體沒有的功能，人們利用各種新興網路技術的新媒體頻道來傳遞龐大災難資訊。在顧佳欣(2009)指出，網際網路本身在莫拉克八八風災中扮演資訊傳達、資源募集調配的重要角色。如何以資訊傳播科技使資訊暢通，是災害發生時的災訊流通的關鍵，網際網路的「互動性」和「資訊空間」可以讓偵測環境的功能更為接近民眾的個人資訊需求(孫式文, 2000)。線上社群網路和社交媒體的特點之一是他們對於信息傳播的潛力，他們在訊息擴散的技術創新，已被有經驗的社會學家研究多年(Rogers, 1995)。

網路現已成為平民化的傳播工具，任何組織不需要昂貴器材和專業製作團隊，就可以上網發布訊息及傳遞聲音影像。透過網路，受到災情影響的人們，也可以簡單地將他們的經驗與需要放到網站上和其他災民分享想法。過去幾年中，社群訊息網路工具（如：Twitter）已在災難時期被用來作為一種常見的溝通工具，這類工具可以跨越國家、時區和文化讓人們分享訊息和知識，參與者可以找尋資料和驗證事件資訊，分享災區和失蹤人口等有關的詳細資訊(Potts, 2009)。

四、 透過機器學習方式進行文字訊息分類

從另一個面向來看，透過網路傳遞多種來源的資訊，固然有助人們進行資料收集，但巨大的文字數量使得人力很難在短時間內有效的解析內容。這個議題引起自然語言處

理（Natural Language Processing）和機器學習（Machine Learning）領域的研究人員注意，過去從簡短文字訊息學習分類的問題，有些研究是利用與比較各種機器學習的演算法（像是 Naive Bayes、SVM、Logistic Regression 和 Decision Trees），來解決識別簡短訊息的垃圾訊息問題(Cormack, Hidalgo, & Sanz, 2007)，或是針對線上問題的對話內容（如：Yahoo Answers 或 Google Answers）進行分類比對(Gupta & Ratinov, 2008)。另有研究者則是利用部落格的文本內容，自動抽取關鍵字和階層式分類改善文章註解的方式(Brooks & Montanez, 2006)，或是透過對醫療方面的文字訊息分類，協助非洲偏遠地區的醫療問題(Munro & Manning, 2010)。

在這些過去的訊息分類研究中，雖有針對簡訊或網路訊息透過機器進行分類，但缺乏災難事件下針對未翻譯的原生語言，同時具有跨不同性質資料集比較的分類經驗。因此，本研究之重要價值在於針對單一事件—台灣莫拉克颱風八八水災，收集了不同類型頻道來源共五個資料集的資訊內容，進行跨性質的資料集分析與訊息分類實驗。

參、系統架構與實作

本研究中所使用的五個資料集，除了二個來自緊急救難單位所提供的電話報案記錄，大部分資料收集自網路的不同型別頻道。相較於緊急救難單位的書寫是由少數的專業人員執筆，因此書寫方式與內容格式較為一致，而網路訊息則是來自多元管道與各類型群眾參與書寫，並非所有使用者都接受過專業的訓練，使得收集得到的資訊存在許多干擾因素。因此，在進行文本分析前，本研究設計一些格式統整以及去除干擾因素的方法，進行資料內容的前處理程序。

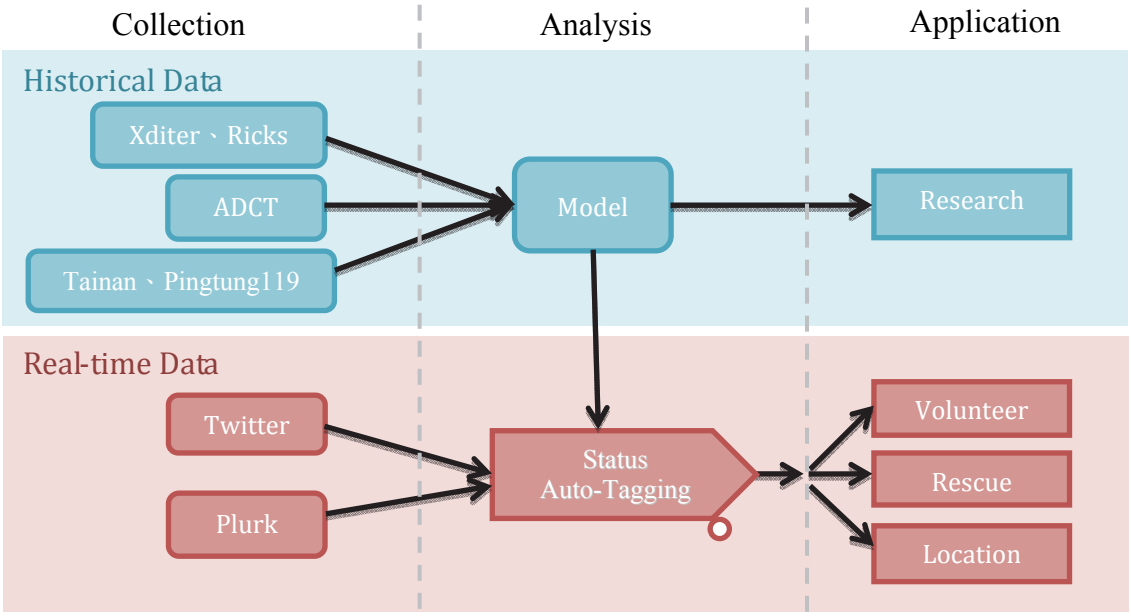


圖 1：自動分類概觀

在資料清理與格式統整之後，研究者以機器學習的方式訓練訊息自動分類器，透過所收集到的五個資料庫內容，初步經過資料前處理、文章特徵抽取，機器學習建立分類模組、驗證分類結果，比較各個資料庫的異同及評估分類器的效果。期盼未來能夠利用分類器，對災難的資訊進行自動化處理，將即時訊息在最短的時間內作初步過濾，提供給各類訊息相對應的單位團體利用（本研究提出的災難資訊自動化分類概觀，如圖 1）。

一、 系統設計

圖 2 為本研究的系統設計概觀與流程，包括撈取網路社群資料、文字辨識、文字轉碼、移除干擾、中文斷詞處理和去除停用詞（Stop Words）等等資料前處理步驟，並將資料儲存至資料庫中保存，而後進行專家文本分類作為機器學習使用。在機器學習前，研究者必須抽取更具意義的特徵資料，作為建立向量空間模型的維度，以 TFIDF 向量模型來做為分類器的訓練，在訓練時期設計交叉驗證（Cross Validation）的方法，對訓練後的各個分類模組進行驗證及比對。

二、 資料來源概論

莫拉克（Morakot）為 2009 年太平洋第 8 號颱風，在 8 月 9 日至 8 月 11 日三日期間為南台灣帶來了罕見的驚人雨量，單高雄市山區即帶來超過 2,500 毫米的雨量，等同三日內降下一年的雨量。在短時間內狹帶超量雨水集中於台灣南部地區，形成大規模水災重創南部地區，也帶來少見的資訊洪流。

水患的災民與其親友於此次水災中，在無法得知足夠資訊的恐慌焦慮下，只能四處求援，大量的求助電話癱瘓了部分縣市消防局緊急應變中心(119)的通訊系統。既有管道壅塞的情形下，許多得不到第一手訊息的人紛紛在各大網路論壇大量留言請求群眾協尋或轉貼分享，引發網路資訊氾濫，這現象引起一些網路使用者與技術人員的注意，在相互引導下組織動員，自發性成立災情資訊頻道，提供莫拉克災情相關的資訊內容張貼和討論。

本研究中使用的資料，即為 2009 年發生在南台灣地區的莫拉克八八風災的各類型頻道資料，利用程式的解析處理，將既有頻道（屏東及台南縣市 119 報案電話記錄）和網路上備援頻道（ADCT 數位文化 Twitter 資料）及浮現頻道（XDite、Ricks）等五個資料來源，整理儲存在資料庫中，做為本次研究的歷史資料集。

這五個資料集的概述如下：

（一） Tainan119 報案電話記錄：由於報案電話記錄屬敏感性資料無法直接獲得數位檔案，研究者經管道取得台南縣於莫拉克風災期間的紙本記錄資料。經過掃描成圖檔再透過機器文字辨識與人工校對後，將資料儲存為逗號分隔文字檔（.csv），再編碼轉換後存入資料庫。該資料集的欄位包括：獲報時間、鄉鎮別、災害地點、災情類別、災害狀況、處理情形，共取得 Tainan119 資料集 2436 筆。

（二） Pingtung119 報案電話記錄：研究者取得紙本報案記錄，經由同樣的資料處理過程存入資料庫，共取得 Pingtung119 資料集 512 筆。

（三） 莫拉克災情資料表（由 Ricks 架設，以下簡稱 Ricks）：為網友自行架設的網站，以 Excel 表單方式供群眾填寫災情需求及回覆。研究者於風災發生期間，透過人力擷取「莫拉克颱風災情資料表」網站資料，儲存為 Excel 檔案格式（.xls）。 資料欄位有：發生時間、推文與否、鄉鎮市、詳細地址、聯絡方式、發生災情、需要協助內容、最新狀態，Ricks 資料集總共取得 4193 筆資料。

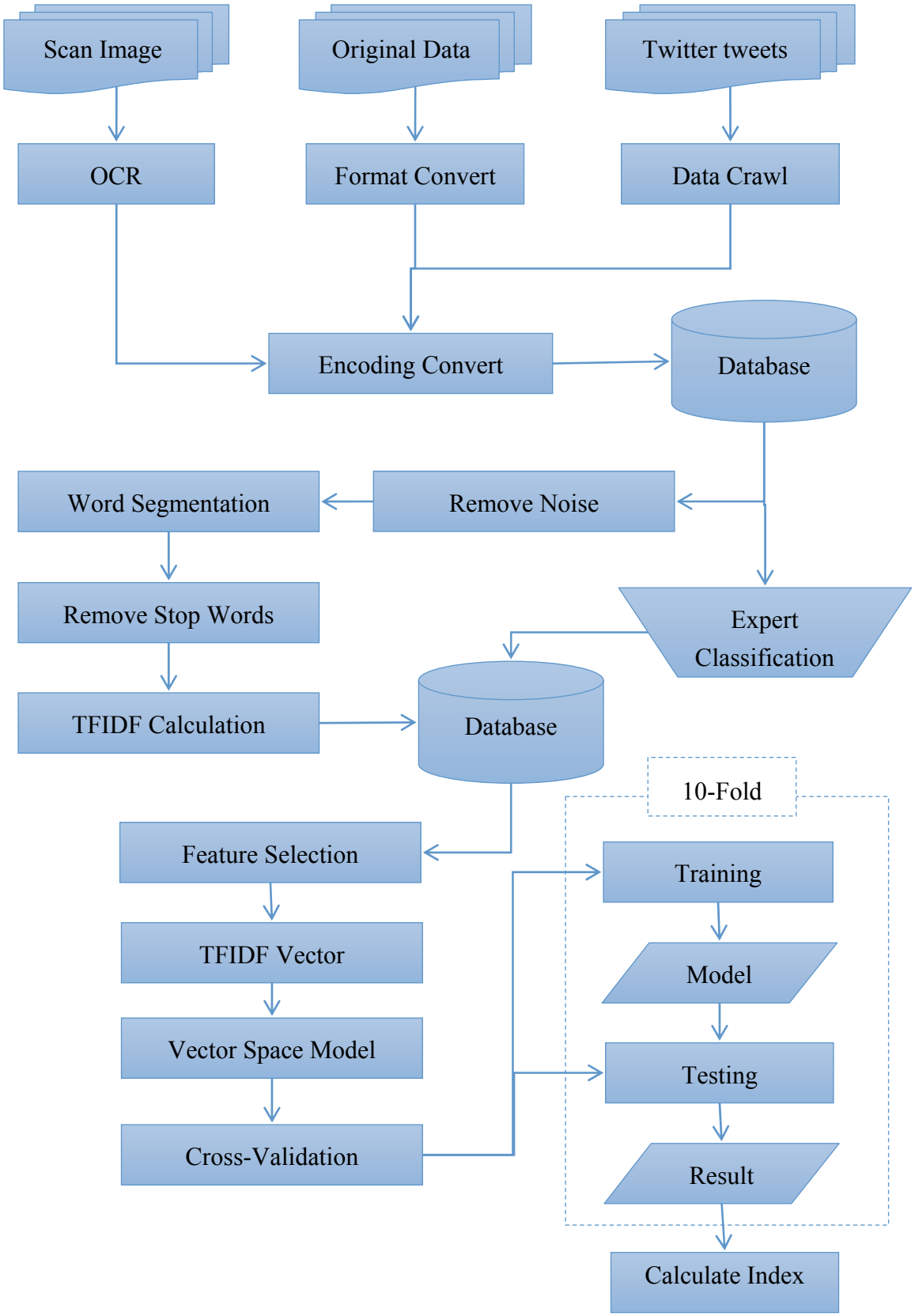


圖 2：系統設計概觀與流程

（四） 莫拉克災情支援網（由 Xdite 架設，以下簡稱 Xdite）：同樣為網友自發性動員架設的網站，為論壇形式的網站，提供多對多的互動方式。研究者經同意取得資料庫檔案（.sql），檔案中包含許多資料表，主要發文和回文內容資料表的資料欄位有：文章標題、文章內容、發文時間、更新時間、垃圾訊息標示。Xdite 資料集總共取得 9499 筆資料。

（五） 莫拉克民間災情網路中心（由台灣數位文化協會 Association of Digital Culture, Taiwan 架設，以下簡稱 ADCT）：台灣數位文化協會資料集是透過協會官方的 Twitter 帳號（@adctnpo）將收集整理過後的災情資訊發布至 Twitter 上，這些資料是將網路災情資訊透過人工確認正確後才發佈。研究者使用 Twitter API 將 2009 年 8 月 9 日至 2009 年 10 月 6 日期間 ADCT Twitter 的發布內容擷取下來，共取得資料 2,099 筆資料。Twitter JSON 資料中包含許多 tweets 和 User 的基本資料，轉換後我們僅保留了幾個欄位內容儲存，分別是 Tweet 編號（id）、tweet 內文（text）、發文時間（created_at）、發文者名稱（screen_name）、轉載次數（retweet_count）。

三、 資料收集與儲存

本研究所使用的資料集包含了風災期間三種頻道（既有頻道、備援頻道、浮現頻道）在災難發生當下資料的內容，研究者在災難後透過不同管道取得原始資料，但除了部分資料為數位檔案外，有些資料以紙本方式儲存。因此，研究者必須先經過圖形文字辨識、文件格式轉換、資料擷取過濾匯入等方式，統一儲存至關聯式資料庫。不同來源的資料因屬性相異，研究者儲存至不同的資料表中，每一筆資料給予一個不重複且獨立的編碼，這些資料我們稱為「歷史性資料（Historical Data）」。

歷史性資料來自不同單位和平台，有部分儲存前已經過文字辨識或資料編碼轉換，可能有亂碼或缺漏出現在文章內容中。為求較好的文字品質以減低文字分析時的錯誤率，在完成儲存後先做資料清查（Data Cleaning）的動作，以人力將所有經過文字辨識（屏東縣及台南縣 119 報案電話記錄）、檔案轉換（Ricks）的資料內容進行校正。

四、 資料前處理

（一） 編碼轉換與移除干擾雜訊

各國依據自己的語言特性，訂立了不同的文字編碼系統。為了減少分析可能發生問題，將收集的內容都轉換為相同的文字編碼。為了避免轉換過程中，因轉換的字集大小不同的問題而遺失部分文字，我們選擇目前網頁上常見的 UTF-8（8-bit Unicode Transformation Format）編碼。

為了豐富文字中的情緒和趣味化，許多使用者會在文字段落中加入顏文字（表情符號）。本研究所使用的資料內容部分來自於網路，內容包含如Σ(゜△ ゜|||)ｼ、(￣.￣)+、(=￣ω￣=)等等的顏文字，為了避免這些符號對分析的干擾，在文字前處理時我們將此類符號先由內容中移除。

（二） 中文斷詞處理與詞庫

詞彙（Word）是最小且有意義的語言單位，任何語言處理的系統都必須要先能夠分

辨文本中的詞才能夠進行更進一步的處理，所以在做文本處理時斷詞便成為不可或缺的技術。中文的文本中句子裡沒有分隔符號，無法直接取出中文詞彙，必須依靠中文斷詞系統將可能的詞彙先行處理。

常見的中文斷詞工具有中央研究院 CKIP 中文斷詞與剖析系統、Yahoo 的斷章取義、mmseg4j 等。考慮未來若發展為即時監測，需注意斷詞速度、使用的數量和具有擴充性，因此在本研究中選擇了 mmseg4j 做為我們的斷詞工具。

mmseg4j 這個工具主要是用來替簡體中文斷詞，不過因為它的詞庫強制使用了 UTF-8 編碼且可自行建立詞庫，所以也可以使用 mmseg4j 來替繁體中文進行斷詞處理。mmseg4j 預設的詞庫來源為大陸搜狗搜尋引擎的詞庫，內容用語比較貼近大陸地區的使用方式，不適合用來做為台灣地區的中文文本斷詞使用。我們以教育部於 1997 年發行的「國語辭典簡編本編輯資料字詞頻統計報告」("國語辭典簡編本編輯資料字詞頻統計報告," 1997)網路版中的〝字頻總表〞與〝詞頻總表〞，重新建立 mmseg4j 中的單一字對應頻率詞庫檔（chars.dic）和核心詞庫檔（words.dic）。

中文斷詞的原則需要依賴詞庫，斷詞系統除了基本詞庫之外，使用者必須附加專屬領域詞庫。在災難發生的情境下，網路使用者發布的訊息文本，可能會異於平常使用的詞彙用語，同時包含大量地理資訊內容。截至目前，尚無針對災難發展相關領域的詞庫資料可直接運用，因此本研究在進行中文斷詞時需自行建立災難相關的詞庫。根據八八風災所收集的資料集內容，研究者採用以下二種方式建立本地風災的災難詞庫。

1. 地理位置詞庫

由於八八風災涉及許多地方淹水或人員受困，因此民眾求助訊息經常包含地理資訊，如：屏東縣那瑪夏鄉，若研究者能以地理位置詞彙做為災難詞庫，可大幅減少需要進一步進行挑出斷詞的內容。因此研究者採用台灣「中華郵政有限公司」維護的郵遞區號檔("3+2 碼郵遞區號 Excel 檔 101/05," 2012)，抽取其中的縣市鄉鎮里鄰及街道等資訊，重新依字詞筆劃排序，做為八八風災的地理詞庫。

2. 建立災難詞庫

以本研究所收集的資料集為對象，透過不斷的隨機抽取斷詞後內容，挑出有意義的詞彙加入詞庫再斷詞，直至隨機抽取的內容詞彙穩定。為了能更容易的挑選這些有意義的詞彙，本研究自行開發設計詞彙收集的網路應用程式，讓研究人員可以容易且直覺地瀏覽斷詞結果與挑選儲存。

（三） 移除停用字

另外，一般文件中常包含語助詞、副詞和連接詞等經常出現的慣用字詞，尤其是發表於網路的一般文章中，這類詞彙出現的頻率相當高，但它們對於文件的分析較不具有意義，在資料探勘的領域中稱這些詞彙為停用字（Stop Words）。在分析前，將符合停用字集合的詞彙，由斷詞後的文件中移除，避免停用字影響後續分析的結果。

中文檢索方面目前仍缺少標準的停用字詞表，在本研究中所使用的停用字集合是參考「中央研究院平衡語料庫詞集及詞頻統計」資料。此統計是根據帶有標記的中央研究院平衡語料庫所計算出的詞頻統計資料，收錄有五百萬個詞。我們取統計資料中頻率最高的前 100 個詞彙，建立本研究用的中文停用字集合。

五、 詞彙重要性指標（TFIDF）

然而，並非所有文件中的詞彙都是一樣的重要，研究者需要一個指標來表示詞彙的重要性，我們選擇以 TFIDF 來做為指標。TFIDF 是用來評估一個詞彙對於一個文件集的重要程度，由詞頻（Term Frequency，TF，式 1）與逆向文件頻率（Inverse Document Frequency，IDF，式 2）組成。

$$\mathrm{tf}(t,d) = \frac{\mathrm{f}(t,d)}{\max\{\mathrm{f}(w,d) : w \in d\}}$$ （式 1）

分子 f(t.d)為該詞彙再文件中出現的次數，分母是該文章的詞彙總數

$$\mathrm{idf}(t,D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$ （式 2）

分子 |D| 資料集中的為文件總數，分母是出現該詞彙的文章數

詞頻是指一個詞彙於文件中出現的頻率，為了避免單純使用詞彙次數會出現的偏差（在一份長文字文件裡出現的次數會高於短文字文件），除以文章中的詞彙總數來標準化數值。逆向文件頻率是用來表示一個詞彙在整個文件庫中的普遍重要程度。如果包含某個詞彙的文件越多，代表這詞彙在文件庫中較不具有分辨的重要性，IDF 的數值就較低。結合詞頻與逆向文件頻率成為 TFIDF 指標（TF-IDF = tf × idf），這個指標會傾向於過濾掉太常見的詞彙，保留較重要的詞彙。

六、 內容分析（人工分類）

另一方面，本研究亦透過人工方式進行不同資料集的內容分析與訊息分類。藉由災難傳播的文獻檢閱，歸納出不同類別的災難訊息，並由傳播科系研究生做為編碼員，經過一定時間的編碼員訓練，以及檢驗不同編碼員之間的信度與效度，達到信度標準後再進行分類，讓編碼員對五個歷史性資料集進行內容分析與分類標示，並由系統儲存編碼員所標示的分類結果，作為機器學習之用。

（一） 分類項目

經過災難傳播的文獻檢閱，研究者將人們在災難期間的資訊需求分為三大類型，包括：資訊、行動、表達，分別意指：提供或要求資訊、徵求或採取行動、表達或討論，若無法歸類到上述三群組則視為第四群組。本研究進一步檢視收集到的八八風災災情資料，發現在這次颱風造成的災情以淹水、道路受阻、土石流為主，這使得前述的三大類型又可進一步區分幾個次類型。

最後，研究者根據莫拉克風災所展現的災情狀況，將災難資訊分為九大類別：（1）資訊：請求協尋失聯人士（2）資訊：提供災難地理位置等情境資訊（3）資訊：轉貼傳媒或政府公告（4）行動：請求救援（5）行動：徵求或提供志工物資（6）表達：討論反應（7）表達：要求網友自律（8）其他：搭便車的公關行銷訊息（9）其他：資訊不完整無法分類。

（二） 時間隨機抽樣分類

在進行內容分析前，研究者先將每個資料庫內容依照發布時間先後排序，並以等距抽樣的方式，抽取十分之一的內容筆數作為分類用的母群資料。為避免分類母群可能會

與分類項目產生時間性干擾偏差，設計程式由母群中隨機抽取尚未分類的內容，提供給編碼人員進行分類編碼，並於顯示內容時隱藏時間資料。

（三） 編碼員信度分析

針對專家建議定義的九個類別的分類原則，製作成編碼參考表（Coding Table）以訓練分類編碼人員。當編碼人員超過一人時，必須檢測不同編碼人員的一致性（Consistency）求取信度（Reliability）的可靠性，確保資料分類編碼的過程中不會受到人、事或工具的影響而產生變化。隨機抽取五十分之一的資料筆數，一式多份給不同的編碼人員進行訓練與計算其信度。信度參考王石番（1991）傳播內容分析法中公式（式3），信度檢定需達到 0.80 以上信度係數標準，使完成分類編碼訓練進行實際分類。

$$\text{Intercoder Agreement (IA)} = \frac{2 \times M}{N_1 + N_2}$$

$$\text{Composite Reliability (CR)} = \frac{N \times (Average\ of\ IA)}{1 + [(N-1) \times (Average\ of\ IA)]} \quad \text{（式 3）}$$

M：編碼完全相同數量，N：參與編碼的人員數

N_1：第一位編碼人員，N_2：第二位編碼人員

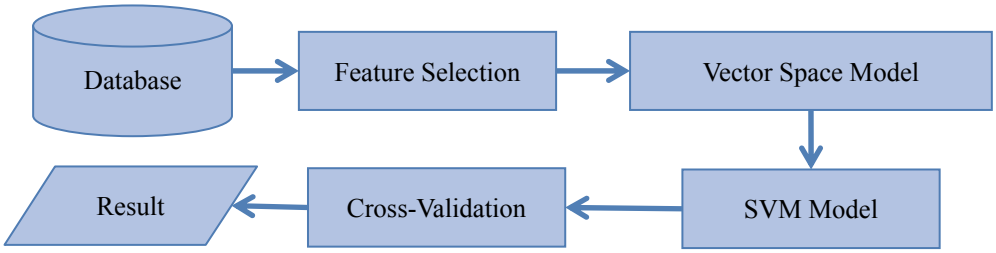
由於每個資料集所具有的資料筆數差異頗大，由 500-9500 筆不等，為了達到後續資料的一致性，研究者先將五個資料集分成二組，並進一步在每一組內資料集抽取數量相近的資料筆數，交由編碼員進行內容分類。第一組是專家書寫的既有頻道，包括台南與屏東 119；第二組是來自大眾書寫的浮現頻道，包括 XDite, Ricks, ADCT；每個資料集完成專家分類的筆數如下表 1：

表 1：各資料集實際分類筆數

	資料集名稱	母群筆數	實際分類筆數
既有頻道	台南 119 報案電話記錄	2436	542
	屏東 119 報案電話記錄	512	512
浮現頻道	XDite	9499	1056
	Ricks	4193	1050
	ADCT	2099	1050

七、 機器學習

機器學習（Machine Learning）關注的是電腦程式所能達到的學習，透過結合統計、機率等方法將經驗累積達到自動學習的效果。只要給定問題的範圍和訓練資料（Training Data），由資料中選擇特徵資訊（Feature Selection），然後建構資料的模型（Model Selection），把模型當做學習的成果，可以用來預測未知資料內容（如類別）



（一）文件特徵選取

特徵選取是訓練分類器模型時，在無損機器學習的效能下，保留下真正對效能指標有影響的特徵資料，來達到降低特徵空間維度的目的。

在研究中，我們採用文件分類時效果較好的卡方分析（CHI-Square Test）特徵選取的方法，利用統計方法計算每個詞彙指標值（式 4），排序後以特定閾值（threshold）過濾的方式，挑選出有影響力的特徵詞彙子集合。

		詞彙	
		是	否
類別	是	A	C
	否	B	D

$$\chi^2(t,\ c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad \text{（式 4）}$$

t：詞彙　　c：類別

A：同時出現指定詞彙與類別的數量　　B：出現指定詞彙但是不在類別中的數量
C：類別中其他非指定詞彙的數量　　D：不在類別中也非指定詞彙的數量

卡方分析（χ2 Statistic measure, CHI-Square Test）的方法又稱為獨立性檢定，是用來確定兩個組別的獨立性，使用在文件分類的特徵選取上，即表示用來觀察詞彙與類別之間的獨立性，越高的數值代表詞彙與類別的相關性越高。

（二）向量空間模型

向量空間模型（Vector Space Model）(Salton, Wong, & Yang)是一個應用於資訊擷取與過濾、索引文件和評估相關性的代數模型。向量空間模型中，將每個文件的重要詞彙作為代表文件的屬性，聯集所有的屬性表示為高維向量空間中的的獨立維度，形成文件的屬性向量。每個維度為文件中的關鍵詞彙，給予每個詞彙一個權重數值，代表詞彙在文件中的重要性，即能表示文件在高維空間中的情形（圖 3）。

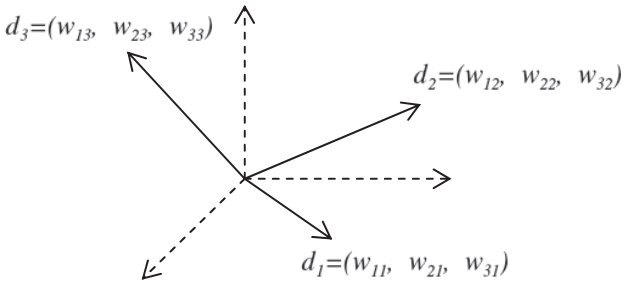


圖 3：空間向量示意

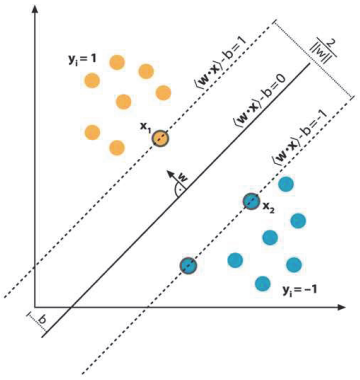


圖 4：支持向量機示意

（三）支持向量機（Support Vector Machine）

本研究所使用的機器學習方法，是一種監督式學習（Supervised Learning）法。在方法的選擇上，以近年來在分類具有高公信力的 SVM 來做為機器學習的模型建立方法。SVM 是以統計為基礎的機器學習演算法，是一種處理二元分類（Binary Classification）的方法，在高維度的空間中尋找一個超平面（Hyperplane）將兩個類別分隔開。這樣的超平面可能有很多個，SVM 的目的在找出最佳的超平面，使兩個類別的邊界（Margin）幅度最大話，能夠明顯區分兩個類別（圖 4）。

（四）交叉驗證（Cross Validation）

本研究中所採用的方法是 10 折交叉驗證（10-Fold Cross Validation），10-Fold 是將資料分割成 10 個子集，其中一個子樣本被留作驗證模組使用，其餘 9 個子集被用來訓練，重複 10 次，每個子集都被驗證一次，平均 10 次的結果得到一個評估值。

（五）績效評估

資訊檢索領域中的文件自動分類的研究中，常見的績效評估指標有正確率（Accuracy）、查準率（Precision）、查全率（Recall）、F 度量（F-measure）。如果單純使用查準率（Precision）、查全率（Recall），可能會因樣本數的變化而有高查準率低查全率，或是低查準率高查全率的情形。而 F 度量是一個調和平均數，同時考慮了查準率與查全率，只有查準率和查全率都高的時候 F 度量的值才會高，能夠作為客觀指標衡量整體的績效。

在研究中有五個不同頻道來源的資料集，每個資料集各進行 10 折交叉驗證（10-fold Cross Validation），將已建立好的文件向量和對應的類別集合分成十份，其中九份作為訓練集，剩下一份作為測試集，依序進行十次訓練，每次都取得 F 度量指標，並計計算指標的平均值和 95%信賴區間（95% confidence interval），作為評估 SVM 的結果。

肆、實驗結果分析

一、 詞彙網絡分析

本研究中為瞭解不同屬性資料集中詞彙彼此間的關係，利用發佈的文字內容建置詞彙共現網絡（Words Co-Occurrence Network）。其中詞彙為網絡中的點（Node），共同出現在同一篇文章的關係則為網絡中的邊（Edge）。

然而，並非所有出現過的詞彙都具有代表文章或分類的高相關性，本研究以詞彙的平均卡方值與最大卡方值，作為用來挑選建置網絡關係詞彙的條件。排序後取出平均卡方值與最大卡方值取前 200 個詞彙，當做共現網絡中的點，並以這些詞彙共同出現在一篇發文的情形，作為共現的範圍建立網絡中的線。以共現值大於等於一做為門檻值（至少有一個線），可以得到如下表 2 資料。

表 2：卡方特徵詞彙共現網絡結構

	AVG-CHI			MAX-CHI		
Dataset	Nodes	Edges	Density	Nodes	Edges	Density
Tainan119	116	1839	0.092	54	348	0.017
Pingtung119	105	1649	0.083	44	286	0.014
Ricks	156	8895	0.447	121	3029	0.152
Xdite	178	13035	0.655	156	5252	0.264
ADCT	137	5756	0.289	134	1351	0.068

從上表中可以發現使用最大卡方值（MAX-CHI）選出來的特徵詞彙資料，在各個資料集中的節點數與連線數皆遠小於平均卡方值（AVG-CHI）。反應出使用最大卡方值選出來的詞彙，其獨特性可能會成為資料分類時的關鍵，但平常是比較少被一起使用的詞彙。既有頻道（Tainan 與 Pingtung119）與浮現頻道（Ricks 與 Xdite）比較，可以發現節點數差異不大，既有頻道的連接線數量卻少了很多，進一步解釋其原因在於救難專家的書寫相較於大眾的俗民書寫，前者在詞彙使用上比較精確，不容易重複使用相同的詞彙。

若觀察幾個資料集的密度指標，當一個網絡中的成員同質性越高越單純，網絡的密度就會較高。在表二中可以發現，在 Ricks 與 Xdite 資料集中的詞彙，具有較高的同質性；然而，同質性高，可分辨性就會較低，可能在機器學習的效果上會較不顯著。

就網絡分析的特點而言，網絡的中心性（Centrality）可以幫助了解節點在整個網絡中的重要程度，在文本中。Betweenness Centrality 反應的是一個詞彙對於其他詞彙出現在同一篇文章中的影響力，數值越高的詞彙表示對其他詞彙的共現影響越大。我們以 Betweenness Centrality 的值做為節點大小，透過模組性分群（Modularity）將詞彙叢聚分群後分色，繪製詞彙共現網絡圖，如圖 5。

由圖 5 可以發現，在資料集中具有高影響力的特徵詞彙皆不同，Xdite 中最具影響力的詞彙為「物資」，恰好能夠呼應資料集的特性與分類的分佈概況。另外，由詞彙分群的情形可以看出同一個群聚中的詞彙幾乎和分類的情形有關連性，如的 Ricks 網絡圖中我們可以看到藍色的群聚（物資、志工、救災、救援）和分類中的「物資志工」類別相近。綠色的群聚（聯絡、知道、消息、家人、聯絡）和分類中的「請求協尋」相近。

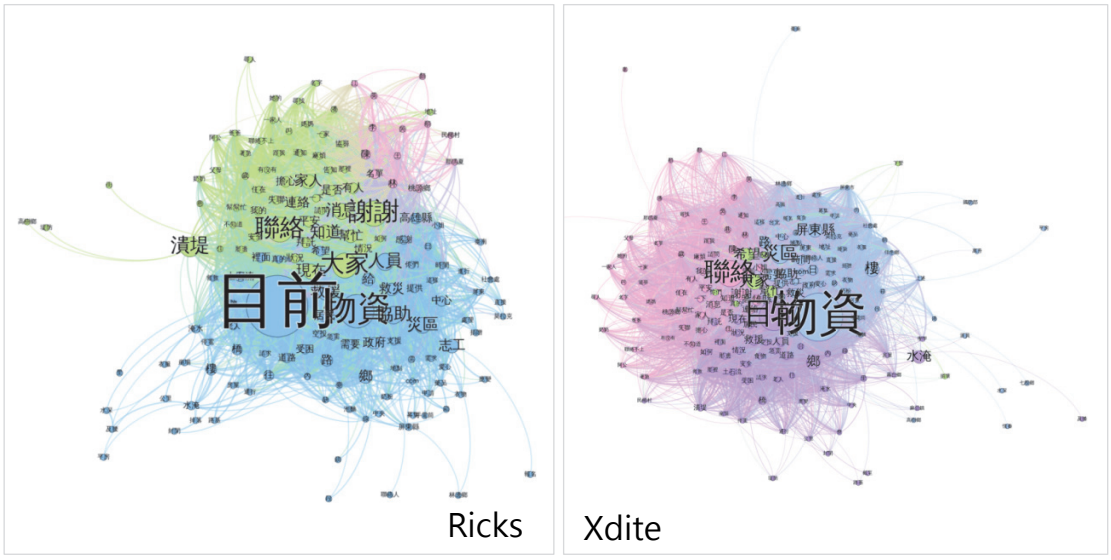


圖 5：卡方值特徵詞彙共現網絡圖

二、 機器學習比較

本研究使用監督式學習中的 SVM 作為機器學習自動分類的方法。以卡方值篩選文件中的特徵資料詞彙，以每個詞彙的 TFIDF 建立空間向量模型進行 10-fold 交叉驗證，得到 10 次 F-measure 計算平均值與 95%信賴區間的結果做為分類的績效評估。

過多的特徵資料可能成為可能會成為干擾因素，太少的特徵資料卻又可能不足以構成分類的條件，多少數量的特徵資料才適合做為機器學習訓練的依據，便成為研究者的一大挑戰。過去研究中，沒有發現針對中文網路資訊多分類的特徵值數量可作為研究參考，因此在實驗的開始，我們取出不同的特徵詞彙數量進行訓練實測，發現特徵詞彙數量為 7000 個時，多資料集的績效可達到最好。因此，在所有實驗中，我們選擇特徵詞彙數量為 7000 個作為機器學習訓練的標準。

（一） 各資料集單獨訓練的結果

	Tainan119	Pingtung119	Ricks	Xdite	ADCT
AVG-CHI	0.83±0.02	0.67±0.04	0.54±0.03	0.49±0.03	0.53±0.02
MAX-CHI	0.83±0.05	0.67±0.04	0.52±0.02	0.49±0.02	0.52±0.02

根據上述結果發現，兩個既有頻道 Tainan119 與 Pingtung119 訓練後的績效表現較好，而浮現頻道中的 Xdite 訓練績效表現較差。探究可能原因，兩個縣市 119 報案電話記錄的內容皆為救難專家書寫，用字遣詞皆經過專業訓練。反之，Xdite 資料集的內文完全為網路使用者自行填寫，也就是俗民書寫的內容，不一定經過 專業訓練，且在緊急狀況下有許多語句不通順及錯別字問題，可能造成中文斷詞與分類上的困難。Ricks 雖然和 Xdite 同樣是俗民書寫，其內容絕大多數為請求協尋，有相對明確的人名或地名，因此提資訊正確性相對來說較高，其機器學習訓練的績效表現也在 Xdite 資料集之上。而 ADCT 則受限於 Twitter 的 140 字數限制，加上大部分的資訊為媒體連結轉貼，所以較難有好的分類做為訓練。

（二） 交叉比對不同資料集分類器

在本次研究，研究者透過各種管道取得不同性質頻道的資料集，所以能夠嘗試將不同資料集訓練而得的分類器，用來預測其他資料集內容的類別，交叉比較不同性質的資料集分類器的績效。

AVG-CHI	Tainan119	Pingtung119	Ricks	Xdite	ADCT
Tainan119-CM	-	0.24±0.00	0.13±0.00	0.03±0.00	0.04±0.00
Pingtung119-CM	0.33±0.01	-	0.13±0.00	0.03±0.00	0.06±0.00
Ricks-CM	**0.76±0.01**	0.27±0.01	-	0.26±0.00	0.09±0.00
Xdite-CM	0.04±0.00	0.02±0.00	0.32±0.01	-	0.07±0.00
ADCT-CM	0.16±0.01	0.05±0.01	0.17±0.00	0.11±0.01	-

*CM：Classification Model

MAX-CHI	Tainan119	Pingtung119	Ricks	Xdite	ADCT
Tainan119-CM	-	0.24±0.00	0.13±0.00	0.04±0.00	0.04±0.00
Pingtung119-CM	0.33±0.02	-	0.13±0.00	0.03±0.00	0.07±0.00
Ricks-CM	**0.77±0.00**	0.27±0.00	-	0.26±0.00	0.09±0.00
Xdite-CM	0.04±0.00	0.01±0.00	0.32±0.00	-	0.07±0.00
ADCT-CM	0.15±0.01	0.04±0.00	0.16±0.00	0.11±0.00	-

*CM：Classification Model

透過上述二個結果發現，屬於相同頻道的資料集間具有些微的分類績效，對其他頻道的效果不好。其中 Ricks-CM 比較特殊，除了 ADCT 之外，對於其他幾個資料集分類也具有一些分類上的績效。對 Tainan119 的資料非常顯著，達到 0.76±0.01、0.77±0.00 相當高的數值。而 Tainan119-CM 對於 Ricks 資料的分類績效卻不好。Ricks-CM 在分類 Tainan119 的資料時，於 C2（提供情境資訊）類別的效果最好，觀察內容，兩個資料集在這個類別的內容，對於地點和狀況用詞都相當簡潔明確。

（三） 合併相同性質頻道資料集訓練分類器

前述的情形，促使研究者想了解資料數量對於訓練分類的影響，如果將訓練資料量放大是不是會有更好的效果，因此將相同頻道性質的資料庫合併訓練。

合併後資料筆數：

Tainan119 + Pingtung119 = 542 + 512 = 1054（筆）
Xdite + Ricks = 1056 + 1050 = 2106（筆）

AVG-CHI	Tainan119	Pingtung119	Ricks	Xdite	ADCT
T+P-CM	**0.95±0.03**	**0.86±0.05**	0.15±0.00	0.04±0.00	0.06±0.00
R+X-CM	0.44±0.02	0.27±0.01	**0.83±0.07**	**0.86±0.04**	0.19±0.00

T+P：Tainan+Ptingtung119　R+X：Ricks+Xdite　CM：Classification Model

MAX-CHI	Tainan119	Pingtung119	Ricks	Xdite	ADCT
T+P-CM	**0.94±0.03**	**0.85±0.04**	0.16±0.00	0.05±0.00	0.06±0.00
R+X-CM	0.48±0.02	0.29±0.01	**0.83±0.07**	**0.85±0.04**	0.19±0.00

T+P：Tainan+Ptingtung119　R+X：Ricks+Xdite　CM：Classification Model

當訓練的資料數量增加以後，對同屬頻道的資料集分類績效明顯提高，但對於不同頻道的分類還是沒有太大的提昇。

（四） 合併所有資料集訓練分類器

另一方面，研究者嘗試將所有資料集合併訓練，以增加數量與平衡類別間的差異。結果在各資料集表現有顯著的提昇。

All-CM	Tainan119	Pingtung119	Ricks	Xdite	ADCT
AVG-CHI	0.93±0.03	0.86±0.05	0.83±0.05	0.87±0.05	0.87±0.03
MAX-CHI	0.93±0.04	0.84±0.04	0.82±0.06	0.86±0.05	0.85±0.04

當資料量變大（4210 筆）預測力明顯提升，表示打散原本不同來源資料集成為一個資料集時，九種分類彼此具有高區辨性，且每個分類集合內的特徵值相似度高，使預測能夠發揮有效預測力。各個資料集差異雖大卻沒有衝突，具有很高的互補性，因此合併這些資料集來訓練自動分類器是安全的。無論是既有頻道的專家書寫或浮現頻道的俗民書寫，在去除停用字僅保留災難詞庫關鍵字後，二者對於同一種類型的訊息描述是相近。這顯示專家書寫比俗民書寫的精準度在於減少贅字與冗言只用關鍵字。

（五） 將分類器當做其中一個編碼員

最後，研究者將合併所有資料集訓練所得的分類器當做第三個編碼員，加入先前的專家文本分類中進行預測，並與先前兩個專家編碼員的分類結果進行複合信度分析，得到：

Intercoder Agreement	Coder1	Coder2
AVG-CHI SVM	0.68	0.72
Coder2	0.8	-

平均相互同意度（Average Intercoder Agreement）$= \frac{0.68+0.72+0.8}{3} = 0.733$

複合信度（Composite Reliability）$= \frac{3\times0.733}{1+((3-1)\times0.733)} = \mathbf{0.892}$

Intercoder Agreement	Coder1	Coder2
MAX-CHI SVM	0.7	0.74
Coder2	0.8	-

平均相互同意度（Average Intercoder Agreement） $= \frac{0.7+0.74+0.8}{3} = 0.747$

複合信度（Composite Reliability） $= \frac{3\times0.747}{1+((3-1)\times0.747)} = \mathbf{0.899}$

兩者複合信度值皆超過 0.80 的門檻值，表示兩種機器訓練後的分類器與兩個專家編碼員的分類具有相當的內部的一致性。

伍、結論

在八八風災之後，研究者收集來自既有頻道、浮現頻道與備援頻道的不同資料集，透過資料處理與斷詞，以機器學習專家文本分類內容，探尋自動化訊息分類的可能性，經過分析與實驗得到以下幾項結論：

一、 以最大卡方值建置特徵詞彙的共現網絡關係，能展現分類代表性

以平均卡方值、最大卡方值建置特徵詞彙的共現網絡關係，藉以觀察不同資料集中詞彙的關係與特徵詞彙對機器學習分類可能的影響。最大卡方值所選擇的特徵詞彙，節點間的關係數量遠小於平均卡方值所篩選的特徵詞彙，其為不常被使用，卻具有分類代表性的詞彙。既有頻道的內容是由救難專家書寫，較不易在同篇內容重複使用相同詞彙，所以在數量雷同的詞彙下，節點的連線數量遠低於以俗民書寫為主的浮現頻道之連接線數量。在網絡規模與網絡密度指標，浮現頻道中使用的詞彙同質性較高，具有較高的同質性，在機器學習的效果較不顯著。

二、 專家書寫的災難訊息其進行自動化分類的績效表現，優於俗民書寫

救難專家書寫的頻道內容，詞彙使用較為精準簡潔，分類的績效表現皆優於俗民書寫的頻道內容。俗民書寫在災難情境下普遍存在錯別字與同音字問題，加上沒有字數條件限制可能造成一篇文章存在兩種分類，導致分類的績效最差。

三、 增加相同性質的資料進行機器學習訓練，可大符提高分類績效

若合併相同性質頻道資料的訓練，可有效提升相同頻道中的分類績效，且救難專家書寫的分類績效較優，因此可以推論，當訓練資料的品質夠好時，分類器能夠有不錯的分類績效。若資料的品質不夠時，可借由增加訓練資料的數量來提升分類的績效。另外，透過合併所有資料的機器學習訓練發現，各個資料集彼此間的差異性雖然相當大，相互之間卻沒有太大的衝突性，可能反而具有較好的互補性。

此外，研究使用合併全部資料集已文本分類的內容做為訓練資料，藉以達到足夠的數量與分類的分佈平均，得到的分類器模組做為第三個編碼員，加入一起分類先前編碼員訓練信度的資料，與兩位人類編碼員再計算平均相互同意度與複合信度，得到超過門檻值相當多的內部一致性。

整體而言，本研究最後經由機器學習自動化分類的實驗，證實災難情境下進行自動化分類資訊的可能性，其應用價值在於日後災難發生時，能透過自動化分類器即時過濾線上訊息，可將篩選過的資訊提供予相對應的救災或援助單位。另方面，此一設計構想未來若能合併事件頻率偵測的感測器，可進一步發展出社群感知器（Social Sensor），將社群感知器放置於不同的的社群網絡位置上，可自動偵測事件的發生與擷取有用的資訊，

以協助政府單位在第一時間制定決策，更有效的擬定傳播策略，或對於目標族群進行宣導。

陸、參考文獻

孫式文. (2000). 網際網路在社會危機中的功能：網友調查研究. Paper presented at the 2000 網路與社會研討會, 新竹.

國語辭典簡編本編輯資料字詞頻統計報告. (1997) Retrieved 01.23, 2012, from http://www.edu.tw/files/site_content/m0001/pin/c11.htm?open

陳百齡, & 鄭宇君. (2011). 災難情境下的新興媒體：莫拉克風災中的浮現頻道. Paper presented at the 中華傳播學會 2011 年會, 新竹，交通大學.

顧佳欣. (2009). 莫拉克效應：災難傳播要善用資源 Retrieved 07.28, 2012, from http://www.feja.org.tw/modules/news007/article.php?storyid=395

鄭宇君（2014.01）。〈向運算轉：新媒體研究與資科技術結合的契機與挑戰〉，《傳播研究與實踐》，4（1）：45-61。

3+2 碼郵遞區號 Excel 檔 101/05. (2012) Retrieved 06.10, 2012, from http://www.post.gov.tw/post/internet/down/index.html

Brooks, C. H., & Montanez, N. (2006). Improved Annotation of the Blogopshere via Autotagging and Hierarchical Clustering. Paper presented at the WWW2006 Conference, Edinburgh, UK.

Cormack, G. V., Hidalgo, J. M. G., & Sanz, E. P. (2007). Spam filtering for short messages. Paper presented at the Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, Lisbon, Portugal.

Gupta, R., & Ratinov, L. (2008, July 13-17). Text Categorization with Knowledge Transfer from Heterogeneous Data Sources. Paper presented at the Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, Chicago.

Munro, R., & Manning, C. D. (2010). Subword Variation in Text Message Classification. Paper presented at the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, Los Ang

eles, California.

Potts, L. (2009). Peering into disaster: Social software use from the Indian Ocean earthquake to the Mumbai bombings. Paper presented at the In Proceedings of the International Professional Communication Conference, Hawaii.

Quarantelli, E. L. (1998). The Computer Based Information/Communication Revolution: A Dozen Problematical Issues And Questions They Raise For Disaster Planning And Managing: Disaster Research Center.

Rogers, E. M. (1995). Diffusion of Innovations. New York: The Free Press

Salton, G., Wong, A., & Yang, C. S. A Vector Space Model for Automatic Indexing. Communications of the ACM, 18(11), 613.