災防數據與機器學習整合應用開發

Application of Combing Social Media and Machine Learning

國家災害防救科技中心

劉致灝[1] 張智昌[1] 蔣佳峰[1]

Liu, Chih-Hao Chang, Chy-Chang Chiang, Chia-Feng

[1]國家災害防救科技中心 災防資訊組

摘要

本計畫以國家災害防救科技中心(以下簡稱災防科技中心)於災害應變期間實行之社群網路蒐整流程為基礎，佐以統計、文字探勘、影像辨識等技術，進行資料分析與加值應用。在災害發生期間，判斷社群網路上民眾發布的文字訊息，是否包含災點資訊，利用訊息中的文字提取地名、地標等地理空間詞彙進行輔助定位；災情照片透過影像辨識技術，以辨識常見商店招牌為對象，輔助空間定位分析，相關定位結果可與文字定位結果進行交叉比對，提升災情訊息的可信度，以及災害位置的精確度。結合地理定位、透過電子地圖之展示，建立輔助性觀測模型，呈現時空間輿論災情分布，作為觀察災情趨勢的參考依據。

關鍵詞：文字探勘、機器學習、圖形辨識、地理資訊系統、社群網路。

Abstract

This project is based on the social network search process implemented by the National Science and Technology Center for Disaster Reduction (NCDR) during the disaster response period, supplemented by statistics, text mining, image recognition and other technologies for data analysis and Value-added application. During the disaster, determine whether the text messages posted by people on the social network contain location information. Our goal is to extract geographic spatial words such as location names and landmarks from those messages. Besides, photos will be identified after using image recognition technology. The signboard is the object to assist spatial positioning analysis, and the relevant positioning results can be cross-compared with the text positioning results to improve the reliability of the information and the accuracy of the location. Combining geographic positioning and displaying through electronic maps, an observation model is established to present the distribution of public opinion in time and space as a reference for observing disaster trends.

Keywords：text mining, machine learning, image recognition, GIS, social network.

一、前言

隨著資訊科技的蓬勃發展，網際網路被視為生活不可或缺的必需品，網路不僅打破了地域性的藩籬，更可以拉近人與人之間的距離。近年來，社群網站成為全球最受歡迎的網路媒介，使用者透過社群網站獲得大量的最新資訊或者與不同背景的網友相互討論、彼此腦力激盪產生多元的新觀點。社群網路是近年興起的熱門議題，也帶動了巨量資料的技術發展，網路使用者藉由資訊散佈、分享及回饋在社群網路傳遞資訊。

在災害發生期間，使用者藉由行動裝置將災情資訊傳遞到社群網站，然而透過社群網站傳遞、分享及回饋的機制，將社群災情資訊收集、分析與統整。社群網路於災害防救的應用，必須快速收集統整來自不同社群網路的災害資訊，建立社群資料中心，再透過資料技術分析，擷取出相關的災害資訊，並且進行地理位置資訊轉換。本計畫擬導入社群網路上所散播的災害訊息資訊，利用社群媒體中，民眾所上傳的文字敘述進行災情的研判，並制定文字資料分析流程，以利未來於實際災害發生時，可利用社群網路上的即時災害文字資訊進行災害的初步位置定位，以輔助災害應變的決策支援。然而目前網路上也流傳著許多假新聞與假訊息，因此除了透過民眾打卡的資訊、訊息中提到的地理詞彙如地名、地標等進行定位，提升災情訊息的可信度以及災害位置的準確度。

本計畫針對社群媒體文字訊息空間定位分析進行研究，導入社群網路上所散播的災害訊息資訊，利用社群媒體中，民眾所上傳的文字敘述進行災情的研判，以利未來於實際災害發生時，可利用社群網路上的即時災害文字資訊進行災害的初步位置定位，以輔助災害應變的決策支援。

目前網路上也流傳著許多假新聞與假訊息，因此除了透過民眾打卡的資訊、訊息中提到的地理詞彙如地名、地標等進行定位，提升災情訊息的可信度以及災害位置的準確度，嘗試導入以人工智慧的方式，針對具有空間資訊的影像與文字訊息進行分析，進一步進行位置的定位，除對於初期災情的彙整能更全面且快速外，對於訊息的真偽，也可透過相關訊息的綜整與比對進行過濾篩選，供災情綜整人員作為判斷參考。

二、研究方法

本計畫主要分為兩個類別進行研究：「文字辨識」以及「影像辨識」。我們將來自社群網路的訊息分為文字訊息與照片，文字部分採用類神經網路訓練模型LSTM進行訓練，將文字中帶有地理性詞彙進行辨識，以利後續定位處理；照片資訊則使用類神經網路訓練模型CNN訓練常見店家招牌，從圖片中找出招牌，結合文字資訊，提升定位之精準性，以及訊息本身的可靠性。

2.1 文字辨識

文字探勘（Text Mining, TM）為處理非結構化的文字資料，透過各種量化技巧，藉由量化後的特性找出文件間的相關性，找出隱藏在其中有趣或有用的訊息。因為文字探勘從文本中提取有用訊息，所以它也稱為文字資料探勘或「從文本資料庫中發現知識」，文字資料探勘通常包括五個步驟：資料選擇、資料清理、資料轉換、資料探勘、結果評估和解釋。文字探勘整合了許多傳統資訊檢索技術，包括關鍵字萃取、文件自動分類、自動摘要等等，以提供對文字處理更強大的功能。

近年來隨著 Twitter、Facebook、Instagram 等大量社群網站的崛起，因此有越來越多的研究應用主題偵測與追蹤技術於社群網站中，研究主要包含分群方法、關鍵字萃取、機率分佈的主題模型。

本計畫使用長短期記憶 (Long short-term memory, LSTM)當作主要的神經網路模型，以社群網站收集新聞以及社群言論為分析對象，進行文字的空間定位分析方法研發，包括災害地點現地媒體資料與相關災害訊息蒐集及彙整、資料的前處理、神經網路訓練與測試、標記相關災害地名與地標，最後檢視定位的結果。

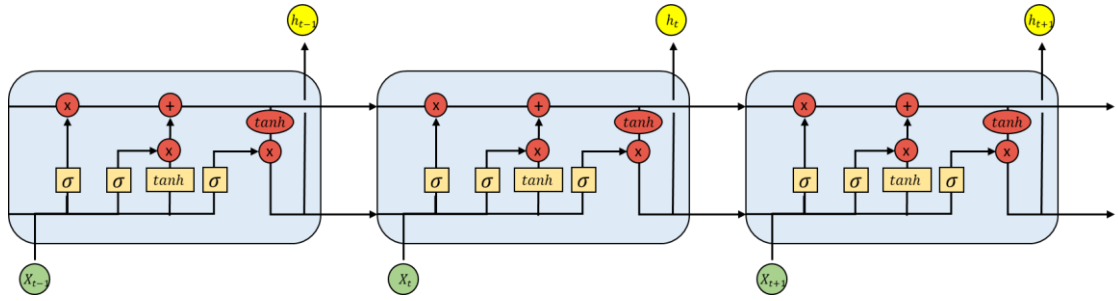


圖 1 LSTM 架構圖

此項目彙整2014年至2019年相關災害事件之災情相片與災情訊息，資料範圍如**錯誤! 找不到參照來源。**所示，包含颱風、豪雨、地震等災害事件，其中以颱風事件居多。

表 1 近六年相關災害時間表

2019年災害		2018年災害		2017年災害	
09/29	米塔颱風	09/14	山竹颱風	10/11	豪雨
08/23	白鹿颱風	09/09	熱帶低壓	09/12	泰利颱風
08/07	利奇馬颱風	08/23	熱帶低壓水災	09/06	谷超颱風
07/17	丹娜絲颱風	07/09	瑪莉亞颱風	08/21	天鴿颱風
05/20	豪雨	06/13	豪雨	07/28	尼莎暨海棠
04/18	地震	02/06	地震	06/13	颱風
				06/01	豪雨
2016災害		2015災害		2014年災害	
09/26	梅姬颱風	09/27	杜鵑颱風	09/19	鳳凰颱風
09/16	馬勒卡颱風	08/06	蘇迪勒颱風	07/21	麥德姆颱風
09/13	莫蘭蒂颱風	07/09	昌鴻颱風	06/15	哈吉貝颱風
07/06	尼伯特颱風	07/06	蓮花颱風		
02/06	地震				

社群災害相關文字訊息，主要使用的資料為發生在臺灣的颱風、豪雨、水災之相關災害資料，資料來源包括Facebook、Mobile01論壇、PTT電子佈告欄、ETtoday新聞雲、蘋果日報即時新聞、中時電子報、工商時報、聯合新聞網、自由時報、伊莉討論區等相關文字訊息。

主要的災害事件包括2015至2019年的颱風、豪雨以及水災資料，共45,603筆原始災害資料，將其進行人工標記的動作，主要的準則為人工判定此留言內是否包含地名或地標，若A留言內有包含地名或地標，則將其人工標記為「有」， B留言內無包含地名或地標，因此將此篇留言人工標記為「無」，14,731筆的資料留言中包含地名或地標；30,872筆的資料留言中沒有包含地名或地標。將原始資料分為80%的訓練資料集和20%的驗證資料集，並使用10,000筆測試資料讓模型測試，2,578筆的資料留言中包含地名或地標；7,422筆的資料留言中沒有包含地名或地標，詳如表2所示。

表 2 LSTM測試結果

實際情況 / 預測結果	留言有地名、地標	留言無地名、地標
留言有地名、地標	2,147	202
留言無地名、地標	431	7,220

持續分析不同社群來源，我們挑選Facebook、PTT、Plurk為來源社群網站的相關災害資料進行收集，使用45,603筆原始資料輸入到神經網路模型訓練，使用80%訓練資料和20%筆驗證資料，並對三組不同的資料來源分別收集1,000筆原始資料進行測試。如圖2所示，在第一組Facebook預測資料集中，共有1,000筆測試資料，其中留言包含地名、地標的有292筆，而無包含地名、地標的有708筆。在第二組PTT預測資料集中，共有1,000筆測試資料，其中留言包含地名、地標的有275筆，而無包含

地名、地標的有725筆。在第三組Plurk預測資料集中，共有1,000筆測試資料，其中留言包含地名、地標的有137筆，而無包含地名、地標的有863筆，表2為三組不同的災害資料來源之預測比較。

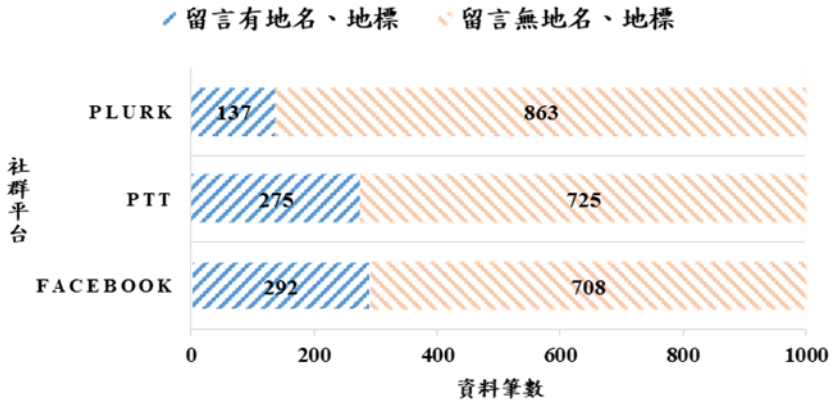


圖 2 不同網站之有效留言數比較

2.2 影像辨識

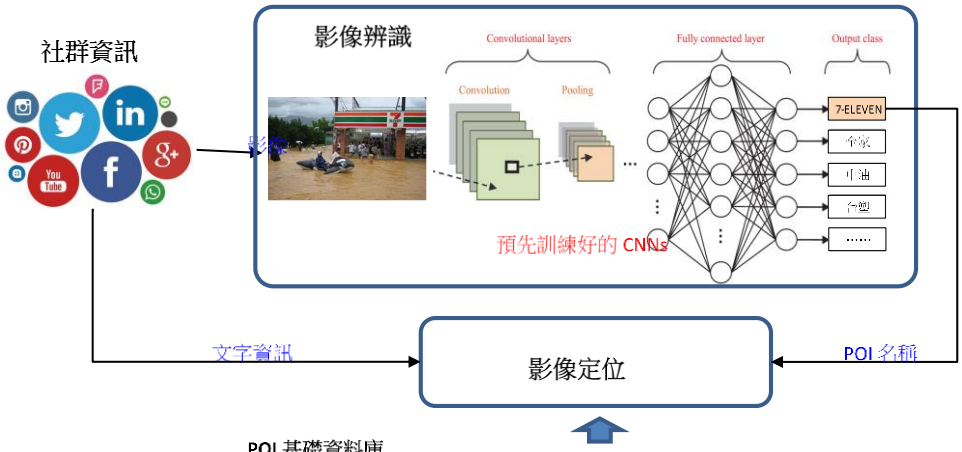
影像辨識(image recognition)的主要是以電腦自動根據影像灰值的分布及變化進行感興趣物件的判識(object identification)，其應用包含影像分類(image classification)、人臉判識(facial recognition)、移動偵測(motion detection)等。在現今非常熱門的先進駕駛輔助(advanced driver assistance)相關研究中，影像自動辨識亦扮演相當重要的角色，如自動駕駛(self-driveing)、車道偵測(lane detection)、行人偵測(pedestrian detection)、及自動停車(autonomous parking)等皆依賴影像自動判識來完成。

近幾年來隨著訊號及影像處理技術的發展，影像自動辨識的能力也越來越強，從傳統需以人工知識進行輔助的半自動辨識，到當今則已經發展出以機器學習(machine learning)技術進行全自動的影像辨識，其中，人工神經網路是機器學習中的一個重要演算法，在眾多深度學習方法中，卷積神經網路(convolutional neural networks, 簡稱ConvNet或CNNs)可以說是最受矚目，也已被廣泛使用的的一個方法。

在神經網路的研究歷程中，ConvNets技術並非是一種嶄新的概念，早在1980年代，已有許多研究提出ConvNets的架構及相關應用，但因為ConvNets通常需要大量的訓練資料，以及較長的計算時間，故相關研究工作及應用較為缺乏。一直到2000年代中期，因為電腦硬體的進步（尤其是GPU的發展）、資料大量的累積（即大數據）、以及許多改進過的演算法，才使得CNNs成為目前神經網路技術中最熱門的技術之一，且成功地應用在影像辨識中。

在人臉的偵測及定位上，以CNNs為基礎的相關研究亦有許多不錯的成果。在衛星遙測影像(remote sensing images)的應用上，Cheng等人(2016)提出一個具旋轉不變性的CNN模型(RICNN)，克服了不同時期、不同角度衛星影像所造成物體偵測的問題。Long等人 (2017) 則利用CNNs於高解析度衛星影像上進行油槽、飛機、交流道等物體的定位及偵測，獲得相當高準確性的成果。

基於上述CNNs於影像辨識上的發展及應用成效，本計畫主要以CNNs為基本研究方法，評估其對災害地點現地照片進行影像辨識的可行性，並與本計畫所蒐集的地標資料進行比對，以完成災害地點的定位。



POI 基礎資料庫

POI ID	類別	名稱	別名	經度	緯度	地址
1	超商	7-ELEVEN	7-ELEVEN 溫州門市	121.530418	25.017671	106台北市大安區羅斯福路三段245號
2	超商	7-ELEVEN	7-ELEVEN 大坪林門市	121.540573	24.983211	231新北市新店區民權路18號
3	超商	全家	全家便利商店台大二活店	121.536089	25.013946	106台北市大安區羅斯福路四段85號
4	超商	全家	全家便利商店新店寶元店	121.541072	24.983364	231新北市新店區民權路19號
6	加油站	中油	台灣中油新店寶橋站	121.537406	24.981288	231新北市新店區寶橋路219號
7	加油站	中油	臺灣中油萬隆站	121.540124	25.005418	116台北市文山區興隆路一段13號
9	加油站	台塑	台亞石油(新店中興路站)	121.543907	24.979987	231新北市新店區中興路三段131號
10	加油站	台塑	台塑政大加油站	121.519886	25.014234	116台北市文山區指南路一段53號

圖 3 影像辨識應用於輔助定位示意圖

本項目的研究課題為：從影像中辨識出特定關鍵字後，再從POI資料庫中找出該影像內容的地理空間坐標，現階段以知名超商做為測試目標，所辨識的超商包含統一超商(Seven_Eleven)、全家便利商店(FamilyMart)、大潤發(RTMart)、以了及全聯福利中心(PXMart)為辨識目標，各家超商的商標(logo)如圖 44所示，各商標圖底下所示的文字為其標籤(label)，而圖5為根據辨識模型依照圖片上LOGO字樣的分類結果。

Seven_Eleven	FamilyMart	RTMart	PXMart

圖 4 不同超商的商標及標籤

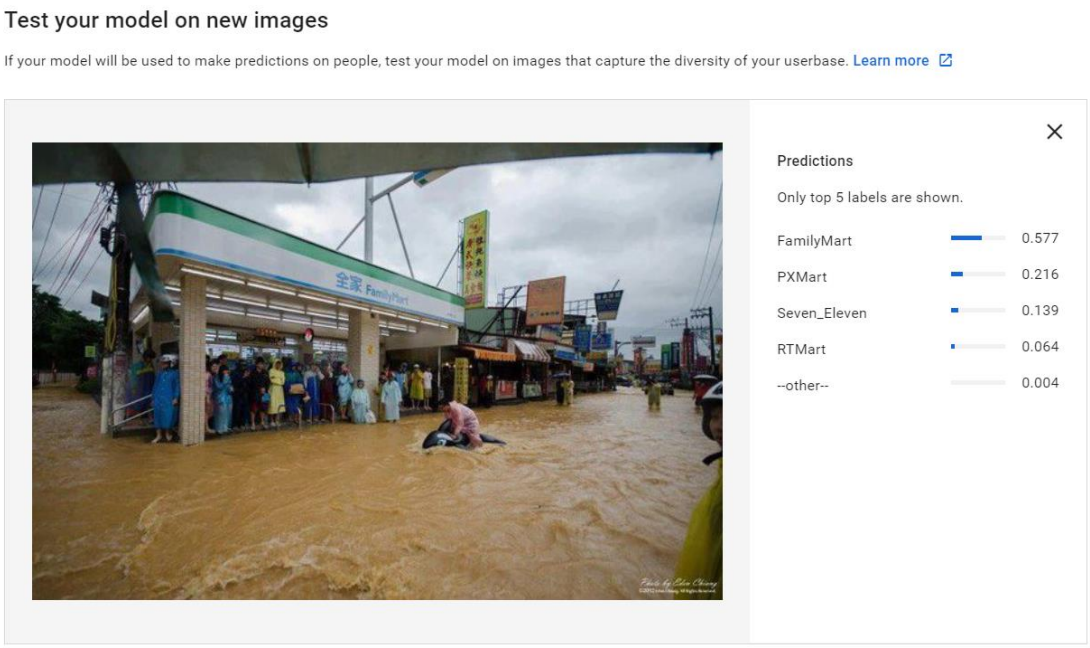


圖 5 辨識結果範例

圖6所示為本次實驗的訓練結果，在監督式的機器學習方法中，一般會將已知樣本分成獨立的三個部分，即訓練集(train set)，驗證集(validation set)、以及測試集(test set)。其中訓練集為訓練模型的主要資料集；驗證集則用來輔助訓練集，以確定網絡結構或者控制模型複雜程度的參數，同時避免模型有過度擬合的現象；而測試集則用來測試最終完成訓練的模型性能，測試集並不參與訓練過程。

我們挑選80%的樣本進行訓練，10%用於驗證、剩下10%用於測試。圖7所示為本實驗之訓練結果，其中精確率(precision)為95.455%，表示誤判的機率很低，而召回率(recall)則高達91.304%，表示漏判的機率約為8.696%。

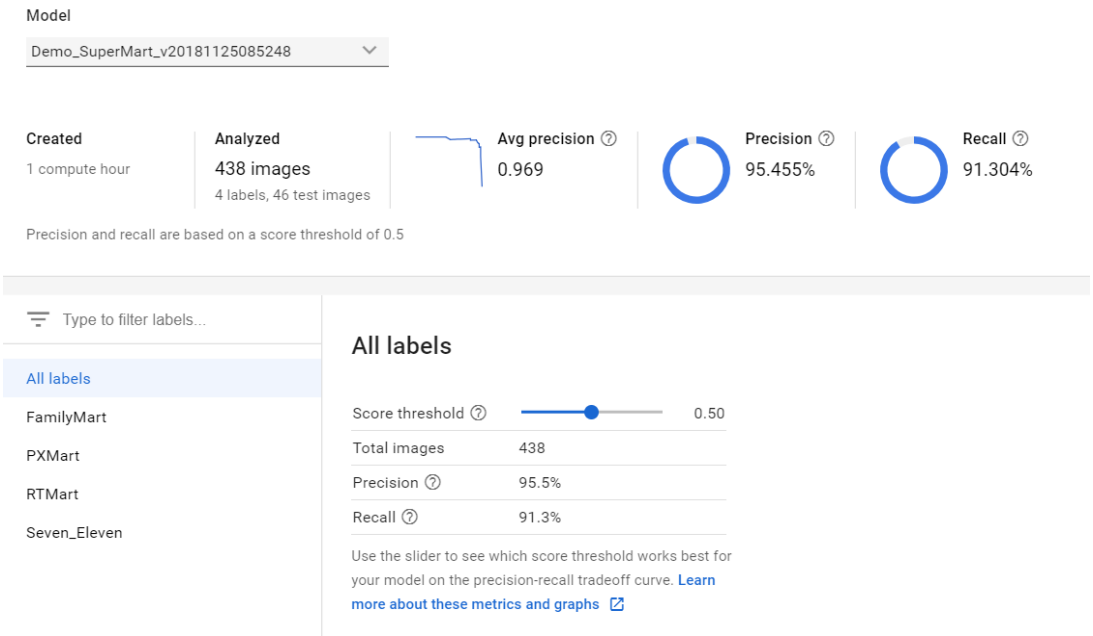


圖 6 完成訓練的模型評估結果

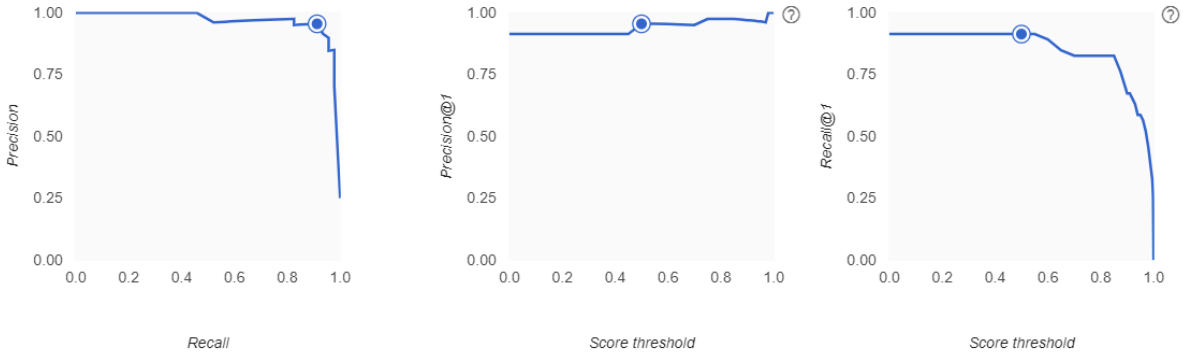


圖 7 精確率與召回率曲線

三、研究成果

我們使用 109 年度 0522 豪雨事件作為本次之應用範例，資料選用範圍如下所示：攀爬目標來源：PTT 批踢踢實業坊、時間範圍：109 年 5 月 22 日至 5 月 23 日。

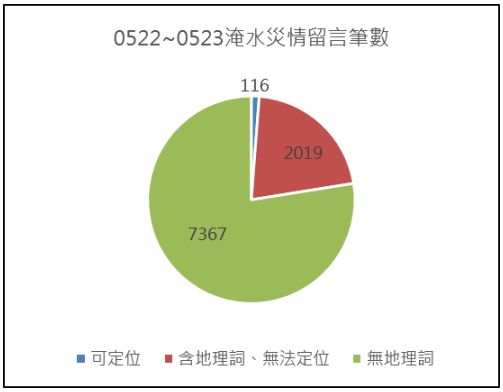


圖 8 豪雨事件於 PTT 留言分布

圖 9 豪雨事件文字雲

共蒐集9502筆留言，含地理詞彙留言數共2019筆，為含有行政區、道路名、地標等敘述之留言者;具備可轉換為GPS座標 (如道路名、地標等) 可定位之有效筆數為116筆，占總數之1.57%，以實務應用的數量為相當稀少，同時也反映民眾在敘述

事件上的可靠性與精確性，此乃自由文本(Free Text)上一項重大的弱點所在。

繼續將這些有效點位資訊，應用於Esri的ArcMap呈現時，我們可以得到一張民眾版的災情輿論展示模型(如圖10所示)，綠色表示民眾的輿論點位，紫紅色則為淹水感測器示警的點位。由此可觀察出：民眾在台南、高雄討論熱度較高，尤其高雄多集中於西部，以仁武、楠梓、三民、苓雅區為大宗。

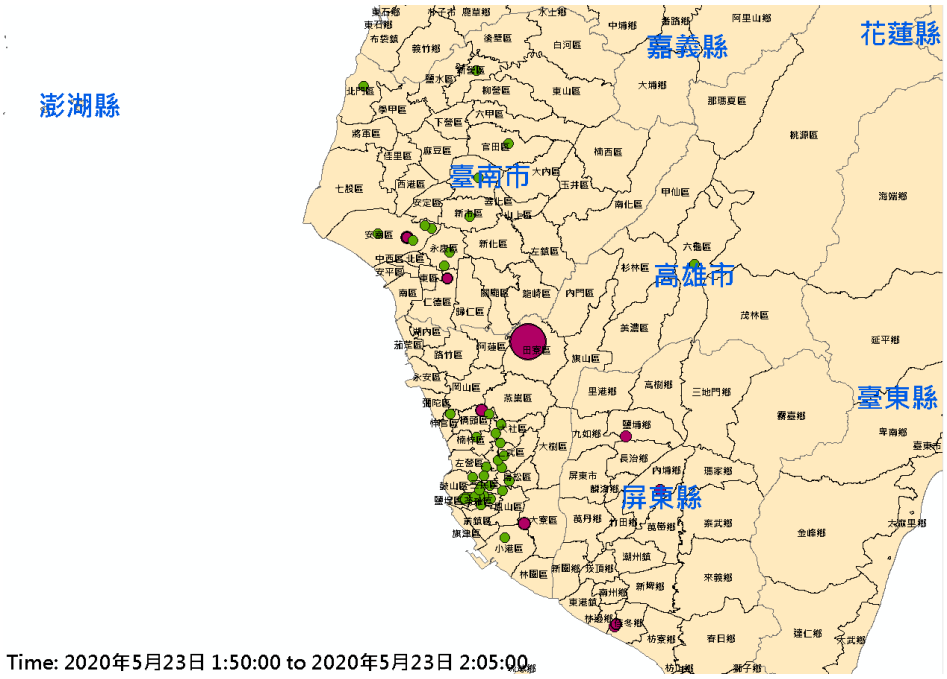
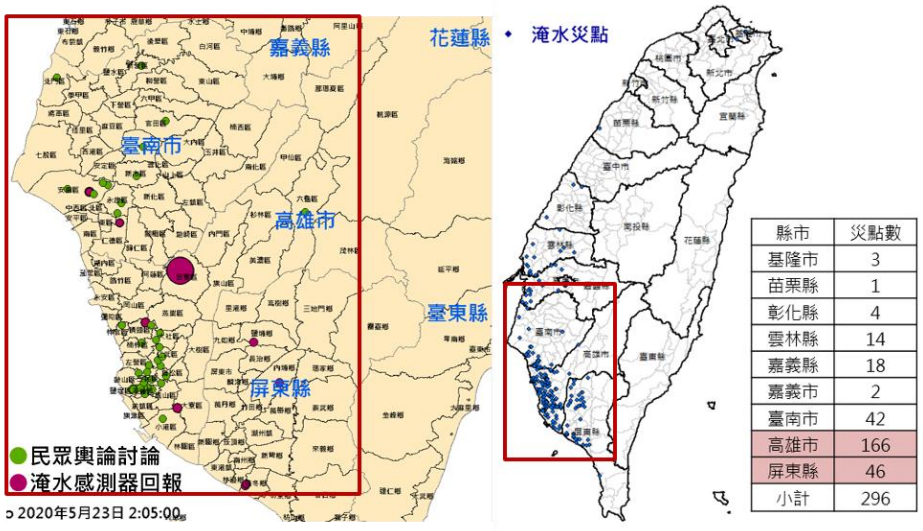


圖 10 社群資料與淹水感測器分布比較

此次事件，災防科技中心亦紀錄於「災害事件簿」之中，我們將災害事件簿之「0522豪雨事件淹水災點分布」與其比較(圖16)，因為在前述的定位流程上，已證明出民眾在地理描述上的不精確性，只能放大觀察範圍，以行政區的尺度進行災情區域的比較，我們定義重疊率公式為：社群輿論與災害事件簿之行政區重疊數量，除以災害事件簿的判定之行政區數量。若套用至本次事件的台南、高雄：台南的行政區重疊率為30%、高雄則為46.7%。



縣市	災點數
基隆市	3
苗栗縣	1
彰化縣	4
雲林縣	14
嘉義縣	18
嘉義市	2
臺南市	42
高雄市	166
屏東縣	46
小計	296

圖 11 社群定位資料與災害事件調查資料

四、結論與建議

本計畫針對社群媒體文字訊息、影像進行定位分析研究，導入社群網路上所散播的災害訊息資訊，利用社群媒體中，民眾所上傳的文字敘述進行災情的研判，並制定文字資料分析流程，以利未來於實際災害發生時，可利用社群網路上的即時災害文字資訊進行災害的初步位置定位，以輔助災害應變的決策支援。

使用長短期記憶模型(LSTM)當作主要的神經網路模型，以社群網站收集新聞以及社群言論為分析對象，進行文字的空間定位分析方法研發，首先蒐集及彙整災害地點現地媒體資料與相關災害訊息，接著進行資料前處理，包括人工標記資料、文本分詞、數字序列的轉換和處理，將資料輸入神經網路進行訓練與測試、標記相關災害地名與地標，最後檢視定位的結果。

影像辨識則由卷積神經網路(CNN)為訓練模型，利用社群媒體中，民眾所拍攝的影像來進行災情的研判，並制定影像分析流程，以利未來於實際災害發生時，可利用社群網路上的即時災害影像資訊進行災害的初步位置定位，以輔助災害應變的決策支援。本年度已順利完成相關文獻與方法彙整、POI基礎資料庫蒐集與彙整、災害地點現地相片與相關災害訊息蒐集及彙整，以及空間定位方法研發、流程設計與定位結果驗證等工作項目，研究結果顯示利用機器學習演算法確實可以進行災害影像的分析及辨識，其正確率皆高達90%以上，其結果可與其他關鍵字結合後進行影像定位及災情位置的展示。

最後再以109年0522豪雨作為實際應用案例，以PTT為資料來源對象，分析5月22日、23日民眾留言內容，分析文字並轉為文字雲呈現，將其定位結果結合水利署淹水感測器進行分布比較，同時，以本中心災害事件簿紀錄之0522豪雨事件，比較其淹水災點統計與分布，作為參考依據。

由此可以從社群網路的角度蒐集大量資料，唯其可靠性與精準性仍為一大問題，在本計畫的未來工作，需考慮如何提升地理詞的辨識精準度、圖片辨識的精確度、招牌種類的廣泛度，以及最終能否有效定位在地圖上，對於進一步的災情掌握為主要難題所在。

參考文獻

1. Adeva, J. G., Atxa, J. P., Carrillo, M. U., & Zengotitabengoa, E. A. J. E. S. w. A. (2014). *Automatic text classification to support systematic reviews in medicine*. 41(4), 1498-1508.
2. Aipe, A., Mukuntha, N., Ekbal, A., & Kurohashi, S. (2018). *Deep Learning Approach towards Multi-label Classification of Crisis Related Tweets*. Paper presented at the Proceedings of the 15th ISCRAM Conference.
3. Bafna, P., Pramod, D., & Vaidya, A. (2016). *Document clustering: TF-IDF approach*. Paper presented at the 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT).
4. Barbier, G., & Liu, H. (2011). *Data mining in social media.* In Social network data analytics (pp. 327-352): Springer.
5. Blunsom, Phil. "Hidden markov models." *Lecture notes, August* 15 (2004): 18-19.
6. Cheng, G., Zhou, P., & Han, J. (2016). *Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images.* IEEE Transactions on Geoscience and Remote Sensing, 54(12), 7405-7415. doi:10.1109/TGRS.2016.2601622
7. Chou, Chien-Lung, Chia-Hui Chang, and Ya-Yun Huang. "Boosted Web Named Entity Recognition via Tri-Training." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 16.2 (2016): 10
8. Dertat, A. (2017). *Applied Deep Learning - Part 4: Convolutional Neural Networks*. Towards Data Science. Retrieved from https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2
9. Graves, Alex, Abdel-rahman Mohamed, and Geoffrey Hinton. "Speech recognition with deep recurrent neural networks." *Acoustics, speech and signal processing (icassp), 2013 ieee international conference on*. IEEE, 2013.
10. Graves, Alex, et al. "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks." *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.
11. Graves, Alex. "Generating sequences with recurrent neural networks." *arXiv preprint arXiv:1308.0850* (2013)
12. Huang, C.-M., Chan, E., & Hyder, A. A. (2010). Web 2.0 and internet social networking: A new tool for disaster management?-lessons from taiwan. BMC medical informatics and decision making, 10(1), 57.

13. Kryvasheyeu, Yury, et al. "Performance of social network sensors during Hurricane Sandy." PLoS one 10.2 (2015): e0117288.
14. Kryvasheyeu, Yury, et al. "Rapid assessment of disaster damage using social media activity." *Science advances* 2.3 (2016): e1500779.
15. Lafferty, John, Andrew McCallum, and Fernando Pereira. "Conditional random fields: Probabilistic models for segmenting and labeling sequence data." (2001): 282-289.
16. LeCun, Y., Bengio, Y., & Hinton, G. J. n. (2015). Deep learning. *521*(7553), 436-444.
17. Liu, P., Qiu, X., Chen, X., Wu, S., & Huang, X.-J. (2015). Multi-timescale long short-term memory neural network for modelling sentences and documents. Paper presented at the Proceedings of the 2015 conference on empirical methods in natural language processing.
18. Long, Y., Gong, Y., Xiao, Z. h., & Liu, Q. (2017). *Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks*. IEEE Transactions on Geoscience and Remote Sensing, 55(5), 2486-2498. doi:10.1109/TGRS.2016.2645610
19. Ma, Xuezhe, and Eduard Hovy. "End-to-end sequence labeling via bi-directional lstm-cnns-crf." *arXiv preprint arXiv:1603.01354* (2016).
20. Nguyen, D. T., Joty, S., Imran, M., Sajjad, H., & Mitra, P. J. a. p. a. (2016). Applications of online deep learning for crisis response using social media information.
21. Olah, C. (2015). Understanding lstm networks.
22.
23. Pennington, Jeffrey, Richard Socher, and Christopher D. Manning. "Glove: Global vectors for word representation." *EMNLP*. Vol. 14. 2014.
24. Sakaki, Takeshi, Makoto Okazaki, and Yutaka Matsuo. "Earthquake shakes Twitter users: real-time event detection by social sensors." *Proceedings of the 19th international conference on World wide web*. ACM, 2010.
25. Y. Y. Huang, C.H. Chung, “A Tool for Web NER Model Generation Based on Google Snippets,” Proceedings of the 27th Conference on Computational Linguistics and Speech Processing, pp. 148–163, ROCLING, 2015.
26. Yubo, C., Liheng, X., Kang, L., Daojian, Z., & Jun, Z. (2015). Event extraction via dynamic multi-pooling convolutional neural networks.
27. Zhang, X., Zhao, J., & LeCun, Y. (2015). *Character-level convolutional networks for text classification.* Paper presented at the Advances in neural information processing systems.