|  |  |  |
| --- | --- | --- |
| **班級：**碩資工一甲 | **學號：**11363123 | **姓名：**葉政杰 |
| **題目：**Future Insights: Harnessing AI and Social Media for Advanced Event and Epidemic Forecasting | | |
| **日期：**2025 / 06 / 03 | | |
| **演講者：**美國 維吉尼亞理工大學 資訊系 呂昌田 教授 | | |
| **關鍵字：**Dynamic Keyword Expansion、Spatial Clustering、Geocoding、Flu Forecasting、Multitask Learning、Lead Time、Supervised、Unsupervised | | |

1. 研究動機與背景

這次演講旨在開發一套結合社交媒體 (如 Twitter) 資料與人工智慧 (AI) 技術的預測系統。此系統可用於即時掌握與預測各類社會不穩定事件，包括抗議、暴動、政治衝突、難民遷移等。此外，也進一步應用於公共衛生領域，協助進行疫情爆發的早期偵測與預測，例如流感或 COVID-19。該系統已獲美國國務院採用，應用於外交政策與國安風險評估。

1. 系統架構與預測流程
   1. 資料來源

該系統的核心資料來源涵蓋各類公共與即時可取得的資料，主要包含：

1. **Twitter 帖文：**用於觀察社會情緒與事件動態的第一手訊息來源。
2. **GSR (Gold Standard Report)：**由具權威性的新聞媒體提供的實際事件報導。
3. **Google Trends / OpenTable / Facebook：**作為搜尋趨勢與人流活動的間接指標。
4. **天氣與地理資料：**用於強化模型對特定地理與氣候條件下事件發生的理解。
5. **疫情監測資料：**如 CDC (美國疾病管制與預防中心) 報告。
   1. 預測流程

預測流程以多階段方式進行：

1. **關鍵字選擇：**每日由領域專家提供或系統自動選擇政治或疫情相關的種子關鍵字作為搜尋起點。
2. **動態關鍵詞擴展 (Dynamic Keyword Expansion, DQE)[1]：**利用多輪次的演算法自動擴展出與種子詞語意相近且具事件代表性的詞彙集合。
3. **語意分析與群體語料萃取：**從社群媒體中萃取使用者語意與關聯性高的貼文，進行次級語料建構。
4. **地點與時間預測：**結合自然語言處理與地理編碼技術判定事件可能發生的地點與時間。
5. **模型預測與模擬整合：**疫情預測部分結合 SIR 模型等模擬方式與即時貼文資料進行融合。
6. 核心技術與方法
   1. Dynamic Query Expansion (DQE)

此方法為一種動態語義擴展技術。系統不依賴固定關鍵字，而是以少量種子詞 (如 protest、corruption) 為起點，在當前時間窗內 Twitter 資料中自動搜尋語義相近且高頻率的詞彙。經由迭代式地搜尋與分析，產生當日最具代表性的事件相關詞集。此方式克服了語言非正式性、詞彙變異快等社群語言的挑戰。

* 1. 地點預測與空間聚類 (Spatial Clustering)[2]

從貼文中使用 NLP 技術擷取地名實體，經由 geocoding[3] 解析轉換為經緯度座標，並藉由兩項原則進行群集分析：

1. Twitter 網路連結度高的使用者分群
2. 空間鄰近的地理單位聚類。這樣的方式能夠對不同城市的事件進行有效區分與預測。
   1. 疫情預測模型融合 (Flu Forecasting)[4]

該系統將社群媒體資料與傳統疫情模擬模型結合，建構一個雙重路徑的預測架構。其中考慮以下四種 Loss Function：

1. **Supervised Loss：**傳統監督式學習中的損失函數，衡量模型預測與真實標註的差異。
2. **Bispace Inconsistency Loss：**評估 social media data 與模擬模型間的預測差異，促進兩者一致性。
3. **Infectious Period Loss：**適用於感染病的時間分布建模，模擬傳染期長短。
4. **Temporal Pattern Loss：**建立疾病爆發時間序列的合理性與連貫性。
   1. 多任務學習 (Multitask Learning)[5]

每個城市被視為一個獨立的預測任務 (Task)。所有模型共享特徵表示層 (Feature Weights) 以提取共性，同時保留區域間差異。此方式解決了大城市資料多、小城市資料稀疏、模型難以訓練的問題。

1. 評估指標與成效

**評估方式：**由獨立第三方 MITRE 組織評估，採用 GSR 新聞作為標準答案。

**評估指標：**

1. **Quality Score (品質得分)：**滿分為 4 分，考量是否準確預測了地點、時間、事件類型、涉入人群等要素。
2. **Lead Time (提前預測日數)[6]：**系統能在事件發生前幾天成功發出警示，最終能提前達 9 天。
3. **Recall (召回率)：**所有實際發生的事件中，有多少比例成功被系統預測出來 (例：0.79)。
4. **Precision (精確率)：**系統預測事件中，實際發生的比例 (例：0.87)。
5. **Probability (信心度)：**模型對事件發生的機率評估。
6. 系統成果與案例
7. 成功預測多起南美洲社會動盪事件 (如委內瑞拉、智利、哥倫比亞)。
8. 初期小城市預測較弱，透過 Multitask Learning 模型顯著改善。
9. 疫情方面，社群資料預測結果與 CDC 官方資料曲線趨勢相符，可提前兩週偵測高風險區域。
10. 分析結論與啟示
11. **結合社交媒體與AI技術可達成即時事件預測：**相較於傳統媒體或統計資料有延遲問題，Twitter 等平臺具備即時反應特性，可作為社會情緒感知的重要來源。
12. **動態關鍵詞擴展 (DQE) 為事件語義建模關鍵技術：**藉由無監督學習與多輪次擴展機制，即時調整語意詞表，能提升召回率與模型敏感度。
13. **社交媒體資料與流行病模擬模型互補互強：**社群資料即時但雜訊多，模擬模型有科學基礎但延遲高，兩者互補可提升整體預測準確度與穩定性。
14. **多任務學習有效處理異質地區間的預測差異**：同時建構多地模型並共享關鍵參數，有助於小城市在資料不足時仍能進行有效預測。
15. **模型融合為提升低資料區預測準確性的關鍵策略：**系統同時整合 supervised[7] 與 unsupervised[8] 模型，在資料稀疏地區可維持高精度，擴大系統適用範圍。
16. 應用潛力與未來發展
17. 可擴展至跨語言、跨文化的全球預警系統：現行以英文、西班牙文為主，未來可延伸至中文、阿拉伯文等語系。
18. 應用於政策部署、國安規劃、公共衛生決策：可作為政府部門、國際組織的風險預警與資源調配工具。
19. 結合其他平臺資料 (如Line、Weibo) 可拓展監測覆蓋層級與群體多樣性。
20. 心得感想

這次演講內容結合了尖端的AI技術與即時社會資料的應用方式，展現了社交媒體在事件預測與疫情監控方面的高度潛力，讓我深刻體認到資料科學在公共政策與國安風險管理中的實務價值。尤其 Dynamic Query Expansion (DQE) 與 Multitask Learning 的技術應用，不僅創新，也具高度可擴展性與實用性。透過具體案例的展示，我能清楚理解不同模型如何在有限資料條件下進行調整與融合，進而提升預測效率。

此外，呂昌田教授也分享其研究團隊與學校在人才培育上的理念與實務策略，提供了關於留學、研究、實習與未來就業方向的多重視角，讓我對跨國學術合作與未來職涯方向有了更多啟發與反思。

1. 參考文獻

[1] T. T. Yuan, and Z. Zhang, "Merchandise recommendation for retail events with word embedding weighted Tf-idf and dynamic query expansion," *41st International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2018.* pp. 1347-1348.

[2] S. Yang *et al.*, “Regional spatial econometric Analysis of carbon footprint of energy consumption based on clustering algorithm,” *Applied Mathematics and Nonlinear Sciences,* vol. 9, no. 1, 2024.

[3] E. Hristov *et al.*, "Geospatial Data Enrichment through Address Geocoding: Challenges and Solutions," *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives.* pp. 239-245.

[4] A. Rodriguez *et al.*, "Steering a Historical Disease Forecasting Model Under a Pandemic: Case of Flu and COVID-19," *35th AAAI Conference on Artificial Intelligence, AAAI 2021.* pp. 4855-4863.

[5] A. Arefeen, and H. Ghasemzadeh, “Cost-Effective Multitask Active Learning in Wearable Sensor Systems,” *Sensors,* vol. 25, no. 5, 2025.

[6] A. Debnath, and B. Sarkar, “SUPPLY CHAIN MODEL HAVING STOCHASTIC LEAD TIME DEMAND WITH VARIABLE PRODUCTION RATE AND DEMAND DEPENDENT ON PRICE AND ADVERTISEMENT,” *RAIRO - Operations Research,* vol. 58, no. 4, pp. 2645-2667, 2024.

[7] Y. Zhang *et al.*, “DAWN: Domain-Adaptive Weakly Supervised Nuclei Segmentation via Cross-Task Interactions,” *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 35, no. 5, pp. 4753-4767, 2025.

[8] H.-W. Wang, and R.-T. Wu, “Unsupervised anomaly detection for tile spalling segmentation using synthetic outlier exposure and contrastive learning,” *Automation in Construction,* vol. 170, 2025.