Future Insights: Harnessing AI and Social Media for Advanced Event and Epidemic Forecasting

演講者

Virginia Polytechnic Institute

and State University

Chang-Tien Lu 教授

記錄人

碩資一甲

11363156

張晉源

心得

這次的演講聽起來非常的有趣，第一次聽到可以透過人工智慧和 twitter 去預測即將發生的事情，並且要發生的事情的準確率還不低，從投影片中可以看到，可以預測的包括社會事件、疫情爆發，這樣透過持續自動化的分析公開資料，能提前預警從政治危機到天災反應等多種大規模事件。在資料處理的方面也有提到兩個問題第一個問題是沒有訓練集怎訓練，和小城市資料比大城市少很多怎辦，可以用預訓練語言模型提取推文語義，再透過偽標籤 + 半監督式學習，用一個小模型幫沒標記的推文貼上臨時標籤，這樣就有一組基礎的訓練資料，然後把人多、資料多的大城市的資料和人少、貼文少的小城市資料放在同一個大模型裡面一起學，共享底層特徵若，如果還不夠，利用地理鄰近關係或地方新聞、Google Trends 作為輔助訊號，結合各方信息來提高模型對小城鎮事件

或疫情的預測能力。藉由這樣收集資料集的方式讓我學到了，當沒有資料集的強況下怎麼去收集資料

Keyword

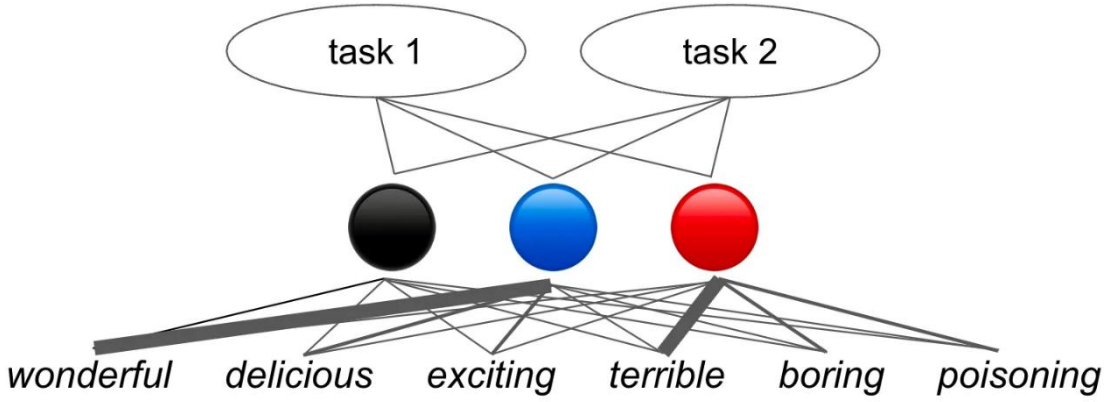
1. Tweet Network Clustering（推文網絡分群）

為了將大量推文 / 使用者依照主題或行為特徵進行分組，方便後續做行為分析、用戶畫像或事件偵測。

1.1 實作方法

先把推文裡的雜訊像 @、#、網址、標點都清掉，再把剩下的詞拆開、還原到基本詞形，接著看一條推文裡出現哪些跟經濟、社交、文化、健康有關的關鍵字，算出跟哪個主題最像，就把它歸到那個類別；然後把同一個人的推文按照這樣的分類統計出他在每個主題裡一共發了多少推文，最後拿這些數字當「特徵」，用 KMeans 把人分群，這樣就能快速找出那些對同樣議題很關心、發文行為相近的一群人。

2. Multi-Task Learning（多任務學習）



讓一個模型同時學習好幾件相關的事，像是把不同城市的抗議預測一起丟進同一個網路，讓底層先學大家共同會用到的訊號，Ex 關鍵字突增代表抗議，然後每個城市再用自己專屬的小層去微調，這樣不只能共享彼此的經驗，還能照顧到每個城市的特殊需求，避免某個小城市資料太少反而學不好。

3. One-for-All 模型

把所有地區當成多個子任務 tasks，將它們的資料同時丟進同一個共享的底層網路進行特徵學

習，再針對每個地區額外分出一小層分支做獨立微調。底層參數讓大城市和小城市都可以共享推文語意、時序趨勢、空間關聯等共通表示，而每個城市的上層專屬分支則負責捕捉該地區特殊的抗議或疫情信號。這種一對多的多任務學習架構既能利用大數據地區的豐富樣本幫助小地區擺脫資料稀疏的限制，也同時保留各地獨有的差異性，從而在全域與在地特徵之間取得平衡，提高事件預報的準確。

4. SEIR Model

SEIR 模型是一種將人口分成易感者（Susceptible）→ 潛伏者（Exposed）→ 感染者（Infectious）→ 移除者（Recovered/Removed）四個階段的傳染病數學模型。首先，易感者在與感染者接觸後，以一定的傳染率 β 轉入潛伏者階段；潛伏者以速率 σ轉為具有傳染力的感染者，感染者則以康復或死亡率 γ 移出到移除階段，

表示不再具有傳染性。透過這四個區隔與參數設定，SEIR 模型能模擬疾病在社區中的傳播曲線，並評估如隔離、疫苗接種等干預措施對疫情峰值、持續時間以及最終感染人數的影響。

4.1 實作方法

在 Python 裡面用 SciPy 的 ODE 求解器來實作 SEIR 模型：先用 NumPy 定義參數β、σ、γ和初始人口分布 S、E、I、R，再寫一個回傳微分方程組 dS/dt、dE/dt、dI/dt、dR/dt 的函式；接著用 scipy.integrate.odeint 讓這個函式隨時間演進並計算各隔離區的人數變化，最後用 Matplotlib 把 S、E、I、R 四條曲線畫出來。

5. . Pseudo-labeling

在少量有標註訓資料之外，利用已訓練的模型先對大量未標註的測試資料做預測，將信心度很高例如預測機率高於某個門檻的預測結果當作「偽

標籤」加回訓練集中，然後連同原有標註訓資料一起重新訓練模型。這樣做可以在不額外花費人工標註的情況下，大幅增加訓練樣本數，並且透過只挑取高置信度的偽標籤避免錯誤標籤誤導模型，有效提升最終模型的性能。

參考資料

[1] Tweets Classification and Clustering in Python.

https://medium.com/swlh/tweets-classification-and-clustering-in-python-b107be1ba7c7

[2] Multi-task learning: what is it, how does it work and why does it work?

https://medium.com/gumgum-tech/multi-task-learning-what-is-it-how-does-it-work-and-why-does-it-work-294769c457bb

[3] Multi-task Learning for Spatiotemporal Event Forecasting

https://ieeexplore.ieee.org/document/10679926

[4] Mastering the SEIR Model: A Comprehensive Guide with Python Code and Real-World Applications Examples

https://medium.com/pythoneers/mastering-the-

seir-model-a-comprehensive-guide-with-python-code-and-real-world-applications-da7584a4fb23

[5]機器學習動手做 Lesson 1 — 善用 Pseudo Labeling 增加訓練資料集

https://flageditors.medium.com/%E6%A9%9F%E5%99%A8%E5%AD%B8%E7%BF%92%E5%8B%95%E6%89%8B%E5%81%9Alesson-1-%E5%96%84%E7%94%A8pseudo-labeling%E5%A2%9E%E5%8A%A0%E8%A8%93%E7%B7%B4%E8%B3%87%E6%96%99%E9%9B%86-95d7d4617c4f