運用機器學習發展有效之新冠肺炎流行趨勢預測方法

Effective Prediction of Spread Trends of COVID-19 Using Machine Learning

*張昭憲 副教授
淡江大學資訊管理學系
jschang@mail.tku.edu.tw

蔣蕙娟 碩士生
淡江大學資訊管理學系
macaron699@yahoo.com.tw

摘要

近年來 COVID-19 蔓延全球，各國紛紛制定各種防疫措施，以防止其擴散與流行。COVID-19 目前雖暫時流感化，但考量病毒變異快速，仍需持續監控以避免其捲土重來。為降低 COVID-19 所引發影響，研究者莫不投注大量心力，配合機器學習發展有效的預測方法。然而，由於疾病具有潛伏期且發展快速，其實用性與準確性仍有改善空間。有鑑於此，本研究以網路社群發文為分析來源，以機器學習方法建立具有延遲特性之預測模型。過程中，我們整理各種與症狀相關之關鍵詞，藉以將社群發文進行轉換。而後，再以潛伏期天數與資料回溯天數為基礎，建立更符合實際狀況之資料集與預測模型。為驗證提出方法之有效性，本研究蒐集美國 CDC 公布之 COVID-19 統計數據，並擷取期間的 Twitter 發文進行實驗。結果顯示，以適當的關鍵詞集分析社群發文，確實可提供高關聯度之流行趨勢預測。當考量潛伏期與資料回溯天數時，可更進一步提升預測準確性。綜合以上成果可知，本研究提出方法能提供具參考價值的流行趨勢預測，做為防疫相關單位決策時之重要依據。

關鍵詞：COVID-19、流行病趨勢預測、深度學習、機器學習

Abstract

In recent years, COVID-19 has spread around the world, and countries have formulated various epidemic prevention measures to avoid its spread and epidemic. Although COVID-19 is temporarily treated as like flu, considering the rapid mutation of the virus, continuous monitoring is still required to prevent its resurgence. In order to reduce the impact of COVID-19, researchers have invested a lot of effort in developing effective prediction methods with machine learning. However, because the disease has an incubation period and develops rapidly, there is still room for improvement in its practicality and accuracy. In view of this, this study uses online community posts as the source of analysis and uses machine learning methods to establish a prediction model with delay characteristics. In the process, we sorted out various keywords related to symptoms to convert social posts. Then, based on the number of incubation period days and the number of data lookback days, a data set and prediction model that is more consistent with the actual situation are established. In order to verify the effectiveness of the proposed method, this study collected COVID-19 statistical data released by the US CDC and extracted Twitter posts during the period for experiments. The results show that analyzing social posts with appropriate keyword sets can indeed provide highly relevant trend predictions. When considering the incubation period and the number of data lookback days, the prediction accuracy can be further improved. Based on the above results, it can be seen that the method proposed in this study can provide valuable popular trend predictions and can be used as an important basis for decision-making by epidemic prevention-related units.

Keywords: COVID-19, Epidemic trend forecast, Deep learning, Machine Learning

壹、 緒論

近年來，COVID-19 蔓延全球，使各國蒙受重大損失，對人民生活與經濟發展造成重大影響。COVID-19 是由新型冠狀病毒(SARS-CoV-2)所造成的疾病，自 2019 年 12 月 30 日出現不明原因的肺炎病例後，至今已擴散超過 200 個國家。截至 2023 年 8 月，全球確診人數已高達 6 億 7 千萬人，死亡人數超過 680 萬人。台灣在初期防控得宜，但迄今也有超過 1 千萬人次確診，其中 17,667 人不幸因此死亡(NCHC[a] 2023)。COVID-19 傳染力極強，雖有人感染後症狀輕微甚至無症狀，但有一定比例的人會引發重症(死亡率約為 2%)，特別是年長者或是慢性疾病患者，對於醫療體系是龐大的負擔(衛生福利部 2021)。更令人擔心的是，某些感染者復原後仍有後遺症(肺功能受損，頭痛等)，對其工作、生活造成重大影響。在大流行期間，COVID-19 造成的影響早已遠遠超過每年秋冬出現的流感，成為讓人聞之色變的全球性傳染病(Pandemic)。

隨著病毒的流感化，目前各國對於 COVID-19 大多已解禁，但考量其變異快速(無法排除其捲土重來的可能性)，相關單位仍持續做密切監控。在此之前，面對 COVID-19 所做的處理對策與經驗，必可做為爾後類似疾病的借鏡，有效降低嚴重的生命與經濟損失。回顧 2020~2022 期間，有鑑於 COVID-19 的影響巨大，各國政府均制定各種防疫措施，以避免其擴散與流行。對於台灣而言，由於初期感染人數相對較低，邊境防疫便顯得格外重要。此外，對於本地民眾，政府也需建立適當的管制措置，限制可能導致流行的群聚活動。很明顯地，封鎖措施將對經濟活動造成嚴重影響，讓許多人收入降低甚至失業。為顧及人民生計，各國政府經常得藉由發放現金或補貼，協助大眾度過難關(中央社 2021)。由於 COVID-19 的中重症治療具有高度不確定定，因此疫苗接種便顯得格外重要。接種疫苗可以預防感染、降低重症比率，有效降低感染人數，因此各國政府莫不積極購買足夠疫苗，全力增加疫苗覆蓋率。

由上述討論可發現，新冠肺炎防控不易，付出的醫療、經濟與社會成本相當巨大。此外，COVID-19 流行爆發後，人們因生命與生活受到嚴重威脅而恐慌，網路社群上更流竄各式各樣的假消息。根據研究發現(Mourad 2020)，在新冠肺炎發生後，社群平台充斥大量未經證實的編造訊息，但發現其數量、點閱率與傳播速度遠高於可靠的真實訊息。令人擔心的是，其中只有極少數訊息來自病毒專業人士。探究其原因，假訊息經常提供符合民眾恐慌心態的編造猜測，因此比真實資訊更容易散播。此外，有心人士也經常利用大眾對於新冠肺炎的關注，將其轉移至不相干的主題(政治、宗教、推銷等)，因此樂於協助散播。凡此種種，政府相關單位需同時顧及各種面向，才能對新冠肺炎進行全面的防治。

為協助解決 COVID-19 所引發的影響，研究者莫不投注大量心力，探討如何發展符合成本效益的方法，做為制訂防疫策略時之有效依據。近年來，由於大數據分析與機器學習方法的蓬勃發展，如何將其運用於 COVID-19 的防治便成為注目焦點。使用機器學習方法進行防疫的優點有成本低、速度快、較少牽涉個人隱私，可做為有效的監測輔助工具。相關的應用如醫療資源調配，社交距離偵測，病患生理資訊偵測，及精準疫調等；亦可對 COVID-19 進行輿情分析，或者其擴散過程進行模擬與預測(Yu et al. 2021; Latif et al. 2020)。其中，資源分配涉及醫療、人力、隔離所容量等資源的最佳配置；確診者接觸史則與其個人之實體足跡與網路數位軌跡相關(Wibowo et al. 2021; Su et al. 2021; Latif

et al. 2020)。事實上，COVID-19 的相關研究相當多樣化，例如，由於需進行封鎖以控制感染，政府在制定政策補貼與振興政策時，也需顧及因工作機會的減少導致失業問題。此外，因隔離政策產生的心理問題(沮喪、恐懼、不安等)，也歸納出官方、企業界、教育界需注意的事項(Yu et al. 2021)。

由上述討論可知，研究者已運用機器學習技術對於 COVID-19 的監測提出許多有效方法，使相關單位獲得寶貴的決策資訊。然而，根據前人研究提出的作法，我們發現仍有以下改進的空間: (1) 新冠病毒具有潛伏期，前人研究很少將此納入考慮，可能影響預測結果之實用性。(2)進行社群發文分析時，所使用之關鍵詞集經需要特別考量，而非單純的斷詞處理，以免影響後續預測模型效能。(3)新冠病毒的流行發展快速，資料的採用需能考慮與預測時間點之回溯天數，才能符合實際狀況。有鑑於此，本研究將發展更有效的方法，結合官方 COVID-19 歷史數據與網路社群發文，運用機器學習方法預測 COVID-19 的確診人數與死亡人數。首先，我們蒐集美國 CDC 公布之 COVID-19 相關數據，並截取期間的 Twitter 發文，建構所需之資料集。當對社群發文進行分析時，我們採用多種症狀單詞作為關鍵字。其次，本研究同時考量疾病的潛伏期與資料回溯天數，建立具有延遲特性的預測模型，期能提升預測準確率。最後，再分別使用深度學習方法 MLP、LSTM 與常用的線性迴歸進行塑模。結果顯示當以 Twitter 發文為資料集時，以線性迴歸表現最佳，可獲得 90%以上之預測關聯度。當以疾病歷史數據為資料集時，則以 MLP 表現最佳，可獲得超過 96%之關聯度。上述結果顯示本研究提出方法之有效性，在配合不同機器學習方法時，預測結果可做為制定防疫政策時之重要參考依據。

本論文後續章節如下: 第二節為相關技術介紹與文獻探討；第三節介紹本論文提出之 COVID-19 流行趨勢預測方法;第四節為實驗結果與討論;第五節為結論與未來工作。

貳、 背景知識與文獻探討

以下介紹與本論文相關之文獻探討、背景知識與相關技術，以利後續章節之討論。

一、COVID-19 的監測與防控

COVID-19 藉由近距離飛沫、直接或間接接觸，帶有病毒的口鼻分泌物傳染給他人，造成大規模流行。病毒可以在寒冷低濕度的環境中存活數小時，所以也可能經由間接接觸傳染。感染後會產生發燒、咳嗽、倦怠、頭痛、喉嚨痛、腹痛、部分嗅覺味覺喪失，嚴重會出現肺炎或呼吸衰竭甚至死亡。根據病例分析，兒童或是青少年得病比率較低且重症比率相對較低。值得注意的是，病毒的型態並非一成不變，而是通過變異不斷演化，如英國出現的 Alpha(B.1.1.7)、印度出現的 Delta(B.1.617.2)、南非出現的 Beta(B.1.351)，以及日本、巴西出現的 Gamma(P.1)等。隨著新型變種病毒的增加，2020 年 1 月最初出現原始病毒株已不再流行。這些變種病毒傳播速度比原病毒更快速，也能幫助病毒避開抗體，使防控與治療更加困難。有關 COVID-19 的監測方式，至少有以下二種 :

(1) 傳統監測系統

當新冠肺炎突然出現時，各國政府似乎不知如何因應，也無法制定有效的因應措施。然而，但經過近幾年的發展，相關單位已發展一套頗為完善的監測機制。例如，台灣 2021 年的邊境管制規定: 入境者需進行 14 天的居家檢疫，期間三次檢驗，之後再進行 7 天自主健康管理(衛生福利部, 2021)。本地民眾常見的方式，則當發現確診病例外，除進行通

報確認外，需進行疫調公布足跡，之後框列接觸者進行隔離。上述傳統方法具有一定功效，但所耗成本相當高。舉例而言，當對一特定區域民眾進行採樣時，需投入大量醫護人力資源，之後的核酸檢驗更是成本高昂。當民眾進行隔離時，政府通常也需補貼住宿費用。凡此種種，讓防疫所需的成本隨著時間大量增加。此外，為了顧及隱私與人權，許多防疫措施無法強制執行，需依賴人民自動自發，也造成人力成本的增加(例如疫調)。

(2) 運用機器學習方法進行新冠肺炎監測

由於傳統監測機制成本相當高，學者便開始研究以機器學習方法輔助監測流程。典型以數據分析為基礎的流程為: 資料蒐集、前處理、屬性擷取(Feature Extraction & Selection)、以學習方法建立預測模型，最後以獲得之模型進行預測。如前所述，運用機器學習方法進行監測的優點有成本低、速度快、較少牽涉個人隱私，因此可做為有效的監測輔助工具。若能事先了解疾病發展趨勢，政府便可早期配置各項資源，使後續的發展獲得良好控制。此外，由於網路社群已成為現代人生活的一部分，透過分析其中的發文以了解民眾的情緒，成為政府了解防疫民情的重要依據。事實上，早在 COVID-19 流行之前，學者便已運用機器學習進行疾病監測。例如，針對流感之流行預測，Lampos 與 Cristianini (2010)分析民眾在 Twitter 之發文，以線性迴歸建立預測模型，以改善人工通報系統之效率。Byrd 等人(2016)亦對 Twitter 發文進行探勘分析，結果再次證實社群發文資料可提供流感就診率之有效預測。Bardak 與 Tan(2015)則蒐集 Google Flu Trend 和 Wikipedia Logs 與流感相關文章，分析後依與疾病之關聯性產生屬性集，最後建立以線性公式為基礎之預測模型。

二、COVID-19 流行趨勢預測

對於 COVID-19 流行之相關預測，學者們已提出各種可行的做法。例如，由於網路社群已成為現代人生活的一部分，Naseem 等人(2021)在社群網站中發掘民眾因 COVID-19 所展現的情緒，並歸納出那些主題最容易被討論，並以多種深度學習方法進行文章分類。Wibowo 等人(2021)則透過分析 Twitter 上的留言，了解民眾是否遵守政府的減少外出政策(Stay Home)。由民眾發言萃取出可能外出的關鍵字(如景點名稱)，配合迴歸與分類樹方法進行預測。對政府而言，疫情時期管控不良訊息造成的影響相當重要。有鑑於此，Liu 等人(2021)分析 Weibo, WeChat 等平台，找出其中的關鍵節點並分析其影響力。此外，該研究也歸納民眾討論主題的演變與情緒的變化，藉以抑制社群中的不良訊息，提供政府在回應網路意見時之參考依據。

與前項研究相反，Gupta 等人(2021)透過社群媒體分析探究民眾是否支持政府各項管制措施。透過自然語言處理(NLP)，分析印度民眾在 Twitter 上的發文，配合 Logistic Regression, Linear SVC 等學習方法建立分類器，了解民眾在期間的情緒的變化。除自行蒐集資料外，有許多研究利用公開資料集進行確診、死亡人數之預測。塑模時常見的方法如 Random Forest、SVM、Regression 等(Gupta et al. 2021; Mandayam et al. 2020)，亦有研究使用分群方法歸納確診人數與發病、死亡人數間的關聯(Nikil et al. 2020)。除了預測感染人數外，Hossen & Karmoker (2020)則為了解飲食習慣是否影響患者的康復，蒐集南亞、東亞、美國、英國、巴西等 COVID-19 流行國家的飲食習慣，以 Random Forest 塑模，對 10 個病例的康復進行預測。

由於醫療資源相當寶貴，許多研究者開始研究如何在隔離期間提早偵測出可能的感染者。例如，Alsabek(2020)等人利用自動聲音辨識(ASR)，將聲音轉換為 MFCC 特徵值，再利用關聯分析(Correlation)分辨病人呼吸、咳嗽聲音，協助醫師進行新冠先期診斷。此外，為進行大量的快速診斷，也建議可使用 Chatbot 與病人對話，降低醫護負擔。同樣地，Vrindavanam(2021)等人將病人咳嗽聲音切割為 Frame，分析其波形，頻率、持續時間等聲紋樣式產生 15 種特徵值。之後，使用 Logistic Regression, SVM, Random Forest 進行分類，希望提早找出未發病的感染者，讓隔離的病人提前知道自己病情的變化，也讓一線醫護能有更多的防護。防疫期間，維持社交距離相當重要，被感染的原因往往是與確診者 Too Close for Too Long(TCTL)。有鑒於此，許多研究者開始運用手機藍芽訊號進行社交距離監測。然而，Su 等人(2021)認為之前作法通常只以強度訊號來推估距離，準確率有待提升(導因於各種干擾)。因此，作者以藍芽 RSSI 訊號強度為基礎，再配合 13 種特徵值(空間、時間、統計等)，以 SVM, RF, Gradient Boosted Machines 等方法塑模，最高可提升近 20%之準確率。為解 COVID-19 未來發展趨勢，Hu (2020)提出結合時序分析與 TDA(Topological data analysis)的方法，考量季節性影響，依照月份對 10 個國家進行 COVID-19 的死亡率之發展趨勢預測。

三、結果評量指標

由於本研究著重於趨勢預測，因此將使用皮爾森相關係數(Pearson's correlation coefficient)來評估官方數據與預測值間的關聯度(Correlation)。其公式如下:

$$r_{Y,Y'} = \frac{\sum_{i=1}^{n}(yi-u_Y)(y'_i-u_{Y'})}{n\sigma_Y\sigma_{Y'}} \quad (1)$$

其中 n 為資料集之資料筆數，其中假設官方數據為 $Y=(y_1, y_2, \dots, y_n)$，模型之預測值為 $Y'=(y'_1, y'_2, \dots, y'_n)$， μ, σ 分別為數列之平均值與標準差。關聯度的絕對值越接近 1，表示 Y, Y'之相關程度越高。使用關聯度可克服模型預測結果的刻度(scale)問題，專注在整體趨勢是否一致。例如，若實際資料為 Y=(100,200,300,400,500,600)，而預測值為 Y'=(15, 31, 44, 61, 75, 89), 表面上 Y'與 Y 差異很大，但二者卻有極高的關聯度。換言之，Y'數列確可反映 Y 的趨勢(大致為第一個值之倍數成長)，具有很高的參考價值。由於本研究是以新冠流行趨勢預測為目標，因此會產生與時間相關之預測值序列。若預測序列 Y'與實際序列 Y 之相關係數絕對值越大，顯示 Y'之變化趨勢與實際序列 Y 越相關。這對於流行病趨勢預測特別重要，因可預先了解流行之變化與轉折(例如確診人數之增減)。縱使二個序列之各資料點數值有差距，只要變化趨勢大致一致，便能提供實用的趨勢預測。有鑑於此，本研究採用相關係數做為評估模型效能之指標(Waldmann 2019; Wu et al. 2020; Teixeira & Fernandes 2014)。

參、 以機器學習方法建立新冠流行趨勢預測模型

本節將介紹本研究提出之方法，以建立有效的 COVID-19 流行趨勢預測模型。內容涵蓋資料集蒐集、資料轉換與模型建立，並介紹如何配合回溯天數與潛伏期，建立更符合實際狀況之預測模型。

一、建立資料集

本研究資料蒐集主要來源如下，將做為後續塑模與驗證之用: (1)考量線上社群已成為人們生活一部分，發文內容可能反應 Covid-19 之客觀與主觀發展潛在因素，因此本研究將蒐集 Twitter 社群在美國地區的每日發文(tweets) (2)蒐集美國疾管中心(CDC)提供的歷史案件與死亡總數，做為預測與驗證之用。資料蒐集期間為 2020/8/3~2021/6/18，時間涵蓋美國地區第一次 COVID-19 流行高峰之上升與下降期(NCHC[b] 2023)。選擇使用美國地區資料做為來源的原因如下: 根據聯合國統計目前確診病例數最多的國家是美國(已超過一億人[1])。此外，在美國 Twitter 的使用相當普遍，其發文資料亦可公開取得，因此將其做為資料蒐集來源。以下繼續說明資料蒐集時的細節:

(1) 蒐集 Twitter 每日發文：

Twitter API 提供二種方式來蒐集數據，分別為 Streaming API 和 Search API，Streaming API 用於採集即時發文，Search API 用於檢索歷史資料。根據所允許擷取資料的時間區間與數量，Search API 又分為三種層級，標準、高級和企業。標準模式提供過去 7 天內發布的推文，高級和企業模式則需要昂貴的價錢支付。基於費用考量，本研究使用有過濾功能的 Free Streaming API。使用 Twitter API 需 Twitter 批准開發人員帳戶，使用活動密鑰(active keys)和符記(tokens)就可以開始蒐集數據[2]。Streaming API 從平台上採集即時公開數據，可獲得推文(tweets)中的基本指標，例如使用者名稱、主要內容、轉發人數、喜歡人次等。此 API 會過濾使用者指定國家的推文，美國地區每日約可獲得 6 千筆。

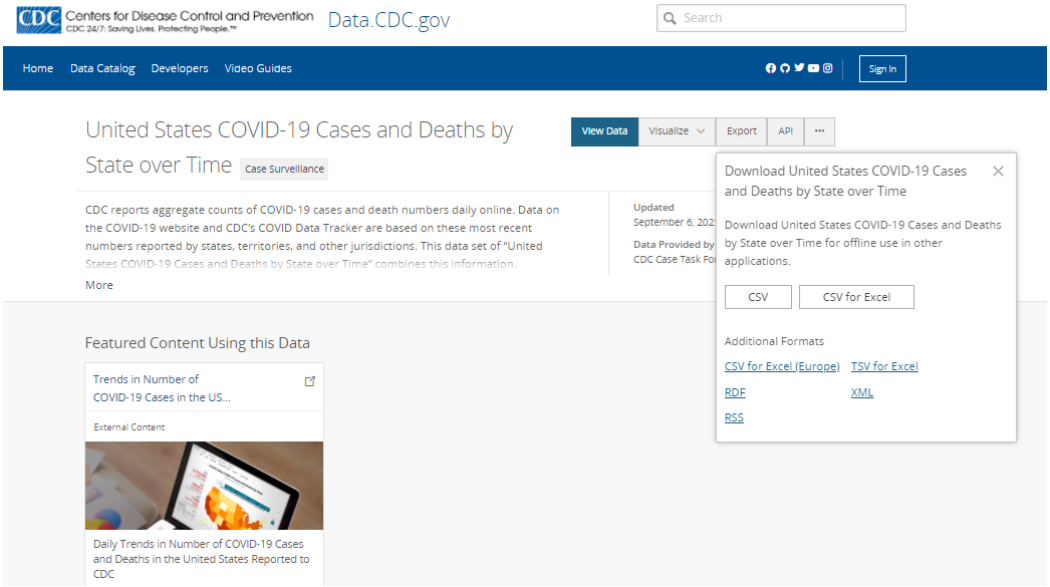


圖 1：美國 CDC 有關 COVID-19 之統計數據

(資料來源: https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36)

(2) 蒐集 COVID-19 案件數與死亡數：

本研究由美國 CDC(Centers for Disease Control and Prevention)下載 COVID-19 之歷史病例數與死亡數(參考圖 1)。除各國 CDC 外，亦有許多公衛組織亦會提供 COVID-19

[1] https://pride.stpi.narl.org.tw/index/graph-world/detail/4b1141ad70bfda5f0170e64424db3fa3

[2] 資料來源: https://developer.twitter.com/en/docs/apps/overview

病例報告。由於 COVID-19 症狀可能不會立即出現，報告和檢測也會延遲，也並非每個感染者都會接受檢測，所以實際病例完整度也會有差異。有鑑於此，本研究將直接採用美國 CDC 公布之數據最為實驗之用。

二、以 Twitter 發文建立 COVID-19 預測模型

為建立以社群網路發文為基礎的預測模型，本研究並非直接將發文斷詞，以一般文章分析的方式來進行，而是彙整與防疫相關之關鍵詞，再將文章轉換為與防疫關鍵詞相關之頻率向量。為此，本研究考量以下四種不同來源，統整 COVID-19 相關之關鍵詞(keywords)，做為後續 Twitter 發文分析之依據:

(a) 美國 CDC 公佈之 COVID-19 症狀：發燒發冷、咳嗽、頭痛、疲勞、噁心嘔吐、肌肉酸痛、流鼻涕、胸悶、腹瀉、失去嗅覺味覺、喉嚨痛等。
(b) Google 2020 年度熱門搜尋：新冠病毒。
(c) 8 月到 12 月 Twitter 文章中出現頻繁之關鍵字：戴口罩、COVID19、疫苗、測試呈陽性、待在家、體溫、沒有症狀、肺炎等。
(d) 英文中跟疫情有關的口語單字：健康篩檢、自我監控、隔離、接觸者追蹤、強制休假等[3]。

根據上述所產生之關鍵詞集(共 25 個)，再將所蒐集之 Twitter 發文進行以下處理，供後續塑模與預測之用：

(一) 資料集的產生

本研究運用 Twitter API 蒐集美國地區的每日公開發文，每日分別約有七千多筆。根據所蒐集的資料，首先找出含有關鍵字的發文，再以人工方式審視出現的關鍵字是否符合所需(確實與 COVID-19 相關)。以單字 fever 為例，本研究所需的意涵需為「發燒」，而有些發文雖包含 fever，但其語意卻是「狂熱」(如 playoff fever)。需注意的是，有些人被感染但沒有出現任何症狀或感到不適，導致當日的 Twitter 發文內容未必能反應至當日 COVID-19 新增病例數。根據上述做法，本研究統計出資料蒐集期間每日關鍵字出現的次數(以美國為例，如表 1 所示)。

由於各關鍵詞出現頻率可能差距很大，因此將其進行正規化，參考以下公式:

$$\mathrm{s}(k_i) = \frac{t(k_i)}{m \times n} \times \mathrm{scale} \quad (2)$$

其中，$t(k_i)$為關鍵字k_i出現在當日所有發文中之頻率，m 為關鍵字總數，n 為當日所有發文總數；scale 則為一整數，其用途為避免前項$\frac{t(k_i)}{m \times n}$數值過小。例如，關鍵字 'cough' 在 2020/8/6 所有發文中共出現了 2 次，則 t('cough')=2。假設關鍵字總數 m 為 25 個，當日發文總數 n 為 7270 篇，且 scale=10^5，則'cough'在當日之正規化頻率為s('cough') = $\frac{2 \times 100000}{25 \times 7270}$。透過上述轉換，再令K = $\{k_1, k_1, k_2 \cdots k_n\}$表示所用之 COVID-19 關鍵字集，則每日資料記錄便可用如下向量來表示：

[3] 來源: 蒐集自天下雜誌之防疫英文與 BBC 防疫相關英文用語

$$Rec_j(K) = \langle s(k_1), s(k_2) \cdots s(k_n), Rate_j \rangle \quad (3)$$

其中，Rec_j代表資料集中第 j 天的記錄，$Rate_j$為 CDC 所提供之當日新增病例數(或死亡數，可依照預測目標調整)。

表 1: 2020 年美國地區關鍵字統計表

2020年8/3~12/31 美國 28個關鍵字	1 發燒發冷 fever feverish chills	2 咳嗽 cough coughed coughing	3 頭痛 headache migraines	4 疲勞 fatigue	5 噁心嘔吐 nausea vomit vomiting	6 肌肉酸痛 muscle aches body aches	7 流鼻涕 runny nose	8 胸悶 chest pressure	9 腹瀉 diarrhea	10 失去嗅覺味覺 loss of smell loss of taste	11 喉嚨痛 sore throat
關鍵字來源	CDC	CDC	CDC	CDC	CDC	CDC	CDC	CDC	CDC	CDC	CDC
總共151天 有幾天是0	87	56	36	110	96	138	139	150	132	112	137
2020/8/3	0	0	0	0	0	0	0	0	0	0	0
2020/8/4	0	1	0	0	0	0	0	0	0	0	1
2020/8/5	2	0	1	1	0	0	0	0	0	0	0
2020/8/6	1	2	3	0	0	0	0	0	0	0	0
2020/8/7	0	1	2	0	0	0	1	0	0	0	0
2020/8/8	1	1	1	1	2	0	0	0	0	0	1
2020/8/9	0	0	2	1	0	0	0	0	1	0	0
2020/8/10	0	1	0	0	0	0	0	0	0	1	0
2020/8/11	0	1	2	0	1	0	0	0	0	1	0
2020/8/12	0	0	2	0	0	0	0	0	0	0	1
2020/8/13	0	0	0	0	0	0	0	0	0	0	0
2020/8/14	0	2	2	0	2	0	0	0	0	1	0

表 2：預測 2020/8/20 美國新增病例(b=5)所需之記錄筆數(黃色區域)

美國	發燒發冷	咳嗽	頭痛	疲勞	噁心嘔吐	肌肉痠痛	流鼻涕	胸悶	腹瀉	確診人數
2020/8/14	0	10.08572869	10.08572869	0	10.08572869	0	0	0	0	53658
2020/8/15	5.317027781	26.58513891	15.95108334	0	5.317027781	0	0	0	0	45028
2020/8/16	5.483207676	0	0	0	0	0	0	0	0	38348
2020/8/17	0	5.628253834	22.51301534	0	5.628253834	0	0	0	0	39389
2020/8/18	0	5.051142821	5.051142821	0	0	0	0	0	0	40107
2020/8/19	0	36.5630713	0	0	0	0	0	0	0	46858
2020/8/20	0	9.950248756	9.950248756	0	9.950248756	0	4.975124378	0	0	47184
2020/8/21	0	9.62000962	9.62000962	0	4.81000481	0	0	0	0	46209

(二) 依照回溯天數建立預測模型

若預測時間點為第 t 天之 COVID-19 相關數值時(如新增病例數)，若使用完整資料集來建立模型，可能導致預測失準。事實上，病毒當前的發展態勢可能與近期的資料較為相關，若將許多早期資料納入訓練，便會影響預測準確率。有鑑於此，本研究將只取用 t 之前 b 天之資料集做為塑模之用，並將 b 稱為回溯天數。因此，若回溯天數為 b 天，則可由資料集中取出$Rec_{t-1}, Rec_{t-2}, \cdots, Rec_{t-b}$，以便建立如下預測模型:

$$M(\mathrm{K}, \mathrm{t}, \mathrm{b}) = L(Rec_{t-1}, Rec_{t-2}, \cdots, Rec_{t-b}) \quad (4)$$

其中，L 為使用之塑模學習方法，M(K,t,b)為則是以此 K, t, b 等參數建立之預測模型。例如，如需預測 2020/8/20 的美國新增病例數，且回溯天數 b=5 天，則由資料集中取出 $Rec_{2020/08/20-1}, Rec_{2020/08/20-2}, \cdots, Rec_{2020/08/20-5}$，也就是 2020/8/15 到 2020/8/19 天的資料向量，而每一天資料向量為$Rec_j(K) = \langle s(k_1), s(k_2) \cdots s(k_n), Rate_j \rangle$。參考表 2，其中所示為 2020/8/14~2020/8/21 由美國 Twitter 所蒐集之關鍵詞之正規化後之出現頻率值。若要預測 2020/8/20 的美國新增病例，且回溯天數 b=5，由資料集中取出 2020/8/15 到 2020/8/19 的資料向量(黃色部分)。

三、建構考量潛伏期之預測模型

根據美國 CDC 官方說明，COVID-19 各類型變種病毒之潛伏期為 1 到 14 天。因此，若預測模型不考量病毒潛伏期，可能影響預測準確率。舉例而言，在社群發文表示自己有輕微喉嚨痛、頭痛的患者，可能在 2 天後才確診。若以他當天發文來判別，則可能誤將其歸類於當日之確診人數中，導致訓練模型之預測失準。因此，除考量回溯天數外，本研究進一步將潛伏期因素納入(稱之延遲因子 d，單位:天)，以如下方式建立資料集：

(1) 將每日之資料向量修改為如下形式:

$$Rec_j(K) = \langle s(k_1), s(k_2) \cdots s(k_n), Rate_{j+d} \rangle \quad (5)$$

其中，$Rate_{j+d}$ 表示第 $j+d$ 天的確診人數。上述公式的改變(與(2)相較)，導因於疾病的潛伏期。換言之，當日的 Twitter 發文(頻率)可能對應至 d 天之後的確診病例，而非當日新增病例。

(2) 依照回溯天數建立延遲預測模型：

根據上述修改，便可建立考量潛伏期之延遲預測模型。若預測時間點為第 t 天之新增病例數，回溯天數為 b 天，延遲天數為 d 天，則可由數據集取出 $Rec_{t-1-d}, Rec_{t-2-d}, \cdots Rec_{t-b-d}$，再帶入公式(8)，以獲得所需之預測模型。

$$M(\mathrm{K}, \mathrm{t}, \mathrm{b}) = \mathrm{L}(Rec_{t-1-d}, Rec_{t-2-d}, \cdots Rec_{t-b-d}) \quad (6)$$

以下舉例說明延遲模型的資料集建立方式：若預測時間點 t=2022/08/12，回溯天數 b=7，延遲天數 d=2，則將採用 2022/08/03 到 2022/08/09 的資料進行訓練。此外，由於延遲因素，第 j 天的資料向量會採用第 j+2 天的新增病例做為最後一個元素，也就是 Rec_{j+2}。

上述資料集建立方式乃以 Twitter 發文為基礎，若僅使用 CDC 公布之歷史數據做為預測之訓練資料，則資料集產生的方式如下。令 H={$h_1, h_2, \ldots, h_m$}，其中 $h_{i,j}$ 表示 CDC 在第 j 天所公布之第 i 項 COVID-19 數據(如新增確診人數、新增死亡人數、累積確診人數、累積死亡人數等)，若延遲天數為 b，則第 j 天之記錄可表示為:

$$Rec_j(H_{j-b}) = \langle h_{1,j-b}, h_{2,j-b}, \ldots, h_{m,j-b}, Rate_j \rangle \quad (7)$$

據此，若回溯區間為 d 天(資料取樣區間)，所獲得之資料集便為{Rec_{j-1}, Rec_{j-2}, …, Rec_{j-d}}，再代入(6)便可獲得所需之預測模型。

四、運用於預測模型之學習方法

為建立有效的預測模型，本研究使用深度學習方法來進行 COVID-19 流行趨勢預測，並使用多層感知機(Multilayer Perceptron，MLP)與 LSTM(Long-Short Term Memory)二種方法來塑模。此外，由於前人研究經常也使用線性迴歸(Linear Regression)來預測流行病趨勢，因此我們也將用其來塑模，並與 MLP 與 LSTM 進行比較。以下介紹這些方法的特性:

(1) 多層感知機(Multilayer Perceptron，MLP):

多層感知機是類神經網路的一種，由一個或多個輸入節點(Input Nodes)，一層或多層隱藏節點(Hidden Layers)，以及輸出節點(Output Nodes)所組成。輸入節點不進行運算，單純做為神經網路與外部資訊連接的管道。隱藏層中的節點則是以輸入節點傳遞的訊息為基礎，依照權重值進行運算，再將其傳遞給後續的節點。最後，輸出節點彙整所有與其連接之隱藏節點輸出資訊，將結果輸出至外部環境。面對多維非線性預測問題，需要使用多層神經網路架構來達成。依照定義，輸入層為所有輸入節點所成之集合，隱藏層則包含多個隱藏節點，而所有的輸出節點則組成輸出層。

(2) 長短期記憶(Long Short-Term Memory，LSTM)

多層感知機並無記憶效應，更無法分辨輸入是否為長期或短期記憶，如此可能降低總體預測能力。考量保存時間序列資料中的長期記憶，因而發展出 LSTM 神經網路架構。參考圖 2，LSTM 之神經元較一般多層感知機複雜許多，且其迴路使其運作具有記憶效應。資訊在神經元中流動時，不同類型的 gate 可決定是否是要將其記憶或遺忘。LSTM 對於與時序相關的複雜系統具有良好的模擬能力，可建立具非線性反饋的預測模型。

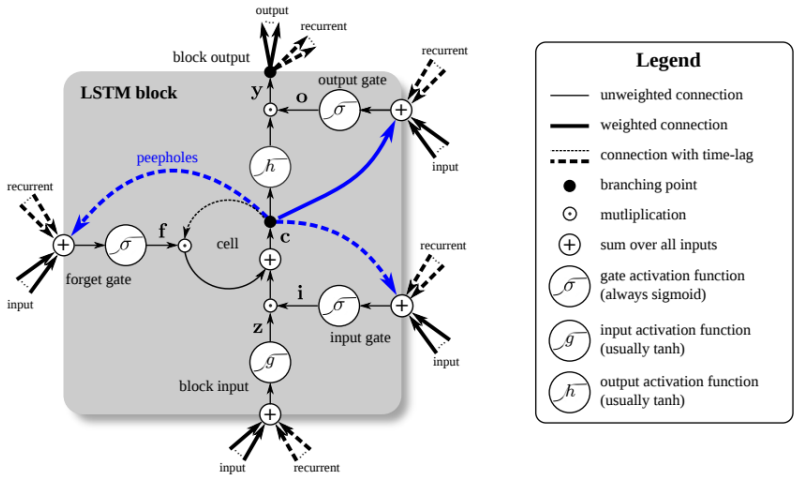


圖 2: LSTM 在隱藏層中的單元(圖片來源: Greff et al., 2017)

(3) 線性迴歸(Liner Regression)

在傳統的趨勢預測研究中，線性迴歸為常用的方法之一 (如 Lampos et al., 2010; 2012; 2013)。因此，本研究也將嘗試以迴歸模型為預測 COVID-19 未來確診、死亡等相關，並與深度學習方法進行比較。建立線性迴歸公式時，需產生各個自變數之權重值與總體的誤差項。參考以下公式，假設因變數 y 與自變數 $x_1, x_2, \ldots, x_n$ 相關，其迴歸公式可表示為:

$$y = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + \varepsilon \quad (8)$$

其中， w_i為各自變數 x_i 之權重，ε則為誤差項。為取得合適w_i與ε，則需以足夠的資料集來進行訓練。當然，由於其線性特性，過多特質迥異資料可能會產生過於「妥協」的迴歸公式，導致預測結果受到影響。

五、預測模型之產生方式

根據上述資料集產生、轉換與塑模考量之介紹，以下將討論如何據以建立深度學習預測模型，以預測 COVID-19 之流行趨勢。此外，本研究也將以常用之分段線性迴歸(piecewise linear regression)塑模，以便與深度學習模型進行比較。

(1) 以 LSTM 塑模

LSTM 是 RNN(Recurrent Neural Network)的一種，除了可考慮與時序相關的資料外(如股價預測、氣溫預測等)，亦可避免因時間間隔太大產生資訊漏失的狀況(參考 2.3 節的介紹)。因此，塑模時通常需產生與自變數、時間相關之二維資料集。據此，假設整理之資料集如表 3 所示，則可將資料集表示為 $D=[r_1, r_2, \ldots, r_m]$，其中 $r_i=[x_{i,1}, x_{i,2}, \ldots, x_{i,n}]$。據此，當考量回溯天數 b 與延遲天數 d 時，所建立之資料集

$$D_{b,d}=[\, q_{b+d+1}, q_{b+d+2}, \ldots, q_m], \text{ where } q_k=[r_{k-d-b}, r_{k-d-b+1}, \ldots, r_{k-d}, y_k] \qquad (9)$$

當然，q_k 亦可用如下更詳細方式表示:

$$\begin{aligned} q_k = [\, & [x_{k-d-b,1}, x_{k-d-b,2}, \ldots, x_{k-d-b,n}], \\ & [x_{k-d-b+1,1}, x_{k-d-b+1,2}, \ldots, x_{k-d-b+1,n}], \\ & \ldots, \\ & [x_{k-d,1}, x_{k-d,2}, \ldots, x_{k-d,n}], y_k] \end{aligned} \qquad (10)$$

根據公式(11)，當以時序相關學習方法塑模時(如 LSTM)，便可將 $D_{b,d}$ 直接做為訓練或測試之用。

表 3: 範例資料集，共有 m 筆記錄，n 個自變數，及 1 個因變數。

Records/Vars	x_1	x_2	…	x_n	y
d_1	x_{11}	x_{12}		x_{1n}	y_1
d_2	x_{21}	x_{22}		x_{2n}	y_2
….					
d_m	x_{m1}	x_{m2}	…	x_{mn}	y_m

(2) 以 MLP 塑模

多層感知機(MLP)為深度學習的基本型式，可透過多層神經元來產生深度學習效果，以模擬高維度非線性函數。與 LSTM 不同， MLP 並無特別考量資料間的時序關係。本研究以回溯天數為基礎之設計，非常適合 LSTM 等時序相關模型，但當使用如 MLP 之學習方法塑模時，則需將資料集進行以下轉換: 首先，定義函數 $C(r_i,r_j)=C([x_{i,1}, x_{i,2}, \ldots, x_{i,n}], [x_{j,1}, x_{j,2}, \ldots, x_{j,n}])=[x_{i,1}, x_{i,2}, \ldots, x_{i,n}, x_{j,1}, x_{j,2}, \ldots, x_{j,n}]$，且 $C(r_1,r_2,\ldots,r_n)=C(C(r_1,r_2,\ldots,r_{n-1}),r_n)$，之後建立 MLP 所需平坦(flattened)資料集

$$DF_{b,d}=[\, qf_{b+d+1}, qf_{b+d+1}, \ldots, qf_m], \text{ 其中 } qf_{b+d+1}=[C(r_{k-d-b}, r_{k-d-b+1}, \ldots, r_{k-d}), y_k] \qquad (11)$$

參考圖 3 之範例，假設 n=3, b=3, d=1，則平坦化後之資料集便如圖中左方序列所示，下方的 y_1, y_2,…則為各筆記錄因變數的實際值。

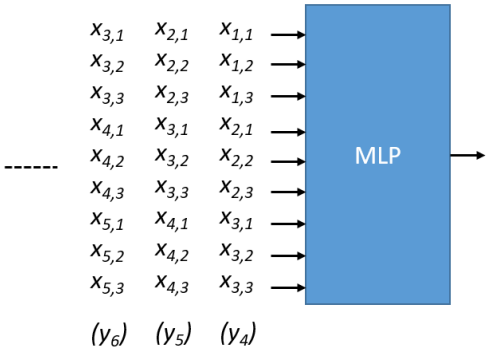
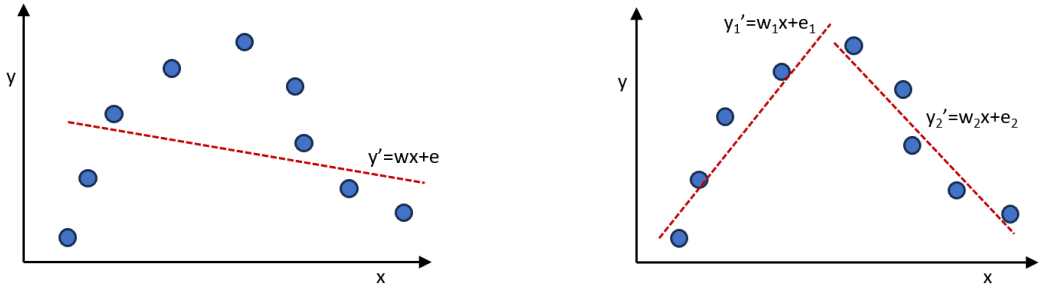


圖 3：將資料集平坦化後，做為 MLP 塑模/測試所需之資料集(b=3, d=1, n=3)

(3) 以分段線性迴歸塑模

線性迴歸為傳統趨勢預測常用方法，但由於實際趨勢變化通常為非線性，經常導致預測結果大幅失準。參考圖 4(a)中的單變數迴歸結果(參考紅線)，為了顧及所有資料點，只好以折衷方式來建立迴歸公式，導致與實際線型差異很大。但當使用分段線性迴歸時(參考圖 4(b))，則可透過建立多個迴歸公式(參考圖中的 y1'與 y2'線)，以模擬實際的資料特性。由圖中可看出，分段迴歸確實可有助於解決非線性資料集的預測。因此，本研究也將採用分段迴歸來建立預測模型，並與深度學習模型進行比較。

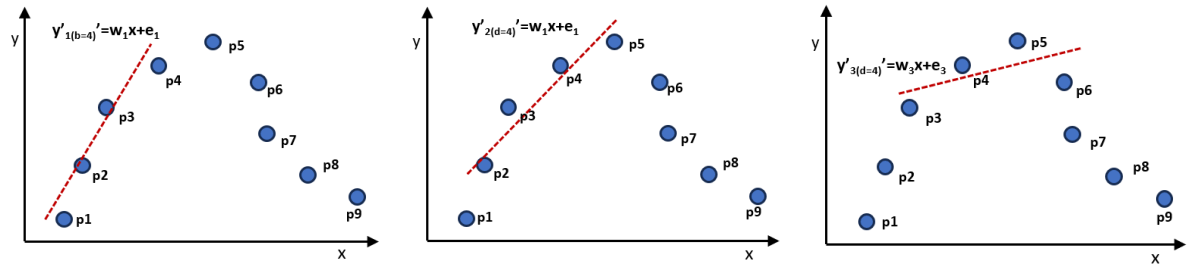


(a)線性迴歸(容易失準)　　(b) 分段線性迴歸

圖 4: 線性迴歸用於趨勢預測

分段線性迴歸在本研究中配合滑動式窗概念(sliding windows)，以回溯天數為視窗大小，並考量延遲天數，動態建立迴歸公式來進行預測。參考圖 5，假設回溯天數為 4，表示將固定以 4 個資料點來建立迴歸公式，而資料點的編號則與資料蒐集的時間相關。因此，一開始將以(p1,p2,p3,p4)來建立 y_1'(b=4)公式(參考圖 5(a))。注意，此處為了說明方便，將延遲天數 d 設定為 0。接下來，依序考量(p2,p3,p4,p5), (p3,p4,p5,p6)，以相同方式產生 y_2'與 y_3'公式(參考圖 5(b), 5(c))。由圖中可發現，這是一種以近期變化為依歸的時間相依溯模方式。其優點為可追蹤近期的變化趨勢，缺點則為仍可能受資料點變化劇烈的影響。例如，圖 5(c)中的 p6 便與預測線有較大的距離。此外，其效能顯然與回溯天數 b 的大小有關。有鑑於此，下一節將探討分段線性迴歸在不同 b 值狀況下的表現，並與深度模型比較。有關上述流程之細節，請參考圖 6 之虛擬碼。參考第 5 行，過程中將建

立 T-(b+d)個迴歸公式，b, d 分別對應至之前所提之回溯天數與延遲天數。而歷次所使用之資料集則以 piecewise()來達成，操作過程則如公式(11)所規範。當執行完畢後，便可將 rlt_list 傳回，以便與實際值進行關聯度評估等運算。



(a) 以(p1,p2,p3,p4)建立 y1'公式　(b) 以(p2,p3,p4,p5)建立 y2'公式　(c) 以(p3,p4,p5,p6)建立 y3'公式

圖 5: 以滑動式窗概念動態產生線性迴歸公式，以預測 COVID-19 相關數據。

```
1   def piecewise_linear_regression(D, b, d):
2       # D: 資料集, b: 回溯天數, d: 延遲天數
3       T = len(D) # D 之記錄筆數，對應至資料蒐集之總天數
4       rlt_list =[] # 記錄歷次預測結果
5       for skip in range(T-(b+d)): # skip 變數範圍: 0~T-(b+d-1)
6           Db_d=peicewise(D, skip, b, d) # 參考公式(11)
7           model = Linear_Regression(Db_d)
8           rlt= model.predict(D[skip+1][:-2], D[skip+d+1][-1])
9           rlt_list.add(rlt)
10      return rlt_list
```

圖 6: 分段線性迴歸塑模與評估之虛擬碼(使用 Python)

肆、 實驗結果

為驗證提出方法之有效性，本節使用實際下載的資料進行測試，實驗將分別針對不同資料集，以分段線性迴歸與深度學習之 MLP、LSTM 等不同方法建立預測模型，再進行效能評估。

一、以 Twitter 資料集建立分段線性迴歸模型之效能評估

本節先以傳統分段迴歸模型配合 Twitter 發文資料集，分析回溯天數與延遲天數對於預測關聯度(correlation)之影響，並討論第三節所提出之各種資料集建立方式之效能差異。在本實驗中，使用之關鍵詞共有 25 個，其中包含: (1)美國 CDC 所提供的 11 個形容 COVID-19 的症狀，(2) 從文章中發現常出現的 8 個單詞，(3)針對 COVID-19 的 5 個常見口語英文單字，(4) 1 個熱門 google 關鍵字。

(1) 回溯天數之影響

首先，我們將驗證回溯天數(b)對於模型效能之影響，並以相關係數做為比較之基礎，資料蒐集期間為 2020/8/3~2021/6/18。以 b=5, 7, 14, 21, 28, 35(單位:天)，對美國 COVID-19'新增病例數'與'新增死亡人數'進行預測。由表 4 的結果可知，當預測'新增病例人數'時，回溯天數 b=7 建立之預測模型可獲得最高的相關係數(0.9414)，顯示預測結果與實

際值高度相關。此外，回溯天數 b=5 則可獲得次高之相關係數(0.9285)。當回溯天數增加時(如 b>=14)，預測效能變明顯下降。與之後雖有回升，但效能仍不如 b=5, b=7 之結果。各種回溯天數對於預測結果之影響趨勢，請參考圖 7。在'新增死亡'人數的預測上，可更清楚看出回溯天數與關連性之變化趨勢。由上述結果可知:

(a) 考量 COVID-19 的流行趨勢變化快速，線性迴歸之預測模型應以近期蒐集的資料為主 (在此為與 COVID-19 相關的近期發文)，方可獲得較佳的預測效能。而天數而論，則在 7 天(一週)左右最為合適。

(b) 實驗結果顯示'新增病例人數'之預測結果，明顯優於'新增死亡人數'。以 b=7 天為例，二者之關聯度分別為 0.9414 與 0.8991，差異相當明顯。此差異應來自 COVID-19 的潛伏期天數較可掌控，但發病後死亡的天數則長短不一，因此較難預料，導致後者的預測結果較差。

表 4：以線性迴歸模型配合各種回溯天數之預測結果(美國)

'新增病例數'預測						
回溯天數	b=5	b=7	b=14	b=21	b=28	b=35
相關係數	0.9285	0.9414	0.8436	0.2116	0.5909	0.7117
'新增死亡人數'預測						
回溯天數	b=5	b=7	b=14	b=21	b=28	b=35
相關係數	0.8746	0.8991	0.782	0.2229	0.6105	0.7266

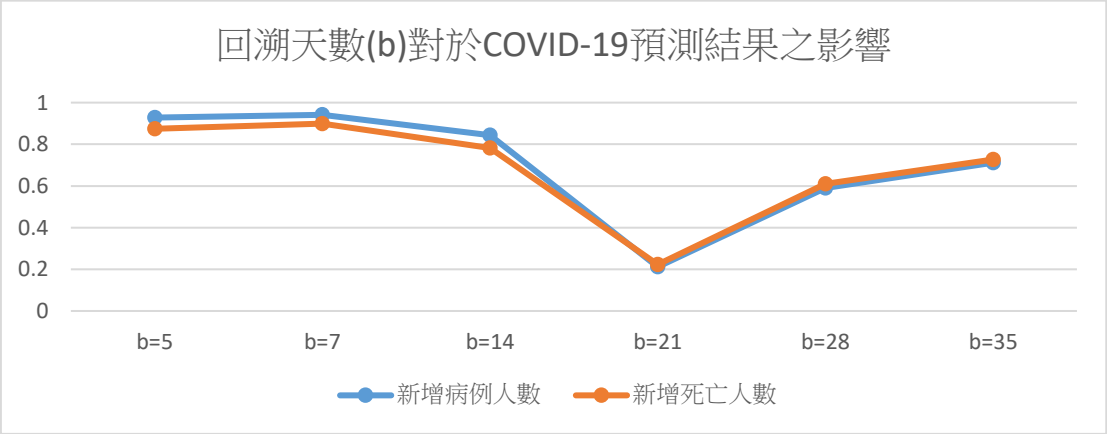


圖 7: 不同回溯天數對於預測結果關聯度之影響(使用線性迴歸塑模)

(2) 考量潛伏期之延遲模型預測效能評估

本節將探討將潛伏期引發之延遲因素加入 Twitter 模型中所產生之影響。實驗時，延遲天數(d)範圍由 1 到 14 天，塑模方法仍採用線性迴歸。參考 5，表中數據為使用美國 twitter 資料集建立延遲模型之預測值真實數據之相關係數。根據上一節的回溯天數實驗，美國資料集以回溯天數(b)以 b=5,7 為最佳，因此表 5 僅呈現這二種回溯天數之實驗結果。觀察表 6 的結果發現，在'新增病例數'之預測上，回溯天數 b=5 延遲天數 d=3 之設定可取得最高之相關係數(0.9127)，而 b=5, d=2 則可獲得次佳的結果。而在'新增死亡

人數’之預測中，則以 b=5, d=5 與 b=5, d=2 之結果較佳。上述結果除說明延遲模型確實能反應潛伏期造成的影響，取得較佳之結果外，也顯示選擇適當延遲天數的重要性。圖 8 為根據表 5 繪製之變化趨勢圖，由圖中可看出，’新增病例數’與’新增死亡人數’的預測關聯度分別在 d=2 與 d=5 之後開始下降，且二者同時在 d=2 時獲得次佳的預測關聯度。此結果顯示大多數人的潛伏期可能為 2~5 天，而 2 天應是合理的基本設定。

綜合上述結果可知，以社群發文來預測新冠肺炎之流行趨勢確實可行。在配合傳統線性迴歸模型時，若能適當設定回溯與延遲天數，可獲得超過 90%之高關聯度。此外，本研究所提及之回溯天數與延遲天數，確實會影響模型的預測準確性。若能仔細分析，更深入了解 COVID-19 之潛伏期與流行趨勢。

表 5：使用 Twitter 發文資料建立線性迴歸延遲預測模型之結果(美國)

新增病例數	b=5	b=7	新增死亡人數	b=5	b=7
d=1	0.8886	0.8867	d=1	0.8435	0.8406
d=2	0.9042	0.8671	d=2	0.8734	0.8188
d=3	0.9127	0.8872	d=3	0.8616	0.8264
d=4	0.8891	0.8501	d=4	0.8612	0.8131
d=5	0.8875	0.8730	d=5	0.8809	0.8360
d=6	0.8533	0.8162	d=6	0.8458	0.8175
d=7	0.7851	0.7732	d=7	0.8084	0.8153
d=8	0.8218	0.8053	d=8	0.7542	0.7383
d=9	0.7891	0.8198	d=9	0.7944	0.7987
d=10	0.8554	0.8239	d=10	0.7958	0.7573
d=11	0.7692	0.7512	d=11	0.7500	0.7295
d=12	0.7836	0.7603	d=12	0.7626	0.6949
d=13	0.7792	0.7337	d=13	0.7679	0.7391
d=14	0.7215	0.6546	d=14	0.6494	0.6163

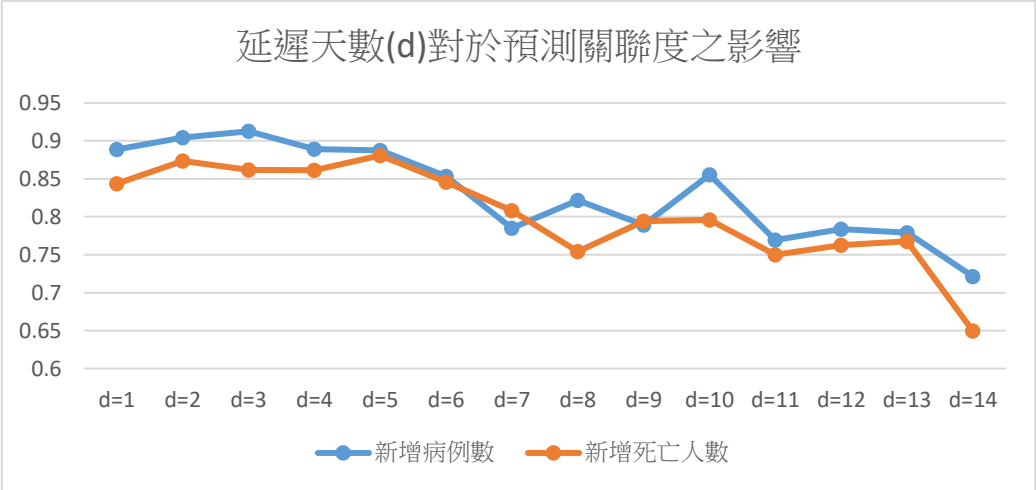


圖 8: 在 b=5(回溯天數)狀況下，延遲天數(d)對預測關聯度之影響(使用分段線性迴歸)

二、以 Twitter 資料集建立深度學習預測模型之效能評估

本節將改用深度學習方法 MLP 與 LSTM 來建立預測模型，並仍使用前述的 Twitter 發文做為資料集，訓練集與測試集之比例為 7:3。MLP 模型共有三個隱藏層，神經元個

數分別為 4096, 64, 32；LSTM 則有二個隱藏層，第一層含有與自變數個數相同之 LSTM 單元(此處對應至 25 關鍵詞)，第二層則為平坦層，含有 1024 個神經元。

首先，我們介紹 MLP 塑模之實驗結果。參考表 6(a)，當預測'新增確診人數'時，在 b=5,7,14,21 與 d=1~5 之 20 種設定組合下，其預測關聯度大約介於 80%~87%之間。雖不如分段線性迴歸之最佳結果(0.9127)，但與實際值仍具有高關聯度。當以'新增死亡人數'為預測標的時(表 6(b))，與之前結果類似，獲得 57%~72%左右之較低關聯度(因感染後死亡時間較難預料)。有關 MLP 預測模型與分段線性迴歸之比較如下:

(a) 傳統的線性迴歸雖然單純，但在短期趨勢預測上，仍可獲得相當不錯的關聯度。而 MLP 因使用多層架構，為使模型中的神經元能有適當參數，通常需有足夠大量的資料來進行訓練。

(b) 回溯天數(*b*)對於線性迴歸模型的表現有明顯影響，其在 *b*=14 時便開始明顯下降。然而，對 MLP 而言，卻因 *b* 的增加而展現更高的預測關聯度。此結果顯示深度學習模型對於如何擬合更多資料，具有更好的學習能力。而當採用單純線性迴歸時，卻可能因同間考慮太長時間區間中的資料，導致其描述能力下降。

(c) 不像線性迴歸，MLP 對於延遲天數 *d* 並不敏感，在 *d*=1~5 的設定中，大多呈現幅度很小的變化，原因應同如(b)所述。

值得注意的是， 以上推論與所使用資料集相關(此處為 Twitter 發文)，不同類型資料集會使學習方法展現不同特性。

表 6: 使用 MLP 配合 Twitter 資料集(美國)塑模之預測關聯度

(a) 預測'新增確診數'

相關係數	預測'新增病例數'				
回溯天數(b)/延遲天數(d)	d=1	d=2	d=3	d=4	d=5
b=5	0.8102	0.8098	0.8082	0.8096	0.8083
b=7	0.8565	0.8620	0.8580	0.8576	0.8588
b=14	0.8691	0.8700	0.8704	0.8707	0.8670
b=21	0.8642	0.8662	0.8641	0.8640	0.8645

(b) 預測'新增死亡人數'

相關係數	預測'新增死亡人數'				
回溯天數(b)/延遲天數(d)	d=1	d=2	d=3	d=4	d=5
b=5	0.5701	0.5717	0.571	0.5695	0.5683
b=7	0.6657	0.6663	0.6662	0.6721	0.6675
b=14	0.7058	0.7076	0.7047	0.7049	0.6697
b=21	0.7218	0.7301	0.7294	0.7346	0.7246

除 MLP 之外，我們也使用LSTM 配合 Twitter 資料集產生預測模型，實驗結果如表 7 所示。由於上述的 MLP 實驗顯示深度學習方法對於延遲天數 *d* 並不敏感，因此只在回溯天數變動，延遲天數則根據 4.1 節之實驗觀察將其固定為 2。由表中可看出，在'新增確診數'的預測上，關聯度約介於 0.57~0.76 之間，而'新增死亡人數'則介於 0.51~0.66 之間，關聯度明顯低於 MLP。其可能原因如下: LSTM 的架構與時序有關，能考量資料之時序對於預測結果之影響。然而，由於感染者之潛伏期與發病後死亡天數均不固定，導致相關發文與預測標的時序關係並不明顯。此外，LSTM 架構較 MLP 複雜，有更多參數需要調校，也是導致在有限的訓練資料狀況下，LSTM 表現不如 MLP 之可能原因。

表 7: 使用 LSTM 配合 Twitter 資料集(美國)塑模之預測關聯度

回溯天數(b)	延遲天數(d)	相關係數	
		預測'新增確診數'	預測'新增死亡人數'
5	2	0.5702	0.5164
7	2	0.7283	0.5444
14	2	0.7618	0.6227
21	2	0.7212	0.6617

表 8：使用三種塑模方法配合美國 CDC 公布 COVID-19 數據之預測關聯度

預測標的*	回溯天數(*b*)	MLP	LSTM	線性迴歸
新增病例數	5	0.9630	0.9472	0.8679
	7	0.9665	0.9477	0.9046
	14	0.9686[b]	0.9458	0.9245
	21	0.9730[a]	0.9461	0.9282
新增死亡人數	5	0.9658	0.7361	0.7991
	7	0.9630	0.8295	0.8810
	14	0.9712[a]	0.8509	0.8979
	21	0.9660[b]	0.8856	0.8985

*延遲天數 *d* 均設定為 2, [a]:最佳結果, [b]:次佳結果

三、以 CDC 歷史數據建立預測模型之效能評估

本節將改以美國 CDC 公布之 COVID-19 相關歷史數據做為資料集，以深度學習方法 MLP 與 LSTM 建立預測模型，並與分段線性迴歸進行比較。資料蒐集期間為 2020/8/3~2021/6/18，資料欄位包含: '病例總數', '新增病例數', '確認病例總數', '可能病例總數', '新的可能病例', '死亡總數', '新增死亡人數', '確認死亡總數', '可能死亡總數, '新的可能死亡人數'等 10 項。實驗時，訓練集與測試集之比例為 7:3，並以考量延遲天數之'新增病例數', '新增死亡人數'做為預測標的。

表 8 為配合美國 CDC 公布 COVID-19 歷史數據，使用 MLP、 LSTM 與分段線性迴歸塑模之結果比較。其中，MLP, LSTM 模型使用之架構與 4.2 節相同，延遲天數(*d*)則根據仍 4.1 節之實驗觀察將其固定為 2。由實驗結果可知，當預測'新增病例數'時，以 MLP 的表現為最佳，關聯度可高達 0.9730(*b*=21)，次高仍為 MLP 在 *b*=14 設定下之 0.9686。為更清楚了解三種模型之效能，請參考圖 9 (a)之比較圖:

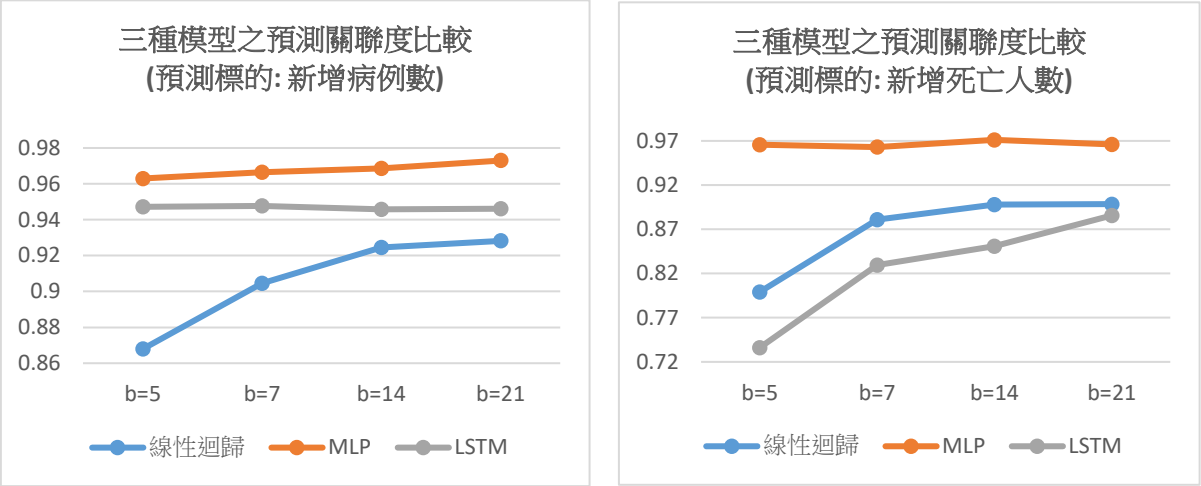
(a) 由圖中可看出，MLP 在各種回溯天數(*b*)的設定下，均可獲得超過 96%之最佳關聯度，LSTM 居次，其預測關聯度也均得超過 94%。線性迴歸所獲之關聯度為其中三者最低，範圍大致介於 86%~92%。由此可知，使用深度學習方法確實可獲更佳、更穩定的預測結果，有助於提升預測結果之參考價值。

(b) 其次，三種模型之結果大致都可獲得超過 90%之關聯度(除線性迴歸在 b=5 設定下)，這也說明使用機器學習方法預測疾病趨勢具有高度的可行性。

當預測'新增死亡人數'時，仍以 MLP 表現最佳，可獲得高達 0.9667 關聯度(*b*=14)，線性迴歸與 LSTM 則分居二、三名。參考圖 9(b)之比較圖，可得知:

(a) MLP 仍在不同 *b* 值設定下獲得一致性的最佳結果，而 LSTM 與線性迴歸之關聯度則明顯下降(與'新增病例數'之預測相較)，與上二節之實驗結果類似。

(b) 但當回溯天數(b)增加時均有改善，再次說明感染後死亡的時間不定(與感染之潛伏期不同)，更長的回溯時間較有機會涵蓋更多可能案例。



(a) 預測標的為'新增病例數'　　(b) 預測標的為'新增死亡人數'

圖 9: 根據表 8 所繪製之三種塑模方法之預測關聯度比較，以 MLP 表現最佳

綜合以上各節之結果，可清楚驗證以機器學習方法進行 COVID-19 趨勢預測之有效性，若能進一步深入發展，應可產生更具實用價值之資訊。深度學習雖然需耗費較高的運算成本，對於不同性質之資料集亦有不同表現，但確實有助於提升 COVID-19 之預測效能。此外，傳統分段迴歸雖然簡單直覺，但在資料量有限或資料變化較不劇烈時，仍不失為一種可行的預測方法。本研究分別以 Twitter 資料集與 CDC 歷史數據建立預測模型，獲得結果也有所差異。實驗結果顯示，以 CDC 歷史數據為基礎之預測模型優於以 Twitter 社群發文為主之預測模型。可能原因為 Twitter 發文雖可反應疾病流行趨勢，但感染者不一定會上線發文，發病後也可能因身體不適，無法及時發文。然而，在疾病尚未大規模流行時，CDC 並不會對其進行各種數據統計，因此也無法獲得大量訓練資料。此時，以社群發文資料為基礎發展疾病流行預測，便顯得格外重要，能讓政府相關單位獲得寶貴的早期預警資訊。

伍、 結論

近年來，COVID-19 對人民生活與經濟發展造成重大影響，更導致無數寶貴生命因此而喪失。為防止新冠病毒的傳染，各國政府紛紛制定各種防疫措施，以保護國民的生命安全。政府制定各項政策時需有多方考量，需顧及成本效益與時效性，期能以合理的資源達成最大防疫效果。有鑒於此，研究者莫不投注大量心力，發展防疫輔助方法，做為相關單位制訂策略時之決策依據。有鑑於此，本研究結合官方 COVID-19 歷史數據與網路社群發文，配合機器學習方法預測 COVID-19 的發展趨勢。首先，我們蒐集美國 CDC 公布之 COVID-19 相關數據，並截取期間的 Twitter 發文，產生混合式的資料來源。當對社群發文進行分析時，我們採用官方公布的症狀單詞作為關鍵字。其次，本研究考量疾病的潛伏期，建立具有延遲特性的預測模型，期能提升預測準確率。最後，我們分別使

用 Linear Regression, MLP 與 LSTM 進行塑模，預測未來可能的死亡與確診人數。實驗結果顯示，本研究提出之方法確實有助於 COVID-19 之流行趨勢預測，做為相關單位在制定策略時的決策依據。

本研究雖然以 COVID-19 為研究對象，但所開發之方法亦可應用至其他傳染病之防控。以登革熱為例，其潛伏期為 3~14 天，病程約 2~7 天，論文中提出之方法架構便可用於其流行趨勢預測。當然，在不同應用中需設計不同之症狀關鍵詞，並假設感染者會在網路社群發文中陳述其症狀。又如，豬瘟、雞瘟等動物傳染病對於經濟影響巨大，亦可對於養殖戶在相關社群之發文進行分析，以早期發現流行徵兆。雖然我們已驗證本研究提出方法之有效性，但仍有以下研究限制: (1)本研究假設感染者會在其常用的網路社群上發文，並反應其身體狀況。然而，由於新型態媒體的快速發展(如 Instagram、TikTok 等)，以文字分析為基礎之預測方法將無法適用於以圖像、影音為主之網路社群。(2)本研究僅對新增病例數與新增死亡人數進行預測，對於住院天數、重症至死亡天數等涉及複雜病程發展之資訊，並無法提供相關預測。(3)本研究雖然考量疾病潛伏期與資料回溯時間，但當潛伏期與病程變化不定時，便可能導致方法效能受到影響。有關本研究所發展方法之未來發展如下: 首先，未來或許能將研究與經濟數據進行連結，相互驗證方式，預測其之間的因果關係。其次，對於 COVID-19 關鍵詞的設定，可透過學習方式來取得，而非事先預設。此外，更多的資料來源(社群發言或官方疾管局數據)對於預測結果應有提升的效果，因此未來亦可進一步檢視不同類型資料集，並以資料融合方式彙整，以產生更有效的預測模型。

參考文獻

1. 衛生福利部疾病管制署，2021，疾病介紹，https://www.cdc.gov.tw/Category/Page/vleOMKqwuEbIMgqaTeXG8A
2. 中央社，2021，立院三讀，紓困特別預算總額上限提高至 8400 億元，https://www.cna.com.tw/news/firstnews/202105315004.aspx
3. Alsabek, M., B., Shahin, I., and Hassan, A., "Studying the Similarity of COVID-19 Sounds based on Correlation Analysis of MFCC," 2020 International Conference on Communications, Computing, Cybersecurity, and Informatics.
4. Bardak, B., & Tan, M., "Prediction of influenza outbreaks by integrating Wikipedia article access logs and Google flu trend data". Bioinformatics and Bioengineering (BIBE), 2015 IEEE 15th International Conference.
5. Gupta, P., Kumar, S., Suman, R. R., and Kumar V., "Sentiment Analysis of Lockdown in India During COVID-19: A Case Study on Twitter," IEEE Transactions on Computational Social Systems, 8(4), 2021, 939-949.
6. Gupta, V. K., Gupta, A., Kumar, D., and Sardana, A., "Prediction of COVID-19 confirmed, death, and cured cases in India using random forest model," Big Data Mining and Analytics 4(2), 2021, 116 – 123.
7. Hossen, Md. S., and Karmoker, D., "Predicting the Probability of Covid-19 Recovered in South Asian Countries Based on Healthy Diet Pattern Using a Machine Learning Approach," The 2nd International Conference on Sustainable Technologies for Industry 4.0 (STI), 2020.

8. Hu, C., "The Topological Properties of COVID-19 Global Activity Time Series Forecasting," The 5th International Conference on Information Science, Computer Technology and Transportation, 2020.
9. Lampos, V., and Cristianini, N., "Tracking the flu pandemic by monitoring the social web," The 2nd International Workshop on Cognitive Information Processing, 2010, pp. 411-416.
10. Lampos, V., and Cristianini, N., "Nowcasting Event from the Social Web with Statistical Learning," ACM Trans. On Intelligent Systems and Technology, Vol. 3, No. 4, Sep. 2012, pp. 72:1~72:22.
11. Lampos, V., Preotiuc Pietro, D., and Cohn, T., "A user centric model of voting intention from social media," The 51st annual meeting of the association for computational linguistics, 2013, pp. 993-1003.
12. Latif, S., Usman, M., Manzoor, S., Iqbal, W., Qadir, J., Tyson, G., Castro, I. and Razi, A., " Leveraging Data Science to Combat COVID-19: A Comprehensive Review," IEEE Transactions on Artificial Intelligence, 1(1), 2020, 85-94.
13. Liu, X., Zheng, L., Jia, X., Qi, H., Yu, S., and Wang, X., "Public Opinion Analysis on Novel Coronavirus Pneumonia and Interaction with Event Evolution in Real World," IEEE Transactions on Computational Social Systems, 8(4), 2021, 1042 – 1051.
14. Mandayam, A., Rakshith A.C, Siddesha, S., Niranjan S. K., "Prediction of COVID-19 pandemic based on Regression," The Fifth International Conference on Research in Computational Intelligence and Communication Networks, 2020.
15. Mourad, A., Srour, A., Harmanani, H. M., Jenainatiy, C., and Arafeh, M., "Critical Impact of Social Network Infodemic on Defeating Coronavirus COVID-19 Pandemic: Twitter-based Study and Research Directions," IEEE Transactions on Network and Service Management, 17(4), 2020, 2145-2155.
16. Naseem, U., Razzak, I., Khushi, M., Eklund, P. W., and Kim, J., "COVIDSenti: A Large-Scale Benchmark Twitter Data Set for COVID-19 Sentiment Analysis," IEEE Transactions on Computational Social Systems, 8(4), 2021, 1003 - 1015.
17. NCHC, "COVID-19 全球疫情地圖", https://covid-19.nchc.org.tw/, last retrieved 2023/Aug.
18. NCHC, " USA COVID-19 每日確診統計圖 ", https://covid-19.nchc.org.tw/2023_country_confirmed.php?mycountry=USA, last retrieved 2023/Dec.
19. Nikil, S., Dalmia, H., and Kumar, P. ,"Covid-19 Outbreak Analysis," The 2020 International Conference on Smart Technologies in Computing Electrical and Electronics, 2020.
20. Prakash, B. A., "Prediction Using Propagation: From Flu Trends to Cybersecurity", IEEE Computer Society, 2016, pp. 84-88.
21. Rakhra, A., Jain, I., Gupta, R., and Bhatia, M. ,"Predicting the Prevalence Rate of COVID-19 Falsity on Temperature," The 11th International Conference on Cloud Computing, Data Science & Engineering, 2021.
22. Su, Z., Pahlavan, K., and Agu, E., "Performance Evaluation of COVID-19 Proximity Detection Using Bluetooth LE Signal," IEEE Access, Vol. 9, 2021, 38891 – 38906.
23. Teixeira, J.P. and Fernandes, P.O., "Tourism time series forecast with artificial neural networks," Tékhne, Volume 12, Issues 1–2, January–December 2014, Pages 26-36.
24. Vrindavanam, J., Srinath, R., Shankar, H., and Nagesh, G., "Machine Learning based COVID-19 Cough Classification Models – A comparative Analysis," The 5th International Conference on Computing Methodologies and Communication, 2021.
25. Wibowo, N., S., Mahardika, R., and Kusrini, K., "Twitter Data Analysis Using Machine Learning to Evaluate Community Compliance in Preventing the Spread of Covid-19," The 2nd International Conference on Cybernetics and Intelligent System, 2020.

26. Wu, N., Green, B., Ben, X. and O'Banion S., "Deep Transformer Models for Time Series Forecasting: The Influenza Prevalence Case," https://doi.org/10.48550/arXiv.2001.08317, Jan. 2020.
27. Yu, Shuo, Qing, Q., Chen Zhang, Ahsan Shehzad, Giles Oatley, and Feng Xia, "Data-Driven Decision-Making in COVID-19 Response: A Survey," IEEE Transactions on Computational Social Systems, 8(4), 2021, 989-1002.
28. Waldmann, P., "On the Use of the Pearson Correlation Coefficient for Model Evaluation in Genome-Wide Prediction," Frontiers in genetics, Sep. Vol. 10, 2019.