運用數據科學杜絕使用者被假消息誤導或遭受網路霸凌

成大統計系暨數據科學所\李政德

透過智慧型手機或上網設備進入社群媒體的世界，如Facebook、Instagram與Twitter，與遠方的親朋好友分享訊息與交流互動，這已經是當前不分國家、年齡與族群的使用者的日常，社群媒體提供便利社交互動功能，使用者僅須透過『點讚』、『分享』或『留言』，就能對訊息表達意見，而社群媒體的訊息演算法推薦機制提高訊息曝光度，因此讓社群媒體具備強大的資訊擴散(Information Diffusion)性質，易於形成意見領袖的看法、以及浮現出具爭議且高度被討論的事件。然而，資訊擴散也使得社群媒體成為兩面刃，有心人士得以利用演算法造成「同溫層效應」(Echo Chamber)來操作對特定議題的輿論，假訊息(Fake Messages)因被演算法放大其能見度並模糊其真偽性，使得一昧接受來自社交圈訊息的使用者暴露在被假訊息誤導的風險中，就如2020年各廠牌COVID-19疫苗的副作用在疫苗公布初期被未經查證的言論給誤導；在此同時有心人士也能透過虛假帳號或自動帳號，以仇恨言論和暴力文字的形式，集體對於目標人士進行所謂的網路霸凌(Cyberbullying)，國內最典型例子是藝人楊又穎輕生事件。

儘管社群媒體充斥假訊息，且不時會有網路霸凌事件發生，但凡走過必留下痕跡，數位科技的進度與網際網路的成熟，讓在社群媒體上使用者的一舉一動都留下了數位足跡，使用者的個人檔案、曾經與哪些貼文與哪些粉絲頁有過互動，甚至互動的文字與時間地點，這些紀錄無時不刻被儲存在社群媒體背後的龐大資料庫中，形成所謂的大數據(Big Data)，讓研究人員得以透過各種數據科學(Data Science)、人工智慧(Artificial Intelligence, AI)與機器學習(Machine Learning)等技術，從資料中挖掘與假訊息和網路霸凌相關的線索與知識，最終使得自動化偵測社群媒體不實訊息，並且準確偵測即將發生的網路霸凌行為成為可能。

在臺灣其實已經有基於大數據的虛假訊息示警平台，最知名的是LINE 聊天機器人「美玉姨」，它基於臺灣事實查核中心與事實查核平台的資料庫，可自動查證可疑的訊息，並且提供對應的事實佐證，其他知名事實查核平台包含「MyGoPen麥擱騙」、「Cofacts真的假的」、「趨勢科技防詐達人」、「臺灣事實查核中心」、「蘭姆酒吐司」與「LINE訊息查證」。這些平台透過群眾智慧與領域專家所共同彙整的不實訊息與謠言資料庫，以人工智慧的自然語言處理技術(Natural Language Processing)進行訊息文本內容比對與數據擴增，輔助人們進行更有效率的虛假訊息判斷。根據資策會的AI鑑識臺灣不實訊息聯防體系技術的統計，在2021年這些自動化事實查核平台已透過人工智慧技術將事實查核的速度提高了三倍，節省了80%的人工查核，並且有效杜絕了超過300種與COVID-19相關的不實訊息。

對於網路霸凌的杜絕與防範，儘管在臺灣目前則尚未有能夠自動偵測網路霸凌行為的平台或軟體，在美國已經有人實現透過AI偵測網路霸凌，並且將其開發成App與瀏覽器擴充功能，而且發明者是一名年僅17歲的少女拉奧(Gitanjali Rao)，她發明了Kindly這項智慧服務，能夠讓正在社群媒體上編打文字的青少年，自動獲得文字是否具有霸凌意圖，並獲得替代文字建議來發表貼文訊息，協助讓青少年族群的網路文字行為更加安全健康。在臺灣，教育部從全民的資訊素養著手，打造了「全民資安素養網iSafe」，讓青少年、家長與社會大眾認識正確且合法的網路行為，從根本面防範網路霸凌的發生。

研究人員持續開發先進的數據科學技術來偵測假訊息和網路霸凌行為，屬於人工智慧與資料科學領域中，自然語言處理子領域的範疇。我們成功大學數據科學所的網路人工智慧實驗室(Networked Artificial Intelligence Laboratory, NetAI Lab)很榮幸在這個課題上有相關研究成果。針對社群媒體的假訊息偵測，現有的方法雖然有一定準確性，但仍有幾個主要的限制，首先是既有技術需要仰賴對於類似議題一定數量的訓練資料，方能進行穩健的模型學習，換言之若出現新興議題或事件，在未能擁有訓練資料下，模型方法勢必難有好的效果；第二，現有演算法需要仰賴社群媒體使用者對於訊息的評論文字與轉傳分享紀錄，方能從群眾對訊息的反饋與互動中學習得知真偽性，然而我們更需要在訊息被張貼發布的當下立即判斷其是否為不實資訊，欲達到偵測即時性便難以擁有足夠使用者反饋紀錄來輔助預測真偽；第三，相較使用者對目標訊息的文字反饋，一般使用者更傾向採用點讚來表達意見，然而現有方法卻忽略這些點讚使用者的長相，譬如短時間累積大量點讚的訊息可能背後存在網軍操作。

我們開發了一基於圖神經網路(Graph Neural Network)的深度學習演算法GCAN，能有效克服上述三項挑戰，在不需要仰賴大量訓練資料、不需要仰賴使用者文字反饋、且能有效利用點讚使用者的屬性，就能做到即時假訊息偵測，我們在社群媒體Twitter上可獲得將近90%的準確性，可在不實訊息

廣為擴散前將其示警給使用者們進行閱讀上的參考。除了具有高度準確性，GCAN演算法也具備可解釋性人工智慧的特性，能夠標示出一則訊息為何會被判斷為虛假資訊，譬如訊息文字經常使用「突發事件」與「協助轉傳」等字眼，加上轉傳該則訊息的使用者多為未經帳號查驗的新創帳號，便有較高的機率是不實訊息。此一成果已發表於The 58th Annual Meeting of the Association for Computational Linguistics (ACL) 2020，該會議為人工智慧自然語言處理頂尖國際會議，截至2022年四月已獲得超過110次的引用，足見此技術在假訊息偵測課題上的國際影響力。

針對社群媒體的網路霸凌行為偵測，我們成大NetAI Lab研究團隊亦有突破性的AI技術開發成果。典型網路霸凌發生於惡意族群對於社群媒體目標使用者所發起的討論串、透過留言評論的形式蓄意進行文字言語上充滿惡意且重複性的攻擊，使目標受到傷害，因此如何有效分析留言文字是否存在惡意，以及找出哪些使用者屬於會發起霸凌，是基於網路霸凌歷史紀錄作為訓練資料、透過數據科學方法偵測網路霸凌行為的關鍵。圖神經網路是一種能夠有效學習個體在人事時地物等不同面向上相互關聯的深度學習方法，近幾年已成為社群媒體分析與文字語意理解的重要技術，因其能有效從錯綜複雜的網路關聯結構中抽絲剝繭，學習出對於預測目標最有用的潛在特徵。我們基於圖神經網路開發了偵測網路霸凌行為的AI演算法HENIN，它能自動學習當前訊息文字與歷史霸凌事件文字間的關聯性，同時學習當前留言使用者族群與歷史霸凌使用者的特徵長相是否雷同，藉此達到當前最優異的網路霸凌偵測效果，在Instagram標準數據集上可獲得將近九成的準確率，並且能夠精準標示出哪些留言最是最可疑的霸凌行為。此成果的研究論文也發表於EMNLP 2020，該會議同樣為人工智慧自然語言處理頂尖國際會議。

儘管成大團隊在社群媒體上偵測不實訊息與霸凌行為已獲得具國際競爭力的研究成果，但我們認為要透過數據科學邁向真實世界可落地被有效應用的假訊息與霸凌行為等失序資訊(Disorder Information)偵測，眼前還有幾個關卡必須解決度過。第一，惡意人士製造假訊息和霸凌行為的手法千變萬化，且所針對的議題也多變，甚至拿出歷史新聞作為佐證，使得現有基於歷史大數據來訓練偵測模型受到嚴重挑戰，要能做到不仰賴標記訓練資料的非監督式學習(Unsupervised Learning)與自監督式學習(Self-supervised Learning)，才有機會以無招勝有招，讓偵測方法自動與時俱進不斷演化進步。第二，道高一尺魔高一丈，不實訊息與霸凌文字不僅能夠被惡意使用者撰寫出來，目前也有惡意AI方法可自動產生能夠騙過偵測演算法的生成式訊息(Generative Messages)，可巧妙地被辨識為在正常不過的訊息，在設計並訓練偵測模型的同時，也必須考量到與惡意AI進行對抗訓練(Adversarial Training)，讓偵測演算法學習與惡意AI提前過招、識得各種編造手法，方能變得不僅準確也更加穩健。第三，偵測各種資訊失序固然是杜絕其擴散的有效手法，但若能搭配真實、正確且乾淨不受汙染的訊息一起讓使用者服用，才是一種積極改善現有網路環境的手法，以假訊息擴散為例，數據科學扮演的角色也得準確推薦正確的訊息給使用者，或者把特定訊息的不同角度立場相關資訊一併呈現出來，把決策權交回到使用者手中，使其判斷不易受到同溫層與片面資訊來源的影響，因此如何在偵測假訊息的同時推薦對應查核過的真實資訊，也會是技術研發上的重要方向。

數據科學仰賴從資料中學習對預測目標有幫助的隱藏線索，而社群媒體正好提供大量的使用者數位足跡，不僅對於分析各種社會現象、了解使用者心理有所幫助，更能引發具有商機的應用，如精準廣告投放、個人化商品推薦，因為每一位使用者的需求、喜好或網路行為目的皆有機會從數據中被探勘拼湊出來，也因如此，對於不實訊息的杜絕和網路霸凌的預警，我們才有機會透過數據科學的手段切入，讓人工智慧成功扮演鍵盤柯南的角色，從使用者數位足跡中學習精準打擊假訊息和網路霸凌。隨著人工智慧技術一日千里的進步和突破，運用數據科學杜絕使用者被假消息誤導或遭受網路霸凌，將指日可待。