利用 AI 技術偵測假新聞之實證研究*

王維菁、廖執善、蔣旭政、周昆璋**

摘要

數位媒體科技日新月異，為人類創造無比便利的溝通工具，卻也讓我們身處於假新聞與假訊息充斥的環境。為了因應假新聞的氾濫，本研究運用人工智慧（Artificial Intelligence, AI）與自然語言處理 NLP 技術，結合深度學習方法，設計出一套新聞可信度評估系統模型。本系統經測試後辨識率最高可達 90.53%。然 AI 決策的可解釋性、人類自主性與社會過度依賴等問題皆是運用 AI 於假新聞上需深思者。本系統除了作為網路使用者在評估新聞是否真實可信之輔助工具，也期盼提醒民眾對於假新聞有所警戒。

關鍵詞：人工智慧、自然語言處理、社群網路、假新聞

* 本文特別感謝學刊審查委員與編委會給予的寶貴建議；一併感謝中央研究院 CKIP Lab 中文詞知識庫小組提供中文斷詞程式供研究團隊使用。本文也感謝科技部研究計畫的支持，計畫編號為 MOST 109-2634-F-003-08。

** 王維菁為國立臺灣師範大學大眾傳播研究所教授兼所長。廖執善為臺灣師範大學大眾傳播研究所博士後研究員；蔣旭政為國立臺灣師範大學大眾傳播研究所副教授；周昆璋為國立臺灣師範大學大眾傳播研究所研究助理。通訊作者為廖執善，Email: csliao@ntnu.edu.tw。

投稿日期：2020/02/24；通過日期：2020/09/21

DOI: 10.3966/172635812021060039002

壹、研究背景與動機

新聞是人類集體社會生活中重要的公共事務資訊來源，協助社會及社會成員依據所獲得的訊息進行正確的決策與判斷，並且是民主社會中監督政府與權力者重要的機制。不論處在何種科技社會生態環境，媒介與傳播是大多數人瞭解外在世界與客觀真實的重要工具，訊息傳播的性質，也建構塑造人們對社會世界之認知（McQuail, 2010；梁瑞祥，2001）。扭曲的訊息與社會圖像，影響媒介傳遞資訊、監督整合社會及社會化過程，威脅個人、國家與社會利益。此外，錯誤虛假的新聞訊息不但影響閱聽人對新聞事件之認知、態度及行為判斷，也可能導致新聞關係人權益受損害（徐佳士，1974；鄭瑞城，1983；羅文輝、蘇蘅、林元輝，1998）。是故，新聞的真實與正確，是社會傳播發揮基本功能的先決條件，也是新聞最重要的原則，新聞內容若虛假不實，可能對社會造成難以估計的傷害，傳播亦失去其存在價值（徐佳士，1974；羅文輝、蘇蘅、林元輝，1998）。

但近十年來，新聞傳播科技、新聞傳播機構以及新聞傳播途徑已有若干變化，人們接獲新聞的管道及傳播新聞的平台有相當比例轉移至網路與社群媒體，新興網路社群媒體興起，建構了新的新聞產業市場、新的新聞內容特性、以及新的新聞消費使用型態，也若干程度重構了新聞與社會、文化及民主間之關係。Twitter、Facebook 等社群媒體平台的新聞傳播與過往主流媒體新聞傳播模式不同，缺乏嚴謹的編輯守門過濾及事實查核判斷就可以在用戶之間中繼內容，也因如此，在某些情況下，沒有追蹤紀錄或嚴謹社會信任聲譽的個人用戶或平台也可能一時間吸引到與主流媒體一樣多之讀者，並對經濟、政治、社會產生一定的影響力，此也是假新聞與假訊息網路氾濫現象的重要基礎因素（Hermida, 2010; Newman, Dutton, & Blank, 2012）。

另外，社群媒體對傳統新聞媒體的影響日積月累，已造成傳統的單一權威訊息來源機制的瓦解，基於個人觀點與親身體驗的相關話題越來越多（Hartley, 2010），而蓋過事實與真相，使得目前的社群媒體出現後真相時代（The Post-Truth Era）現象，意指人們憑著個人立場篩選訊息，不再思考事件本身的真實性（Macdonald, 2018／林麗雪、葉織茵譯，2018）。相關現象表徵了當代政治生態的若干解構，傳統被認為具權威的消息來源真實性一再被質疑，取而代之的是小道

消息及未經新聞專業守門的資訊大量流竄，佔據輿論市場上風。新聞的「真實」（Reality）不再根據「事實」（Truth），反而受大眾觀感影響。如同許多其他新興科技，一種新科技或是新的技術普及對社會的衝擊以及後續的問題通常是在科技與技術普及之後才逐漸浮現，現今各種假新聞已埋伏潛藏在各種社群媒體中，社群媒體帶來的「假新聞」約在 2016 年美國總統選舉期間受到各方重視，眾多假新聞訊息出現被認為可能左右影響了選舉結果（Haldevang, 2017）。而假新聞及假訊息造成的混亂確實也隨著社群媒體的興盛快速地在世界各地發生，如 2017 年網路假消息指卡達政府支持恐怖主義，因此導致阿拉伯國家大規模與其斷交，經調查後發現，引發爭議的卡達官方訊息是由於卡達國家新聞局遭駭客入侵，被蓄意植入錯誤之資訊，已造成足以影響外交及國家安全的嚴重危機（李士傑，2018）。而臺灣也屬於身陷假新聞與假訊息傷害之國家。臺灣 2018 年 9 月強颱「燕子」襲擊日本時，眾多臺灣媒體引述中國「觀察者網」報導，引發國人抱怨駐日代表處，最終前駐日大阪辦事處長蘇啟誠自殺（蘇威全、方璇，2019 年 3 月 5 日）。同年九合一選舉期間出現大量假新聞與假訊息充斥社群網站甚至新聞媒體，政府亦多次警告敵對國家可能利用臺灣的公共言論自由來誤導與達成社會混亂之目標（Horton, 2018）。利用網路社群發達及缺乏監管的社群媒體平台，有系統、有組織地以產業鏈模式，大量製造假新聞與假訊息，似乎已成為民主社會的重大威脅。假新聞假訊息攻擊民主選舉非美國或臺灣獨有，近年進行重要選舉的國家，從北歐瑞典、南美巴西、東南亞馬來西亞，幾乎都無一能置身其外（乾隆來，2018 年 11 月 7 日）。因此，假新聞假訊息帶來的問題除了「事實」逐漸顯得不再重要，「事實」的地位失墜後所帶來的言論意見極端化與社會分裂和混亂，造成民主社會的巨大傷害，且當選舉結果得以境外化、商業化、目的化操作後，民主制度岌岌可危，虛假新聞與訊息導致的問題已非個人社群的私領域問題，已然嚴重威脅民主社會基礎。

面對假新聞興起，許多國家與組織致力發展各種解決手段，包括成立新聞事實查核組織、推廣網路與媒體素養教育、法規要求平台責任等。而近年人工智慧（Artificial Intelligence）因大量數位資訊數據易於蒐集、自然語言處理、類神經網絡分析等深度學習技術之成熟，大幅提升其實際效能，故其優越的文本分類能力也逐漸被寄望用來協

助分類判斷網路與社群媒體上的假新聞。運用人工智慧系統來協助分類判斷假新聞的優勢在於，將大大提升假新聞防制效率（賴燕芳，2018）。

因此基於上述，本研究針對網路上來源不明的新聞，規劃繁體中文新聞之 AI 可信度評估系統，本研究認為若能夠藉由人工智慧與大數據方法，提供讀者對未經驗證新聞之警示，提高閱讀者的警醒認知，或許可以減少假新聞的傳播與社會影響，並或讓社群媒體平台的管理者更能掌握假新聞的訊息生產與傳播組織化的來源，進一步進行平台的媒體自律把關。因此本研究開發繁體中文新聞 AI 可信度評估系統，並驗證測試其辨識未經驗證新聞之效能，希望為臺灣的假新聞問題帶來另一種解方或思考。

故本論文的研究目標在於為臺灣開發一套 AI 新聞可信度評估系統。由於人工事實查核新聞的速度較為緩不濟急，本團隊希望藉由語言特徵與傳播路徑等特徵訓練出穩定的 AI 新聞可信度評估模型，並透過此系統新聞的可相信程度（也可以當作是某種類型之「新聞可信度評分」），將測試用的新聞範例文章（某篇新聞內容）載入系統後，經過程式處理並判斷該新聞的可相信程度為多少百分比。我們預計的系統效能為辨識新聞的可相信指數至少高達 70% 以上，以協助網路使用者與網路平台辨別新聞真實性，而不盲從相信並成為謠言散佈者。本研究也希望相關結果能夠提供產官學各單位，在假新聞因應議題上不同的思考方向，並提供對大眾另一種數位傳播素養教育工具，並希望未來能與社群網站平台合作，提供遏止假新聞傳播之可能方法。

貳、文獻探討

一、假新聞的定義與分類

本研究採用的假新聞定義乃來自何吉森（2018）以及 Allcott & Gentzkow（2017），其指出「假新聞」乃指刻意以模擬、類似新聞的形式，來傳遞虛假及錯誤的訊息，目的在誤導大眾，以獲取政治或經濟利益，且假新聞應可證實為假。至於假新聞的訊息範圍如：

（一）故意製造的錯誤訊息、圖片或視頻，旨在傳播後誤導他人
（二）篡改訊息、圖片或視頻，以及分享老照片但聲稱是新照片，來

欺騙他人

（三）沒有惡意的諷刺或改編，但是愚弄他人的信息

另在假新聞的分類上，根據 First Draft 的分類（Wardle, 2017），可將假新聞分為七類如下：

（一）Satire or Parody：No intention to cause harm but has potential to fool.（諷刺或嘲諷：無意造成傷害，但有可能愚弄他人。）

（二）Misleading content：Misleading use of information to frame an issue or individual.（誤導性內容：誤導性地使用信息來描述問題或個人。）

（三）Imposter content：When genuine sources are impersonated.（冒名頂替者的內容：冒充真實來源時。）

（四）Fabricated content：New content is 100% false, designed to deceive and do harm.（虛假內容：新內容是 100% 虛假的，旨在欺騙和造成傷害。）

（五）False connection：When headlines, visuals or captions don't support the content.（錯誤連接：當標題，視覺效果或字幕不支持該內容時。）

（六）False context：When genuine content is shared with false contextual information.（錯誤的上下文：當真實內容與錯誤的上下文信息共享時。）

（七）Manipulated content：When genuine information or imagery is manipulated to deceive.（操縱的內容：操縱真實的信息或圖像進行欺騙時。）

而 Tandoc, Lim, & Ling 於 Digital Journalism 中，針對 2003 年至 2017 年間使用「假新聞」一詞的 34 篇學術文章進行研究，將這些定義從兩個維度：事實級別和欺騙級別，區分出假新聞為六種類型：（1）「新聞諷刺」（News Satire）、（2）「新聞模仿」（News Parody）、（3）「新聞謊言」（News fabrication）、（4）「照片合成」（Photo manipulation）、（5）「廣告與公關」（advertising and Public Relations）、（6）「宣傳」（Propaganda）（Tandoc, Lim, & Ling, 2018）。Zannettou, Sirivianos, Blackburn, & Kourtellis 則於 Data and Information Quality 期刊中，針對網路上的虛假訊息，包括謠言、

假新聞、點擊誘餌、惡作劇等，提出以下八種虛假訊息的分類：（1）虛構的、（2）宣傳、（3）陰謀論、（4）惡作劇、（5）偏見或單方面、（6）謠言、（7）點擊誘餌、（8）諷刺新聞。Zannettou 等人也提出了五種傳播背後的動機：（1）惡意意圖、（2）影響力、（3）搬弄是非、（4）利益、（5）熱情。（Zannettou, Sirivianos, Blackburn, & Kourtellis, 2019）。

綜合以上所述以及相關研究提出之看法，本研究歸納假新聞之分類範疇，常有以錯誤之程度區別之虛假程度，以及依作者之欺騙動機等兩個維度的重要分類概念。

二、假新聞辨識系統相關文獻

運用人工智慧運算邏輯建立分類器，以分辨網路與社群媒體上的假新聞、假訊息、謠言或特定立場類型之新聞，國際上已經陸續出現相關研究。如 Pérez-Rosas 等人提出一種內建於線上新聞網站中假新聞的自動辨識系統（Pérez-Rosas, Kleinberg, Lefevre, & Mihalcea, 2018）。Pérez-Rosas 等人將採用三種類型的假新聞，第一種是嚴重的捏造（即關於虛假和不存在事件的新聞項目或名人八卦等訊息）；第二種是惡作劇（例如，通過社群媒體提供虛假訊息，意圖被傳統新聞網站蒐集）；第三種是諷刺（即模仿真實新聞但含有諷刺和荒謬的幽默新聞）。在資料集收集方面則採用兩種做法，在真實新聞方面收集主流媒體、組織性、且合法的新聞資料集，並歸類為六個不同領域（體育、商業、娛樂、政治、技術和 教育）的，這些新聞來源是來自主要在美國的各種主流新聞網站，例如：ABC News、CNN、USA Today、New York Times、Fox News、Bloomberg、CNET，每一個領域各收集 40 條新聞，合計 240 條新聞。然而，在假新聞方面採用眾包假新聞的作法，設置了一個 AMT（Amazon Mechanical Turk）的任務，要求工作人員盡可能在他們的寫作中模仿新聞風格，以便獲得具有同質寫作風格的新聞，其做法為要求工作人員生成給定特定新聞的假版本，並避免不切實際的內容與保留新聞中所提到的名字，總共收集了 240 條假新聞（與真實新聞的數量相同）。第二種作法為只收集單一領域的新聞（名人新聞），Pérez-Rosas 等人試圖從網際網路中選擇公眾人物（例如：演員、歌手、社交名流和政治家）的新聞，因為他們經常成為謠言、惡作劇和虛假報導的目標。資料來源包括娛樂周

刊、人物雜誌、RadarOnline 等線上雜誌以及其他小報和娛樂導向的出版物。為了確定給定的名人新聞是否合法，文章中的聲明使用「GossipCop.com」之類的八卦檢查網站進行評估，並且還與來自網絡上其他娛樂新聞來源的訊息進行交叉引用。數據的組成是成對收集的，一篇文章是合法的，另一篇是假的，合計共 500 篇新聞文章。從 AI 技術面來看，該系統檢核方式乃運用詞彙（lexical）、句法（syntactic）與語意（semantic）三者的資訊組合，及文本可讀性的特徵，並採用 TF-IDF 演算法、SVM（Support Vector Machine）分類器與 LIWC（Linguistic Inquiry and Word Count）工具，其假新聞自動辨識系統辨識準確率約達 76%。實驗結果顯示與合法及主流媒體組織的新聞內容相比，假新聞的內容確實與一般新聞存在重大差異，主要給予了兩方面提醒：一是依賴語言學驅動的方法，要區分真實與虛假新聞內容，有必要查看假新聞的詞彙、句法與語意。第二是假新聞的作者比合法新聞的作者使用更多的副詞、動詞與標點符號，顯示一般新聞與假新聞因寫作目的的不同，其內容特質確實也會有所差異。研究結果也發現，政治、教育、科技領域的偵測準確性較高。相比之下，體育、商業和娛樂等領域的就沒有這麼明顯，而在表現最好的模型中，其準確度已與人類檢核假新聞的表現相當。

Wynne 和 Wint 提出了一種假新聞檢測系統，該系統考慮了在線新聞文章的內容，使用兩個 kaggle 數據集（真實或偽造新聞數據集和虛假新聞檢測數據集）測試 n-grams 模型是否能夠檢測虛假新聞。其中真實或虛假的新聞數據集包含 6,256 篇文章（一半被標記為真實，而另一半則為虛假）。另一個數據集則為虛假新聞檢測數據集，包含 4,009 篇文章（內含 2,137 個欺騙性新聞和 1,872 個真相新聞）。在 AI 技術方面，採用單詞 n-gram 和字符 n-gram 分析研究兩種機器學習模型，並結合帶有詞頻倒置文檔頻率（TF-IDF）的字符 n-gram 技術，實驗結果發現字符 n-gram 模型優於單詞 n-gram 模型，而梯度增強分類器使假新聞辨識準確度可高達 96%（Wynne & Wint, 2019）。Khattar 等人則提出了一種使用於檢測假新聞的端點到端點網路模型 MVAE（Multimodel Variational Autoencoder）多模式變異自動編碼器，利用微博和 Twitter 收集兩個標準的假新聞數據集進行了廣泛實驗。實驗結果顯示，在這兩個數據集上，搭配 Word2Vec 模型、VGG-19Net 網路、LSTM 與 Bi-LSTM 演算法，將 MVAE 與其他六種

方法 Textual、Visual、VQA（Visual Question Answering）、Neural Talk、att-RNN（attention RNN）、EANN（Event Adversarial Neural Network）進行對比，MVAE 平均比其他方法的精度高 6% 左右，而 F1 分數則只有 5% 左右（Khattar, Goud, Gupta, & Varma, 2019），為之後的模式方法提供不錯的方向建議。

Potthast 等人則運用超黨派新聞（左翼和右翼）之立場特質與主流新聞寫作風格之差異作為探究切入點，希望建立風格或假新聞的辨識器（Potthast, Kiesel, Reinartz, Bevendorff, & Stein, 2018）。該研究判定新聞為假的標準是由 BuzzFeed 的專業記者進行了手動事實檢查確認的。他們的數據集來自九個知名的出版者，其中包含三個左翼、三個右翼和三個主流新聞網站，總共 16,274 篇新聞報導。Potthast 等人所蒐集的語料庫包含 299 條假新聞，其中 97% 來自左翼右翼出版者。其研究貢獻是提出並展示了一種通過 Unmasking 評估文本類別之間樣式相似性的新方法（一種最初乃為驗證作者身份的元學習方法），其結果揭示了左翼新聞和右翼新聞的寫作樣式比起主流新聞，有更多的相同之處。並且他們使用隨機森林分類器（Random Forest Classifier）來區分是否為超黨派新聞，他們的實驗表明，基於樣式分類器（Style-based Classifier）的方式可以以約 75% 的準確度來區分超黨派新聞，但當他們使用相同的分類器來識別虛假新聞或真實新聞時，準確性僅約 55%。

Shu 等人（Shu et al., 2019）研究社群媒體上的用戶個人資料對於虛假新聞偵測的作用，並使用 Fake News Net 假新聞基準數據資料庫，蒐集兩個事實查核平台 Politifact 與 Gossipcopy 資料，在用戶數量上分別為 159,699 與 209,930 人次，真假新聞各佔 50%，以 Politifact 為例，真假新聞各為 361 篇文章，以 Gossipcopy 為例，真假新聞各為 4,513 篇文章，結合隨機森林演算法將 UPF（User Profile Feature）特徵表示法與其他四種 RST（Rhetorical Structure Theory）、LIWC、RST 串聯 UPF、LIWC 串聯 UPF 特徵表達法做對比，實驗結果發現 LIWC 性能比 RST 好，RST 串接 UPF 比 RST 或 UPF 性能好，而能有效辨識出真假新聞。Katsaros 等人（Katsaros, Stavropoulos, & Papakostas, 2019）則利用三個公開可用的資料集：Liar, liar pants on fire（訓練集為 10,269 條）、Signalmedia 信號媒體（1,000,000 則新聞）、Kaggle 資料集 Getting Real about Fake News（13,000 篇文

章），提出針對假新聞檢測的八種機器學習演算法的綜合性能評估，即 L1 正規化邏輯回歸、C 支持向量分類、高斯樸素貝葉斯、多項式樸素貝葉斯、決策樹、隨機森林、多層感知器和卷積神經網絡，並配合 TF-IDF 與詞嵌入等技術。實驗結果發現，比較幾個不同的數據集和各種績效指標上的有效性和效率，發現類神經網絡中的卷積神經網絡是性能最好的演算法，但缺點是需要大量的訓練時間。

而 Vosoughi 等人調查並收集從 2006 年至 2017 年，在 Twitter 上已發布且透過 6 個獨立的事實查核組織確認過判定為假新聞之推文，總計約 12,600 則推文，作為研究資料集（Vosoughi, Roy, & Aral, 2018）。從假新聞的定義角度來看，其判定 Twitter 推文為假的標準是透過 6 個獨立的事實查核組織確認的。而從新聞的虛假類型角度來看，Vosoughi 等人將所收集的資料集之假新聞分為三大類型：真實貼文、錯誤貼文、混合貼文（部分真實、部分錯誤）。其研究結果主要有二大項，第一為依據推文的性質分為七大類，且根據實際的謠言比例從高到低，依序為政治、都市傳說、商業、恐怖主義、科學與技術、娛樂與自然災害。第二是在所有類別的推文中，虛假性質的推文都比真實性質的推文，其傳播速度要來得更快、更遠、更深與更廣泛，即使是在有機器人帳號活動的情況下，其主要結論仍舊沒有改變，這個研究告訴我們假新聞的社會傷害力實不容忽視。

Jin 等人之研究則是針對選舉期間的假訊息與謠言。其團隊對兩位美國總統候選人 Hillary Clinton 與 Donald Trump 的追隨者的謠言推文進行了全面分析（Jin et al., 2017），其判定 Twitter 推文為謠言的標準是採用主流的謠言揭穿網站 Snopes.com 上檢查過的謠言作為客觀的樣本。從 AI 技術面來看，Jin 等人採用 TF-IDF（Term Frequency-Inverse Document Frequency）、BM25（Best Matching）、Word2vec（Word to vector）、Doc2vec（Document to vector）等四種演算法，其謠言推文辨識率 accuracy 分別為 79.5%、79.9%、55.7%、65.8%，不同演算法呈現了不同的辨識效能。Jin 研究發現對於 Hillary Clinton 與 Donald Trump 的追隨者而言，排名前 10% 的用戶發佈了約 50% 的謠言推文，排名前 20% 的用戶發佈了約 70% 的謠言推文。兩個追隨者團體都有發佈關於他們最喜愛的候選人和對手候選人的謠言，候選人支持者會以謠言作為消極競選策略，也會散佈有關對手的謠言。Allcott 與 Gentzkow 同樣針對美國總統大選，收集 2016 年選舉過程中

假新聞在 Facebook 上的文章（Allcott & Gentzkow, 2017）。Allcott 等人將假新聞定義為有意且可證實是虛假的，可能誤導讀者的新聞文章。在資料庫蒐集方面，採用三個獨立的第三方名單（資料收集期間為 2016 年大選前三個月），包括 Snopes（www.snopes.com）網站上擷取標註有 Donald Trump 與 Hillary Clinton 之所有文章；其次，從事實查核網站 PolitiFact（www.politifact.com）中抓取有 2016 年美國總統大選標籤的文章；第三從新聞媒體 BuzzFeed（www.buzzfeed.com）整理的 21 條在 Facebook 上引起廣泛關注的假新聞文章列表。而從新聞傳播的角度來看，Allcott 等人發現假新聞文章起源於以下三種類型的網站。第一種為建立網站來刊登故意製造和誤導性的文章，例如 denverguardian.com。這些網站的名稱通常會被命名為與合法新聞機構的名稱相似。第二種為在諷刺性網站上面刊登的文章，如果僅僅從上下文來看的話，可能會被解釋為事實，例如 wtoe5news.com。第三種為在事實文章（通常帶有偏見）和一些虛假文章之間混合呈現的網站，例如 Endingthefed.com。還有一個很重要的特徵，就是提供這些虛假新聞的網站，其網站存活時間通常是短暫的，因為在 2016 年大選前夕，這些帶有特定目的的網站已不復存在。

Bode & Vraga 則特別針對社群媒體例如 Facebook，在糾正錯誤訊息中可能扮演的角色，其研究課題為社群媒體的使用是否有助於強化或糾正錯誤訊息（Bode & Vraga, 2015）。Bode 等人在此將假新聞視為錯誤訊息，並說明錯誤訊息或事實誤解定義為存在或相信客觀上不正確的訊息。為了研究社群媒體可能影響錯誤認知的建立與糾正方式，Bode 等人開發出一種實驗設計，該設計利用 Facebooks 內部的一項功能，當用戶於 Facebook 上點擊 Facebook 中的鏈結時，該功能會提供相關文章的連結（即該鏈結會顯示在鏈結下方），該鏈結是由匹配故事中的主題之演算法生成的。Bode 等人藉由向用戶顯示包含錯誤訊息的帖子，然後操縱相關故事來確認、糾正或同時確認與糾正錯誤訊息。調查結果表明，測試該 Facebook 內部功能在增強或打擊誤解中所起的作用是取決於這些相關文章所提供的訊息，也就是說當相關文章糾正了錯誤訊息時，錯誤認知會大大減少。

國內研究部分，目前尚無較具體的假新聞機器辨識相關研究，但類似研究如鄭麗珍等人提出了虛假評論的機器檢測模式，以協助消費者面對購物網站上虛假且刻意誤導消費者的購物評論（鄭麗珍、江彥

孟、游政憲，2019），該研究採用傳統的文本挖掘技術來處理文本數據，包括詞袋 Bag-of-words、潛在語義分析 LSA（Latent Semantic Analysis）和用於詞表示的 Word2vec 以及使用機器學習來訓練模型以檢測偽造的評論，包括 SVM，深度神經網絡 DNN（Deep Neural Network），卷積神經網絡 CNN（Convolutional Neural Network）和長短期記憶 LSTM（Long Short-Term Memory）等技術，並建立多個以「寫手為基礎應用深度學習偵測虛假評論的模型。實驗結果顯示使用詞袋模型表現是最差的，使用 LSA 和 Word2vec 表現的效能很接近。在各種分類器的識別效果中，發現 LSTM 配上 Word2vec 效能表現的最好，而使用淺層的神經網 SVM 配上 Word2vec，和使用 DNN 配上 Word2vec 效果很接近，研究結果顯示分類器初步確實可以達到辨識虛假評論以保護消費者的目標。

故綜合國內外研究來看，運用人工智慧演算法來分類文本，以區辨虛假內容，包括假新聞、謠言、虛假評論等，已逐漸被證實是可行以及極具發展潛力者，而在中文世界與臺灣也還未有相關的研究、嘗試或開發，因此這也將是本論文對臺灣學界最主要的貢獻。

參、研究方法與設計

由於臺灣目前並沒有具規模的假新聞資料庫或資料集，現今經由人工查核的不實新聞數量也尚難以達到足夠規模可以用來訓練程式，因此本研究僅能使用假新聞主要來源的內容農場作為虛假新聞之資料集，以及經由主流新聞媒體編輯守門把關的新聞為一般新聞資料集。我們的邏輯基礎基於前一章節假新聞的定義，是「刻意以模擬、類似新聞的形式，來傳遞虛假及錯誤的訊息」，因此主流新聞組織的新聞雖然也有錯誤的可能，包括客觀錯誤或主觀錯誤，但其並非是「刻意以模擬、類似新聞的形式，來傳遞虛假及錯誤的訊息」，且假新聞與一般純淨或專業新聞學所訴諸的社會目的並不相同，在內容的語言特徵上也會有所不同（P´erez-Rosas et al., 2018），因此程式應可區辨內容農場的虛假新聞與一般主流新聞組織專業把關後的專業新聞，以下也針對本研究的資料集資料來源和研究方法進行詳細說明：

一、資料集資料來源

本研究關注網路上未經證實新聞內容之機器識別，故研究蒐集網路上之新聞平台、內容農場所提供的假新聞與一般新聞做為資料來源，而社群網站或即時通訊 APP（例如：Facebook、Instagram、LinkedIn、Twitter、Line、WeChat）個人間或小群體間私領域的訊息交換與傳散，則尚不在本研究之研究範圍內，未來若有機會可持續發展並納入考量。本研究資料集中一般新聞與假新聞的蒐集範圍與定義說明如下：

（一）一般新聞以臺灣傳統四大報「聯合報」、「自由時報」、「蘋果日報」、「中國時報」的網站平台作為資料收集範圍。
（二）假新聞則以四個知名的假新聞內容農場（怒吼、密訊、琦琦看新聞、寰球軍事網）為資料收集範圍。
（三）資料庫 schema 欄位包括：新聞標題與新聞內文為必要欄位。新聞作者、出處（包含新聞網站名稱與該篇新聞網址鏈結）、新聞發布日期與新聞分類為非必要欄位。

本研究以新聞媒體其編採體制之完整作為一般新聞資料庫建立之參考，本研究以爬蟲程式抓取主流媒體之四大報作為一般新聞資料庫來源，正是依據其長年以來建立之媒體信譽以及完整的專業編採體制。

假新聞資料庫則是以抓取內容農場之新聞為主，選取內容農場當作本研究的假新聞資料來源的理由是，內容農場乃指為了創造流量、賺取網路廣告分潤所建立的網站，它們多以各種合法、非法手段大量生產文章，原創性內容少且內容的真實性難以確認，由於內容農場並不主動管理內容，許多文章是由侵權盜用、抄寫、改寫而來，常摻雜浮濫內容與不實訊息，形成漏洞，因此，在缺乏新聞媒體的編輯規範和過程中，無法確保新聞內容的準確性和可信性。在此須注意的是內容農場之新聞虛假類型與程度彼此不同，有的是各處剪貼，有些採用無可信度之消息來源，又有些是直接造假，因此，本研究將所收集的內容農場新聞與四大報內容交叉比較，若有發現新聞標題與內容本文完全相同的新聞，就將其排除於假新聞資料庫之外。資料庫建構過程中，真假新聞資料的豐富度與乾淨程度將影響往後發展辨識系統與深度學習之結果，因此在資料庫如何建構以及如何分配這些資料用於學

習與驗證是需要謹慎進行的。

二、方法與設計

本研究使用 Python 語言與 Tensorflow 工具開發一套新聞可信度評估系統，系統說明如下：

（一）問題描述：輸入一篇符合 schema 格式的新聞，透過新聞可信度評估系統模型，輸出該文章為假新聞的可相信程度。

（二）每一篇 schema[1] 格式，其欄位包括：id（新聞識別碼）、website（新聞所屬網站名稱）、date（新聞發布日期）、title（新聞標題）、content（新聞內文）、url（新聞網址）。

（三）系統輸入：訓練資料集為大量符合 schema 格式的一般新聞與假新聞。

（四）系統輸出：辨識每一筆資料為假新聞的可相信程度。

從 AI 新聞可信度評估系統架構面向上，本研究的新聞可信度評估系統採用監督式學習（Supervised Learning）方式，監督式學習的步驟如圖 1 所示，說明如下：

在數據前處理方面，首先，我們先討論決定一般新聞與假新聞的網址，並確認資料庫與資料表所需要的欄位，然後撰寫爬蟲程式將所需要的資訊標籤抓取下來寫入資料表內，並設定固定時間與頻率，定期執行爬蟲程式，將我們所需要的原始資料存入資料庫當中。

在探索式數據分析方面，本研究所需要的步驟流程如下 ：

（一）數據清理：從我們的資料庫中擷取需要的欄位，檢查每筆紀錄是否有欄位值為空的，若有則排除該筆紀錄，接著將所選取的資料匯出成帶有 UTF-8 BOM 的 CSV 格式檔案，提供給程式用來當作原始資料來源的檔案。

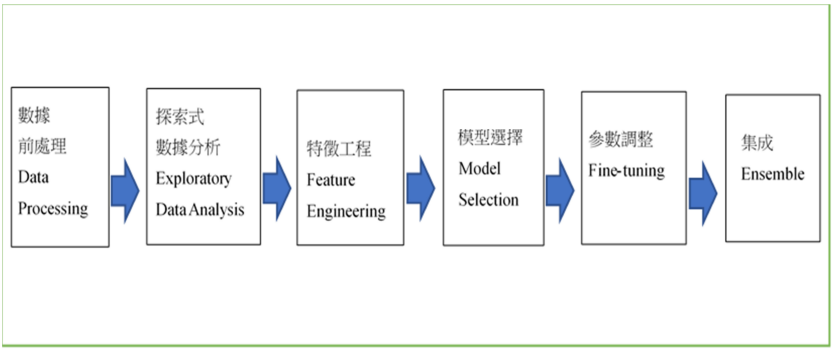
（二）特徵萃取：我們將來源檔案中的文章標題與文章內文當作字串，並採用中研院中文詞知識庫小組所開發之 CKIP 斷詞暨實體辨識系統，先做斷詞再剖析，並執行 POS tagging（Part-of-speech 詞性標記）與 NER（Name Entity Recognition 命名實體識別），接著計算每個斷詞之後的詞性頻率，將異常值（Outliers）排除，在此異常值之門檻定義為詞性頻率低於總筆數 10%，剩下的皆收集用來當作特徵參數。

（三）建立模型：我們採用 Keras Sequential 順序模型（是一種多個網路層的線性堆疊模型），並指定輸入層、隱藏層與輸出層的尺寸參數，提供給編譯模型使用。

（四）訓練與驗證模型：我們採用 model.fit 與 model.evaluate 方法，並輸入相對應的參數值。

（五）資料視覺化：將模型訓練完畢後產生的數據，使用 matplotlib 指令將數據繪製成曲線圖，並儲存成圖片且將檔案放置於 Result 資料夾中。

圖 1：監督式學習步驟



資料來源：本研究製作

在新聞內容檢核方面，本研究透過人工智慧（AI）作為偵測未經驗證新聞的工具，並使用語言特徵與訊息傳播路徑兩種偵測方法取徑，語言特徵中我們使用 NLP（自然語言處理）的文本分析，包括文字的分詞詞性、語塊、斷詞與上下文分析，結合 DNN（深度神經網路）技術。由於 DNN 與過去機器學習需要由人來提供規則的方式相較，DNN 只需要在設計好的神經網路中，藉由訓練資料與參數的微調設定，機器就能自行學習、找出特徵，並藉由 GPU 等硬體支援加速調校訓練模型，來提升 DNN 模型的精準度。

從程式碼的 model.add、model.compile 與 model.fit 等三個函式當中，可以得知其參數設定，並針對所使用的參數細節分別說明如下:

（一）model.add 所使用的參數有輸入層的特徵維度 dim（84）、輸入層節點 units（10）、輸入層激勵函數 activation（relu）、隱藏層節點 units（10）、隱藏層激勵函數 activation（relu）、輸出

層節點 units（1）、輸出層激勵函數 activation（sigmold）。

（二）model.compile 所使用的參數有損失函數 loss（binary_crossentropy）、最佳化方法 optimizer（rmsprop）、訓練指標 metrics（accuracy）。

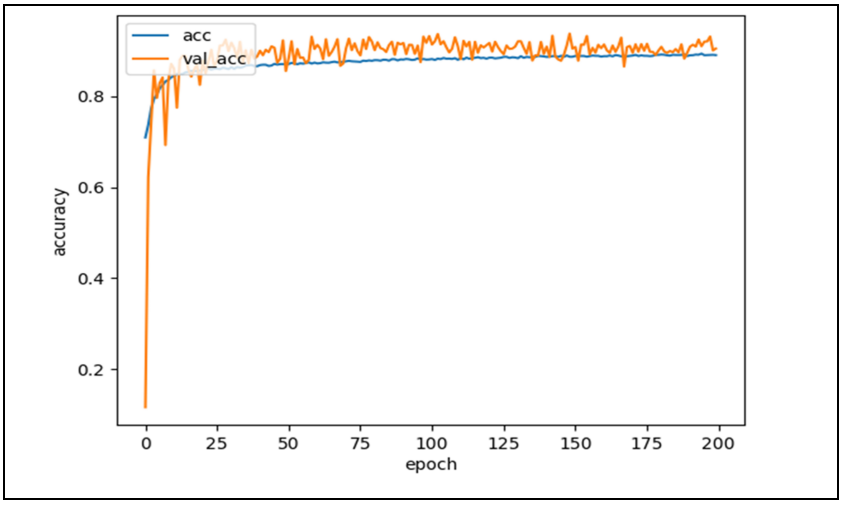
（三）model.fit 所使用的參數有批次尺寸大小 batch_size（100）、訓練回合次數 epocs（200）、驗證集切割比例 validation_split（0.3）。

本研究使用深度神經網路 DNN 的做法，並採用 logistic regression（羅吉斯回歸）的方式，在隱藏層中使用 ReLU Activations（激活函數）。接著，在模型編譯方面採用 binary_crossentropy losses（損失函數），並將輸出定義為 Sigmoid Activations，使得輸出值可以變成一個二元分類問題。在深度學習最佳化演算法上，我們採用 RMSprop Optimizers（Root Mean Squared Prop 梯度均方根下降），藉由計算微分平方加權平均數來慢慢丟棄先前取得的梯度歷史值。此方法的優點為消除梯度下降中的擺動，這樣一來就能夠防止學習率在找到最佳解之前過早地減小，而順利地收斂到最佳值。

肆、研究成果

我們的研究數據規模有 10,000、20,000、30,000、40,000、50,000 筆，經過我們多次的調整訓練模型參數，發現以 20,000 筆數據表現最佳，也就是說訓練集 14,000 筆（包括真假新聞各佔 50%，也就是 7,000 筆），測試集為 6,000 筆（包括真假新聞各佔 50%，也就是 3,000 筆），其中所建立的模型參數為 dim（84）、node（10）、lr（0.001）、hidden_layer（10），其中藍色線條（acc）表示訓練集的正確率，橘色線條（val_acc）為測試集的正確率，經過 200 回合的訓練後，實驗結果辨識率最高為 90.53%，[2] 本系統的訓練模型正確率分布圖如圖 2 所示：

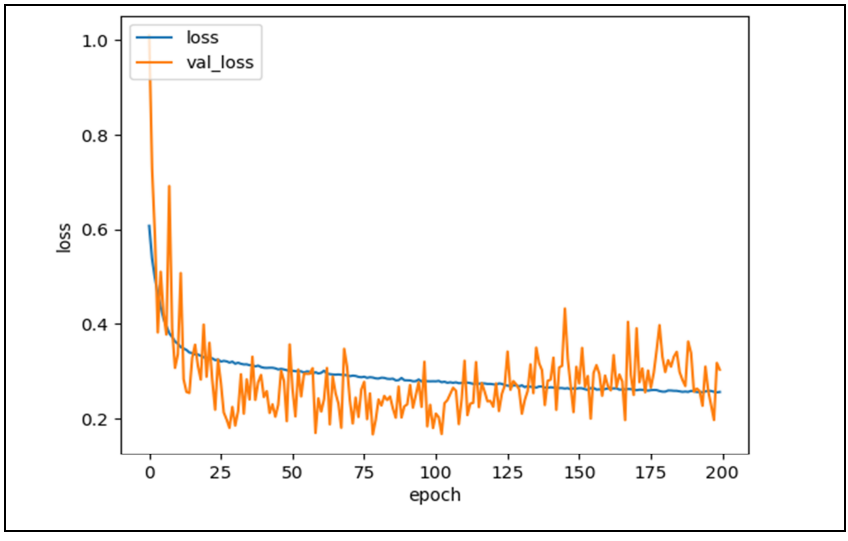
圖 2：模型準確率分布圖



資料來源：本研究製作

從 loss 損失率來看，其中藍色線條（loss）表示訓練集的損失率，橘色線條（val_loss）為測試集的損失率，經過 200 回合的訓練後，其收斂程度趨於穩定，本系統訓練模型損失率分布圖如圖 3 所示：

圖 3：模型損失率分布圖



資料來源：本研究製作

此外，DNN 深度學習訓練完成後，我們將該模型儲存起來，並額外整理出 800 篇新聞（資料庫內之一般新聞與假新聞各佔 400 篇）與 300 篇非資料庫內之新聞網站之新聞，提供給我們的模型做測試，測試結果如下表 1 與表 2：

表 1：資料庫測試集一般新聞與假新聞平均可信度列表

	一般新聞	假新聞
筆數	400	400
平均新聞可信度 %	89.95%	33.71%

資料來源：本研究製作

從上述表 1 之檢測結果指出，一般新聞檢測的新聞可信度約近 90%，但假新聞的可信度百分比僅約 33.7%，一般新聞與假新聞本程式給予的可信度百分比明顯差距很大，代表本研究 DNN 深度學習模型訓練成功，程式可辨別一般新聞與假新聞。

除了使用資料庫來源的資料數據做為測試集，我們也尋求外部資料來源來進行測試，以證明本程式的實際運用性與效能，我們分別使用非資料庫來源的中央社、東森新聞雲以及新聞內容農場「每日頭條」各一百則新聞進行測試，測試結果如表 2：

表 2：各項非資料庫內新聞網站之新聞可信度列表

新聞網站名稱	中央社	東森新聞雲	每日頭條
筆數	100	100	100
新聞可信度 %	81.32%	74.306%	12.22%

資料來源：本研究製作

從表 2 我們發現，中央社的新聞可信度最高，東森新聞雲的新聞可信度略低，但仍達七成以上，而新聞內容農場「每日頭條」的新聞可信度卻很低，平均僅約 12.22%，再次證明我們的系統能夠準確地偵測出新聞可信度，可作為協助讀者對新聞可信度之判斷。

另外，在我們的深度學習訓練過程中，我們採用中研院的 CKIP 斷詞系統將新聞內文作 POS 詞性分析之後的數據收集起來，人工觀察並進一步解讀與分析這些數據，可得知虛假新聞內文在普通名詞

（109 個／篇）、逗號（49.6 個／篇）、副詞（48.7 個／篇）數量級上，比一般新聞內文在普通名詞（74 個／篇）、逗號（33.4 個／篇）、副詞（26.8 個／篇）平均數量級上要高出 1.5~2 倍，顯示出虛假新聞內文在這三種用詞頻率上要比一般新聞來的多。此外，新聞內文經過 POS 詞性分析後，也發現從數量級差異來看，以「的之得地」差異倍數最大（一般新聞 12 個／篇、虛假新聞 30.5 個／篇），地方詞[3]次之（一般新聞 21.9 個／篇、虛假新聞 46.3 個／篇），顯示虛假新聞的新聞寫作贅詞率高，寫作精實度與專業新聞寫作仍有差別，但假新聞地方詞數量也遠高於一般新聞，其原因為何目前我們尚難釐清原因，仍需待更進一步的探究才能說明。

而除了經自然語言處理後一般新聞與虛假新聞間所展現出的差異之外，新聞長年以來的發展，已經是一種高度結構化的文體，為符合社會對其傳播資訊、反映事實功能的期待，傳統正式新聞組織講求純淨精實的新聞寫作、正反併陳、中立客觀等寫作技巧，藉由組織守門過程，這些原則與對新聞的要求已被記者內化融入每日編採寫作中。有別於媒體組織新聞的專業守門實踐，虛假新聞有著截然不同的目的，傳統新聞目的在於不涉價值判斷地傳遞有關事實之訊息，再由閱聽人自行進行評斷，但虛假新聞則是無法擺脫欺瞞、煽動與試圖影響人認知與態度之強烈意圖，兩者在目的上即有明顯不同，這也使得二者在呈現新聞內容與寫作手法上必須要有差異，因此藉由分析新聞寫作風格、寫作特質、新聞價值取向、新聞品質、確切的新聞消息來源、新聞作者、新聞組織與背景揭露等，確實可以明顯觀察到一般新聞與虛假新聞間之差別。

此外，在新聞編採與表現上，為達客觀中立、正反並陳之標準，新聞學要求對爭議性議題盡量呈現不同立場之價值意見，以達到形式中立與多元立場價值呈現，然而虛假新聞之目的在於誤導並煽動，使閱聽人相信，通常不會正反並陳，只會呈現單方面說辭，且用詞通常都煽動或誇大，引述之消息來源經常不明確或錯誤。在新聞寫作風格上，新聞記者的用字用詞等修辭層面，也提供了辨別一般新聞與虛假新聞的判斷基礎。

經由我們的系統得到之結果進行人工觀察，一般新聞與虛假新聞間之差異，可整理如表 3：

表 3：一般新聞與假新聞之差異整理

一般新聞	虛假新聞
● 自然語言處理部分（依新聞內文的斷詞與 POS 詞性分析來看）	
前三名依序為普通名詞（74 個/篇）、逗號（33.4 個/篇）、副詞（26.8 個/篇）	前三名依序為普通名詞（109 個/篇）、逗號（49.6 個/篇）、副詞（48.7 個/篇）
前十名依序為普通名詞、逗號、副詞、動作及物動詞、地方詞、狀態不及物動詞、介詞、括號、專有名詞、“的之得地”	前十名依序為普通名詞、逗號、副詞、地方詞、動作及物動詞、“的之得地”、狀態不及物動詞、介詞、句號、括號
從頻率來看比虛假新聞少	從頻率來看比一般新聞多，約有 1.5~2 倍
從數量差異來看，以「的之得地」差異最大（一般新聞 12 個/篇、虛假新聞 30.5 個/篇），地方詞次之（一般新聞 21.9 個/篇、虛假新聞 46.3 個/篇）	

資料來源：本研究製作

伍、結論與討論

國際上已陸續出現相關研究利用人工智慧優異的文本分類能力，來分類辨識假新聞、謠言與虛假評論等之相關程式系統開發，但國內尚無假新聞分類辨識系統，因此本研究旨在發展一套協助讀者評估來源不明或可質疑之新聞其可信度之程式，希望能在相關領域上為臺灣帶來有意義的討論與貢獻。

根據本研究的研究結果，繁體中文的假新聞可信度評估系統是可行且深具發展潛力的，目前實驗結果精確率最高可達 90.53%。而資料庫測試集測試檢測結果發現，一般新聞經檢測後的新聞可信度約近 90%，但假新聞的檢測其可信度百分比僅約 33.7%，一般新聞與假新聞本程式給予的可信度百分比明顯差距很大。而使用外部資料的測試集進行測試，中央社新聞測試後可信度最高，東森新聞雲測試後新聞可信度略低，但仍達七成以上，而新聞內容農場「每日頭條」測試後新聞可信度很低，平均僅約 12.22%，證實我們的人工智慧系統確實能準確分辨一般新聞與虛假新聞之不同，其偵測判斷新聞的可信度之準確程度，應能協助讀者對於新聞內文之語言特徵分析上新聞是否可

信之建議。此一研究將會把臺灣的假新聞議題或對假新聞之因應帶到一新的領域，一個是否願意嘗試運用人工智慧，並讓其介入人類社會生活認知判斷之領域。

但對於引進人工智慧介入人類的社會生活甚至社會認知，仍需要有一定的警醒與反思，包括本研究雖從程式判斷的結果進一步分析一般新聞與虛假新聞在新聞文本語言學特質、新聞寫作特徵、新聞相關呈現上等之差異，但這均是運用人類邏輯去分析，並不一定是機器判斷的邏輯或依據，而從這些文本語言之差異，程式與機器也並不會提供我們解釋的原因或機器判斷的因素，是故當我們要採用人工智慧的判斷來輔助我們人類的判斷甚至取代人類的判斷時，這樣的依賴其社會結果的負面性，包括過度依賴機器或程式，未來是否造成人工智慧的獨裁以及人類自主性與主體性的喪失，甚至人類思考判斷能力的下滑等，可能都是需要進一步深思與考量。

此外，人工智慧新聞可信度判斷程式對於未來新聞學與新聞業之影響會是什麼？被評比為低可信度的新聞內容組織就會因此失去民眾的信任與喜愛？AI 新聞可信度評估能否反饋為對新聞採寫時新聞正確性與專業性的要求？問題可能並不一定如此簡單，答案可能也會是難以預料的，因此這樣的程式應用對新聞學與新聞環境、產業、生態的影響等等，現今我們可能都尚難以評估，但卻仍應謹慎以對。

而現今幫助人們區分假新聞和一般新聞的現有方案仍依賴單純基於機器或單純基於人力的方法，但在電腦科學、人工智慧與心理學研究中已經表明單純依賴上述兩種方法均有其侷限性。然而假新聞的判斷方式並非僅有判斷內文語言風格，在本研究對於真假新聞之觀察中，真假新聞由於其傳散的目的以及作者和背後組織之不同，有著多種差異。在內文層面中假新聞較會有奇觀式的敘述方式、敘事性的寫作方式、情緒性用詞、誇大用詞、人身攻擊、妖魔化社會群體、陰謀敘事、甚至邏輯缺陷等，並且誤導式標題、誘餌式標題、甚至是要求閱聽人留言等特徵，上述也都是有違專業新聞學之準則，是傳統新聞較不會出現，但假新聞經常出現的。並且真新聞在處理新聞資料來源上會較為謹慎，並且會詳述消息來源；除此之外對於不同立場之訊息也多會盡量呈現，讓內文正反併陳。

而在組織標示或揭露上，正式新聞組織與虛假新聞網站組織性質會有不同，組織目標也不同，組織架構與成員特質也有所差異，組

織、成員與組織文化也造就真實與虛假新聞間之差異，與是否願意揭露之差別。一般而言，正式新聞組織依靠社會信任生存，為保持良好信譽，會有完整嚴謹的編採制度對新聞內容進行較嚴格的查證與校對，語法、拼字、用字與標點符號使用錯誤會較虛假新聞少上許多。並且正式新聞媒體組織由新聞從業人員組成，並對其生產的新聞內容負責，故新聞通常會明確標示採訪撰寫記者之姓名，即便記者不願記名，往往也會明確標註新聞組織之名稱以示對新聞負責。除此之外，正式新聞組織不吝於向公眾展示新聞組織背景，其網站上往往會有明確的組織負責人訊息、正式且固定的聯絡資訊，相反的，虛假新聞網站之目的在於欺瞞，且由於其非法性，通常都會隱瞞組織背景資訊，提供非正式組織使用的聯絡資訊。正式新聞組織為維持社會信任，會有正式且固定的網域名，但虛假新聞網站為避免被辨識出來，經常申請新的網域名，會有著不斷變動、與網站名無關或是看起來非常複雜的網址等。而以上諸多差異，都可證實假新聞的內容、寫作、特質與相關呈現，均與一般新聞確實有所不同。

這些真假新聞之間的差異可做為未來人工智慧對於分辨新聞可信度的發展方向，然而有些特徵現階段較難以藉由人工智慧判斷，如內文邏輯的正確性、是否在文中正反併陳、內文結構上是否保持新聞寫作的倒金字塔格式等。並且有些特徵並非皆適合設計進入人工智慧系統之中，有些特徵比更適合藉由人類進行判斷，例如新聞網站組織的資訊以及對於新聞內文資訊的事實查核。也因此在辨識真假新聞的領域中，人類仍有相當的重要性。

也因此 Okoro 等人（2018）提出一種混合模型（hybrid model），結合上述兩種檢測方法以檢測社群媒體上的假新聞，研究結果指出，混合式的方法較上述兩種方法單獨使用時更為準確。Okoro 等根據深思可能性模型，分作中央路徑與周邊路徑兩種訊息接收處理模式；研究認為，假新聞容易被相信可能是因為假新聞容易透過周邊路徑處理模式觸動人們，這表示需要一種支援系統，幫助人們判定哪些新聞具有欺騙性或者真實性。Okoro 等人提出的人機混合式模型（machine-human system），根據社群媒體新聞素養教育工具，以十個考量項目作為判斷假新聞的基準，包括對新聞標題保持懷疑態度、仔細查看網址、注意新聞來源、注意不尋常的格式、注意照片、檢查日期、檢查證據、參閱其他報導、報導是否只是玩笑、有些報導

是否故意錯誤等判斷面向。

且如黃俊儒（2018）指出，防範假新聞可以依賴 AI 判別、人工事實查核、立法遏止、生產者端需於文中附上第三方資料來源、以及推廣數位與資訊識讀教育等，但他認為端賴單一方向的解決辦法都可能不切實際。而打擊假新聞的難處在於不同政治或價值立場對個別假新聞的想法價值甚至定義等都可能有所不同，這已經不是一個資訊正確與否的問題，而是一個政治議題，恐很難依靠機器來協助解決，必須回到政治社會領域去面對。因此我們必需認知 AI 或科技技術可能無法徹底解決假新聞的所有問題，也未必能提供最好的解決方案，因此若僅是為了使用 AI 技術而使用 AI 技術，很可能會適得其反。因此多方位面向的方案包括技術、法律、道德、與教育等合作之因應方式，可能可以協助技術無法完全解決的問題，因人類社會的問題仍要回到人類身上來自己解決，AI 僅能視為工具或助手，而非人類政治社會問題的最終解答。

因此分辨假新聞的最佳作法，機器可能不該反客為主，最終仍需重回到人類讀者自己的判斷能力上著手，對新聞片段抱持懷疑，並有消息比對的能力，若因為太過相信機器，導致人們放鬆警惕，更願意相信自己所讀的內容，最終可能帶來更嚴重的威脅與不確定性。因為自始至終人自己才是社會問題之所在，人為何選擇相信假新聞？偏見、立場、價值均乃出自於人心，即使告知事實，人們可能仍不願意改變原來的相信，無法說服那些已經相信虛假訊息的人。人工智慧很難改變人們根深蒂固的偏見與選擇性理解，這必須從文化、社會與教育面向的解決著手。因此，我們能夠推理出數位資訊識讀教育與社會民主教育在面對假新聞議題時有其最終重要性。當數位世代崛起，民眾被區分為數位移民及數位原住民時，數位資訊與社會民主素養應為國家優先的社會教育目標，以培養具備新的民主社會素養能力的公民。

陸、研究限制與建議

由於國內目前缺乏繁體中文類型的假新聞辨識系統相關研究，因此，本研究的參考文獻以英語系之真假新聞機器辨識相關研究為主，而本研究由於在新聞分類上採取隨機取樣方式，不特別限定某些類型

新聞，如此也增加模型訓練上的複雜程度與困難度，但適用性也會較廣泛。此外，雖然本研究的新聞可信度評估系統辨識率高達 90% 左右，但是，由於此類型數據強烈受到資料來源之限制，亦即，若訓練的數據來源有所變更，例如收集非四大報的新聞或其他內容農場的新聞，可能就會影響系統辨識率，因此，未來須繼續擴張系統資料來源之廣度，以增加辨識系統的實用性，此亦是本研究未來繼續努力之目標，不過經由本研究，已證實藉由人工智慧系統辨識中文假新聞之可行性，此乃本研究最主要的價值，未來可繼續推廣相關應用，以對社會實務或學術研究創造更多貢獻。

其次，本新聞可信度評估系統之研究，與其他利用 AI 來偵測不同語系假新聞，或運用 AI 進行社會科學研究者，遭遇相同的問題，亦即人工智慧深度學習的黑盒子現象，也就是系統僅能告訴我們結果，但卻無法提供相關的解釋原因與其判斷之因素，也因此，檢測新聞真假程度是一回事，但如何向使用者解釋其虛假原因卻極具挑戰，如果我們不能提供一個適當的說明來解釋一則新聞被判斷為虛假的原因，那麼民眾可能也會因為缺乏對研究團隊或科技中立性的信任，使得民眾缺乏使用辨識系統的動機，或造成人們不接受系統辨識的結果，使得其影響範圍與效果有限，因此，本研究團隊提供下列兩點建議：

一、透過所收集的新聞原始資料以及經過數據清理與過濾產生的多項特徵，給予舉證範例進行一定的說明與解釋，以便呈現在輸出結果上，當作解釋辨識結果之輔助與使用說明，用來說服並提高使用者對於假新聞人工智慧辨識系統的信任程度。

二、對於機器辨識結果，隨機取得足夠樣本後，委請訓練有素且具社會公信力的新聞工作者和經驗豐富的編輯人員團隊檢視其合理性，並適時提出質疑，以做為機器學習之修正與反饋。

最後，礙於高品質標籤的虛假新聞資料仍然處於稀缺不足的狀態，除了建立更大規模的數據集外，我們也應當關注如何將無監督、半監督方法應用到假新聞檢測系統中，這也是相關研究未來可持續努力之處。

註釋

1 為了方便資料庫蒐集時識別出此篇未經驗證新聞的來源與出處，也因為擔心 url 鏈結會因時間的關係而無法鏈結，因此 id、website、date、url 皆僅當作輔助用途使用，而非判斷主要項目，主要項目為 title 與 content。

2 由於我們採用 RMSprop 梯度均方根下降的最佳化演算法，藉由梯度的計算（對於神經元的權重 weight 與偏差 bios 作偏微分）以及學習率的乘積，來不斷地更新權重與偏差值，以及經過數學邏輯與工程上的推導，我們的訓練模型在經過 200 回合的訓練後，損失函數（loss function）穩定地下降到趨近並收斂到最佳值時，其正確率顯示的值即為 0.9053。

3 「地方詞」乃中研院中文詞知識庫小組所開發的繁體中文斷詞系統 CKIP 中 POS（Part-of-Speech）Tags 詞性標注的分類之一，其意義為表地方概念的體詞，包括地方名稱、行政單位、以及可以看成是地方的各種機構等。

參考書目

何吉森（2018）。〈假新聞之監理與治理探討〉，《傳播研究與實踐》，8(2)：1-41。

李士傑（2018 年 5 月）。〈亦真亦假的假新聞〉，《科學人》，195：28。

林麗雪、葉織茵譯（2018）。《後真相時代：當真相被操弄、利用，我們該如何看？如何聽？如何思考？》。臺北市：三采文化。（原書 Macdonald, H. [2018]. *Truth: How the Many Sides to Every Story Shape Our Reality*. London, UK: Bantam Press.）

徐佳士（1974）。〈我國報紙新聞「主觀錯誤」研究〉，《新聞學研究》，13：3-36。

乾隆來（2018 年 11 月 7 日）。〈不是臺灣獨有－大選假新聞成全球流行病〉，《今周刊》。取自 https://www.businesstoday.com.tw/article/category/80398/post/201811070021/不是臺灣獨有%20%20 大選假新聞成全球流行病

梁瑞祥（2001）。《網際網路與傳播理論》。臺北市：揚智。

黃俊儒（2018 年 10 月 1 日）。〈繪製一張「打假」地圖：假新聞的類型與攻略〉，《聯合報鳴人堂》。取自 https://opinion.udn.com/opinion/story/6077/3396950

鄭瑞城（1983）。《報紙新聞報導之正確性研究》。（國科會專題研究報告，NSC 71-0301-H004-001）。臺北市：國立政治大學新聞研究所。

鄭麗珍、江彥孟、游政憲（2019）。〈應用深度學習技術於網路虛假評論偵測〉，《電子商務學報》，21(2)：229-252

賴燕芳（2018 年 10 月 5 日）。〈制止假新聞傳播，谷歌、FACEBOOK 如何使用好 AI 這把劍？〉，《新浪財經頭條》。取自 https://t.cj.sina.com.cn/articles/view/2540408364/976b8e2c02000t7qh

羅文輝、蘇蘅、林元輝（1998）。〈如何提升新聞的正確性：一種新查證方法的實驗設計〉，《新聞學研究》，56：269-296。

蘇威全、方璇（2019 年 3 月 5 日）。〈探討大阪處長蘇啟誠之死 NHK：假新聞危害社會〉，《蘋果日報》。取自 https://tw.appledaily.com/new/realtime/20190305/1527380/

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 Election. *Journal of Economic Perspectives*, *31*(2), 211-236.

Bode, L., & Vraga, E. K. (2015). In related news, that was wrong: The correction of misinformation through related stories functionality in social media. *Journal of Communication*, *65*(4), 619-638.

Haldevang, M. D. (2017, October 17). Russia's troll factory also paid for 100 activists in the US. *Quartz*. Retrieved from https://qz.com/1104195/russian-political-hacking-the-internet-research-agency-troll-farm-by-the-numbers/

Hartley, J. (2010). *The uses of digital literacy*. Piscataway, NJ: Transaction Publishers.

Hermida, A. (2010). Twittering the news: The emergence of ambient journalism. *Journalism Practice*, *4*(3), 297-308.

Horton, C. (2018, November 22). Specter of meddling by Beijing looms over Taiwan's elections. *The New York times*. Retrieved from https://www.nytimes.com/2018/11/22/world/asia/taiwan-elections-meddling.html

Jin, Z., Cao, J., Guo, H., Zhang, Y., Wang, Y., & Luo, J. (2017). Detection and analysis of 2016 US presidential election related rumors on Twitter. In L. Dongwon, L. Yu-Ru, O. Nathaniel, & T. Robert (Eds.), *Lecture Systems* (pp. 14-24). Berlin, DE: Springer-Verlag.

Katsaros, D., Stavropoulos, G., & Papakostas, D. (2019, October). Which machine learning paradigm for fake news detection? In *IEEE/WIC/ACM International Conference on Web Intelligence (WI '19)*(pp, 383-387). New York, NY: Association for Computing Machinery.

Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019, May) MVAE: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference (WWW '19)*(pp. 2915-2921). New York, NY: Association for Computing Machinery.

McQuail, D. (2010). *McQuail's mass communication theory.* London, UK: Sage.

Newman, N., Dutton, W. H., & Blank, G. (2012). Social media in the changing ecology of news: The fourth and fifth estates in Britain. *International Journal of Internet Science*, *7*(1), 6-22.

Okoro, E. M., Abara, B. A., Umagba, A. O., Ajonye, A. A., & Isa, Z. S. (2018). A hybrid approach to fake news detection on social media

[Monograph]. *Nigerian Journal of Technology*, *37*(2), 454-462.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news. In *Proceedings of the 27th International Conference on Computational Linguistics*, (pp. 3391-3401). Santa Fe, NM: Association for Computational Linguistics

Potthast, M., Kiesel, J., Reinartz, K., Bevendorff, J., & Stein, B. (2018). A stylometric inquiry into hyperpartisan and fake news. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, (pp. 231-240). Melbourne, AU: Association for Computational Linguistics

Shu, K., Zhou, X., Wang, S., Zafarani, R., & Liu, H. (2019, October). The role of user profiles for fake news detection. In *IEEE/WIC/ACM International Conference on Web Intelligence (WI '19)*(pp, 436-439). New York, NY: Association for Computing Machinery.

Tandoc, E. C., Lim, Z. W., & Ling, R. (2018). Defining "fake news". *Digital Journalism*, *6*(2), 137-153

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146-1151 doi: 10.1126/science.app9559

Wardle, C. (2017, February 16). Fake news. It's complicated. *FIRST DRAFT*. Retrieved from https://firstdraftnews.org/fake-news-complicated/.

Wynne, H. E., & Wint, Z. Z. (2019, December). Content based fake news detection using N-Gram models. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services (iiWAS2019)*(pp. 669-673). New York, NY: Association for Computing Machinery.

Zannettou, S., Sirivianos, M., Blackburn, J., & Kourtellis, N. (2019). The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality* (*JDIQ*), *11*(3), 1-37.

Empirical Research on Fake News Detection Using AI Technology

Wei-Ching Wang, Chih-Shan Liao,
Hsu-Cheng Chiang & Kun-Chang Chou*

Abstract

Digital media technology is progressing rapidly, providing convenient communication tools for mankind. However, it is also exposing people to volumes of fake news and messages. In response to the flood of fake news, the authors of this study designed a news credibility system model that combines artificial intelligence (AI) technology, natural language processing, and deep learning to perform multidimensional and multiangle data analyses. Our system reached 90.53% in the confidence index for identifying unverified news. However, the decision interpretability of AI, human autonomy, and social overdependence are all points to consider when using AI to detect fake news. AI can be used as an effective auxiliary tool by Internet users to verify news credibility and can hopefully prompt vigilance against fake news.

Keywords: artificial intelligence, natural language processing, social media, fake news

* Wei-Ching Wang is Professor at the Graduate Institute of Mass Communication, National Taiwan Normal University, Taipei, Taiwan.
Chih-Shan Liao is Postdoctoral Fellow at the Graduate Institute of Mass Communication, National Taiwan Normal University, Taipei, Taiwan.
Hsu-Cheng Chiang is Associate Professor at the Graduate Institute of Mass Communication, National Taiwan Normal University, Taipei, Taiwan.
Ken-Chang Chou is Full-time Assistant at the Graduate Institute of Mass Communication, National Taiwan Normal University, Taipei, Taiwan.