許文錦、李牧衡、呂明聲 (2022),「運用 BERT 深度學習模型於衛教謠言檢測之研究」,*資訊管理學報*,第二十九卷,第一期,頁 27-44。

運用 BERT 深度學習模型於衛教謠言檢測之研究

許文錦*
國立中央大學資訊管理學系

李牧衡
國立中央大學資訊管理學系

呂明聲
國立中央大學資訊管理學系

摘要

近年來網路上流傳大量衛生教育謠言已為醫療照護人員帶來嚴重困擾,不實謠言除了誤導病患、造成恐慌外,更可能造成錯誤用藥、延誤治療等嚴重後果,更甚者,還可能造成醫病關係惡化,傷害醫師與醫院形象, 因此發展新的闢謠與防範機制成為急需研究的重要議題。關於謠言防範,醫療機構雖提供正確衛教宣導, 但侷於人力與資源不足, 成效有限, 而人工智慧技術快速發展為此問題帶來新的解決方向。本研究運用深度學習技術 BERT (Bidirectional Encoder Representations from Transformers)發展新一代衛教謠言檢測系統。實作上,首先收集包括衛福部等 15 個政府與民間單位衛教謠言資料,經資料前處理後,運用 BERT 結合 Bidirectional Long Short-Term Memory (BiLSTM) 進行模型訓練與驗證,最後將此 BERT-BiLSTM 闢謠模型佈署於 Line 聊天機器人上供使用者試用。實驗結果顯示本研究提出之 BERT-BiLSTM 模式可準確辨識出 90%的衛教謠言,本研究成果將有助於緩解醫護人員衛教解說時間不足、與滿足民眾查證衛教資訊之需求。

關鍵詞:BERT、深度學習、衛生教育、謠言檢測

* 本文通訊作者。電子郵件信箱:hsuwc@mgt.ncu.edu.tw
2021/07/05 投稿;2021/08/28 修訂;2021/09/28 接受

Hsu, W.C., Lee, M.H. & Lyu, M.S. (2022). A study on using BERT to identify health care rumors. *Journal of Information Management*, 29(1), 27-44.

A Study on Using BERT to Identify Health Care Rumors

Wen-Chin, Hsu*
Department of Information Management, National Central University

Mu-Heng, Lee
Department of Information Management, National Central University

Ming-Sheng Lyu
Department of Information Management, National Central University

Abstract

The circulation of numerous health care rumors on the internet in recent years has become extremely troublesome for medical and healthcare personnel. False rumors have misled patients, created panic, and even resulted in medication misuse and delayed treatment. They may also cause deteriorating doctor-patient relationships and tarnish the images of doctors and hospitals. Developing new mechanisms to dispel and prevent rumors has thus become an imperative issue in research. Although medical institutions promote accurate health education, limited labor and resources limit their effectiveness. The rapid progress in artificial intelligence technology is offering a new direction to solve this problem. This study employed BERT (Bidirectional Encoder Representations from Transformers), a deep learning technique, to develop the new generation alert system for health care rumors. First of all, we collected the materials of health care rumors from 15 private organizations and government agencies, including the Ministry of Health and Welfare. After data preprocessing was done, we used BERT combined with Bidirectional Long Short-Term Memory (BiLSTM) to conduct a model training and validation. Lastly, we deployed the BERT-BiLSTM model in Line chatbot so some LINE users can try it. The experiment result showed that the proposed model can identify as high as 90% of the health care rumors in social media. The research results will help to alleviate the burden relating to healthcare promotion by health professionals and to meet the public's needs to search for the correct health information.

Keywords: BERT, deep learning technique, health education, rumor detection

* Corresponding author. Email: hsuwc@mgt.ncu.edu.tw
2021/07/05 received; 2021/08/28 revised; 2021/09/28 accepted

壹、前言

衛生教育(Health education)(簡稱衛教)為醫療照護工作重要環節。以醫院為例，醫院衛生教育的目的在於引導病患、病患家屬、醫護人員朝向更健康的生活型態及提昇生活品質的所有計畫之設計、執行及評價(Sharma 2016)。其中，「衛教資訊的正確性」對病患康復與生活品質又特別具有關鍵性的影響，例如不當飲食習慣與藥物服用方式會明顯降低治療效果(Bastable 2016)。然而隨著網路社群媒體快速成長，越來越多的網路謠言(online rumors)已明顯增加衛生教育宣導的困難性，也造成醫護人員困擾與工作量增加(Wang et al. 2019)。錯誤衛教訊息可能引起個人面與社會面的嚴重問題。個人面上，錯誤訊息可能導致病患恐慌甚至拒絕服用正確藥物或自行停藥甚至改服錯誤偏方，導致個人健康上的損失與經濟、時間成本的支出。而在社會面上，除卻整體社會的相關成本增加外，醫護人員需要花費更多時間來溝通與安撫，過程中也可能惡化醫病關係(Sjöström et al. 2019) 。再與目前台灣的醫療機構普遍面臨人力不足、工時過長的情況相結合所造成的服務品質下降、壓縮看診時間等問題很可能進一步整體醫療環境的惡化(陳文信 2010; 曾芬郁、郭瑞祥 2009)。

病患與民眾無法在就醫時得到足夠正確衛教資訊時，多數人會上網搜尋資訊。然而網路資訊經常較為片段且來源不清，無論是病患或醫護人員，普遍的憂慮都是無法確保網路資訊的正確性(Sjöström et al. 2019)，因此「查證網路衛教資訊的真偽」是當前許多病患的主要需求。有鑑於民眾的需求，政府與民間單位已開始提供查證網路資訊的服務，例如食品藥物管理署食藥闢謠專區，雖然這些服務可滿足病患部分需求，然而目前這些查證服務多屬於被動式查詢，使用者需逐一查詢較為費時，且單一單位所擁有的衛教資訊有其局限。為此，許多學者建議可嘗試以人工智慧系統來自動檢測以及過濾不恰當的網路資訊 ，以達到節省人力，傳播正確衛教的目的(Li 2019; Wang et al. 2019)。綜合上述，為「緩解醫護人員時間不足」與「滿足病患查證衛教資訊的需求」，本研究運用 BERT (Bidirectional Encoder Representations from Transformers)深度學習技術發展智慧衛教謠言檢測系統，用於輔助醫護人員進行衛生教育，減輕醫護人員工作負擔。此系統目的為：(1)協助醫護人員在診療時給予適當衛教;(2)協助民眾進行衛教謠言檢測，取得正確衛教資訊。本論文第二節將探討衛教網路謠言的定義、防範方式、人工智慧偵測謠言等面向的相關研究，第三節說明研究架構與流程。第四節描述模型訓練、驗證與效能評估之研究成果，最後於第五節提出研究結論與限制。

貳、文獻探討

一、衛教謠言

「謠言(rumor)」在學術文獻的定義方式相當多(DiFonzo & Bordia 2000; Goh et al. 2017; Oh & Lee 2019; Paek & Hove 2019)，本研究整合過往研究結論並將「網

路衛教謠言」定義為「透過網路廣為傳播、未經官方證實的醫療照護(Health care)相關訊息，是一個可能為真實，亦可能是虛假的陳述」。而「官方」是指涉及謠言的當事人或具有公信力的政府單位(例:食藥署、衛生署)、專業人士(例:醫師、藥師)或第三方單位(例:世界衛生組織 WHO)。根據產生意圖，謠言可分為二種類型：(1)造謠(disinformation):故意發布假消息以達到某特定目的，例如誇大藥品療效的商業謠言;(2)誤傳(misinformation):人們因為錯誤認知或未經深入思考或查證，不經意就把訊息傳播出去，例如醫師口碑或醫療手術謠言等(Wang et al. 2019)。衛教謠言透過網路的匿名性、易傳遞性以及無須審查等三個特性快速且廣泛地影響著社會(Huilai, Fang, & Junjie 2016; Liang & Yang 2015)。因此，制訂制約與防護的機制是必須且刻不容緩的。

商業謠言(business rumor)(包括醫藥相關訊息)是最常被傳播的謠言類型之一(Vosoughi, Roy, & Aral . 2018)，而台灣地區最常被傳播的謠言類型則包括衛生與健康(33%)、消費安全(26%)與八卦(Gossip, 16%)等與一般大眾生活密切相關的訊息(汪志堅、駱少康 2002)。若訊息接收者誤信謠言，則可能因錯誤決策(例：延誤就醫)而導致嚴重後果(例：錯失治療良機)(Oh & Lee 2019)。過往研究顯示病患對病症的理解程度與其治療意願呈現顯著正向相關，當病患接收到不實衛教資訊或對於症狀有誤解時，作出不利於其健康的決策比例明顯上升(Matthews et al. 2002)。一份針對線上癌症討論社群用戶進行的問卷調查顯示，超過 60%的受訪者相信未經證實的癌症資訊，尤其若資訊來自於親朋好友時，多數受訪者願意相信這些資訊，即使這些資訊並無醫學根據(DiFonzo et al. 2012)。有鑑於此，有許多學者呼籲政府與相關單位應建立一套闢謠與防護的制度與方法來保護民眾健康以避免重大傷害(Liang & Yang 2015; Vosoughi et al. 2018)。

二、闢謠與防護方法

關於衛教謠言的闢謠與防護方法的研究，目前主要可分為四類，包括：

(一) 由謠言產生原因下手，試圖減少謠言產生。過去研究發現「焦慮感、不確定性、來源可信度、對事件結果的涉入程度、趣味性」是網路謠言產生的主要因素 (Bordia & DiFonzo 2017; Li & Chong 2019)。散播類似「長期使用鼻腔噴霧劑會導致嗅覺喪失」之訊息可能導致接收者心理認知不平衡進而產生焦慮感。要降低謠言造成的心理影響，可能必須於醫療過程中加強宣導正確衛教資訊。

(二) 由謠言散播過程下手，試圖阻斷謠言散播。Pendleton (1998)指出謠言散播一般會經歷孕育期(parturition)、散播期(diffusion)與控制期(control)三個階段。孕育期指謠言產生，通常起因於人們對關心的人或事有焦慮感，且真相難以驗證時產生，可能是惡意製造或自然形成；在散播期，謠言會因自身特性與傳遞媒介的不同而有程度不一的擴散效果；第三階段為謠言的控制，謠言在流傳一段時間後會因其重要性等因素消失而自然消退，或真相已被驗證而逐漸平息。一般而

言，在孕育期阻斷謠言產生是較佳策略，尤其在網路時代，一但上網幾乎難以阻斷。

(三) 強化病患健康素養(health literacy)，由病患自行辨識資訊真偽。學者指出健康素養是「一種能力，使個體能獲取、解釋並理解健康資訊，並能運用資訊促進健康」(Nutbeam 2008)。網路衛教謠言雖非一定錯誤，但大多未經證實，且內容常混入似是而非的佐證，容易讓病患誤以為真(Lo & Chiu 2015)；由於廣為流傳的衛教謠言通常與病患關注的事件或流行病相關(例:高血壓、癌症)，容易使病患認定該資訊是重要的，從而將資訊再分享給其他病友，使得謠言不斷被傳播。若病患能具備一定健康素養，便能對謠言進行思辨，也能應用相關資訊品質工具評估真偽，以促進健康(DiFonzo et al. 2012)。

(四) 運用資訊科技技術自動辨識與闢謠。由於「難以在短時間內查證」與「謠言數量龐大」是網路謠言的重要特性，再加上目前世界各國(包括台灣)醫護人員工作量大，且人員短缺的現狀，想藉由傳統衛教宣導來阻絕謠言有實際執行上的困難, 於是許多學者嘗試以資訊技術進行自動謠言檢測，以求在短時間內達到防範謠言擴散的目的，例如深度學習模型 RNN(Recurrent Neural Network), LSTM(Long Short-Term Memory)與 GRU(Gated Recurrent Unit)(Li, Cai, & Chen 2018; Yu et al. 2017)。

綜上所述，由於網路與社群媒體的特性，每個使用者均能夠發文與轉傳，因此要從「阻斷謠言產生與傳播」著手在現實上是不可行的，而「強化病患健康素養(health literacy) 」與「自動辨識與闢謠」相對較為可行，也是目前許多學者建議的方向，而本研究即是「提供正確衛教資訊強化病患健康素養」與「建構智慧型闢謠系統」為研究方向。

三、自動化謠言檢測

為了建構智慧型闢謠系統並提供正確衛教知識，必須先了解自動化謠言檢測(Automatic detection and verification of rumors)所使用的相關技術與趨勢進行了解，方能提出實質上解決「滿足病患查證衛教資訊」以及「減輕醫護人員負擔」兩個本研究所關注的問題。「自動化謠言檢測技術」以機器學習(Machine Learning)檢測法為主流 (Bondielli & Marcelloni 2019)。此研究領域中，謠言檢測任務被視為一個二元分類問題，即判斷一則訊息是否為「真實」或「虛假」 (Cao et al. 2018)。

在 2016 年以前，相關技術主要研究方向是「引導資訊系統透過歷史謠言資料集學習，並歸納謠言的通用特徵或規則，最後自動檢測謠言真偽的技術」。其中，特徵擷取工程 (Feature engineering) 是由原始謠言資料中辨識出更好的資料特徵用以提升檢測模型正確率，方法包括資料前處理、特徵選取與降維等。Castillo, Mendoza, & Poblete (2011)使用 Decision Tree 分類器對社群媒體 Twitter 的謠言進行分類，其擷取並分析了 Twitter 多種文本與 68 種使用者特徵(例:訊息

包含之超連結數量、情感值分數、朋友數量、回覆數量等)，其分類正確率達 73.9%。其它類似研究還包括 Yang et al. (2012)、Kwon et al. (2013)與 Zhao, Resnick, & Mei (2015)等。

2016-2020 年時期，相關技術開始以深度學習(Deep Learning; DL)檢測法為主，DL 是引導電腦模擬人腦的運作方式(即類神經網路)來學習謠言資料，並運用所學檢測謠言真偽的技術。它的特色是在反覆訓練模型的過程中能夠自動修正已獲得的經驗值(即神經元權重值)，即同等於和人類一樣具有學習並總結歸納的能力 (Moein 2014) 。深度學習可大幅減少特徵工程所需時間，獲取更多隱藏卻有意義的謠言特徵，也能分析謠言上下文及內容的變化型式(Nguyen, Li, & Niederée 2017)。其中，較著名的研究是 Ma et al. (2016) 使用 RNN 於檢測社群媒體 Twitter 的謠言，結果其檢測正確率(約 88%)明顯優於早期的機器學習技術。因此近年來(2016-2019)自動化檢測任務之研究大多採用深度學習相關技術來進行，其結果也確認深度學習法的優越性(Singh, Rana, & Dwivedi 2019)。此時期其他的重要研究包括 Yu et al. (2017) 、Zhou et al. (2018) 、Li et al. (2018) 與 Asghar et al. (2019)等。

雖然謠言檢測技術已有長足進步，然而謠言本身也在不斷進化，例如近期許多謠言不但蓄意捏造，且將假資訊混入真實訊息裡；同時，造謠者也刻意在話題選擇、撰寫風格方面加以複雜化。研究認為，必須使用更先進的謠言檢測技術才能處理不斷進化的錯誤謠言(Asghar et al. 2019)。Asghar et al. (2019)指出 BERT (Bidirectional Encoder Representations from Transformers)技術的應用可以解決上述謠言日趨複雜的問題。該技術係 2018 年由 Google 公司為了使人工智能更深入理解人類語言所推出之技術。其著重在自然語言處理並由約 33 億個字的語言文本訓練而成。其二個重要核心概念為雙向 Transformer 與其所包含之 self-attention 機制。首先，BERT 的雙向 Transformer 模型能有效解決過去深度學習模型(例:RNN)無法平行處理資訊，處理效率較低的問題，而 self-attention 機制則負責處理輸入語句中所有單詞與單詞間的向量關係，進而實踐「可根據單詞上下文或整體語句之結構理解自然語言意涵」之目的(Vaswani et al. 2017)。過去 Word2Vec 或 Glove 等模型的作法是將自然語言轉為向量的詞嵌入(word embedding)，其缺點是不重視上下文關係及無法判別一詞多意，而 BERT 的雙向 Transformer 模型可改善上述缺點，研究者發現 BERT 在依據各詞彙上下文關係以及語法(syntax)、語意(semantics)的情況下，更能理解自然語言之結構與資訊(Jawahar, Sagot, & Seddah 2019)。實務上，Devlin et al. (2018)也發現，即使面對高度精細複雜的語句，BERT 的理解與檢測能力依然優異。因此，本研究將採用 BERT 來建構智慧衛教謠言檢測系統。有關於 BERT 技術的細節可參考 Devlin et al. (2018)、Vaswani et al. (2017) 以及 Wolf et al. (2019)。

參、研究方法

一、研究流程

本研究之流程分為三個階段如圖 1 所示，第一個階段為取得衛教相關謠言資料，範圍包含政府官方(如食藥署闢謠專區)與民間(如台灣事實查核中心)等資料庫，目的是建立作為訓練 BERT 的基礎。第二階段將運用 BERT 深度學習模型分析於第一階段蒐集之衛教謠言資料集，並訓練虛假謠言檢測模型，最後依模型計算結果驗證檢測效果。第三階段則是以第二階段之成果為基礎，針對醫護人員或一般民眾及病患提供使用者介面，使其能夠以簡單快速的方式查詢衛教謠言的真假。

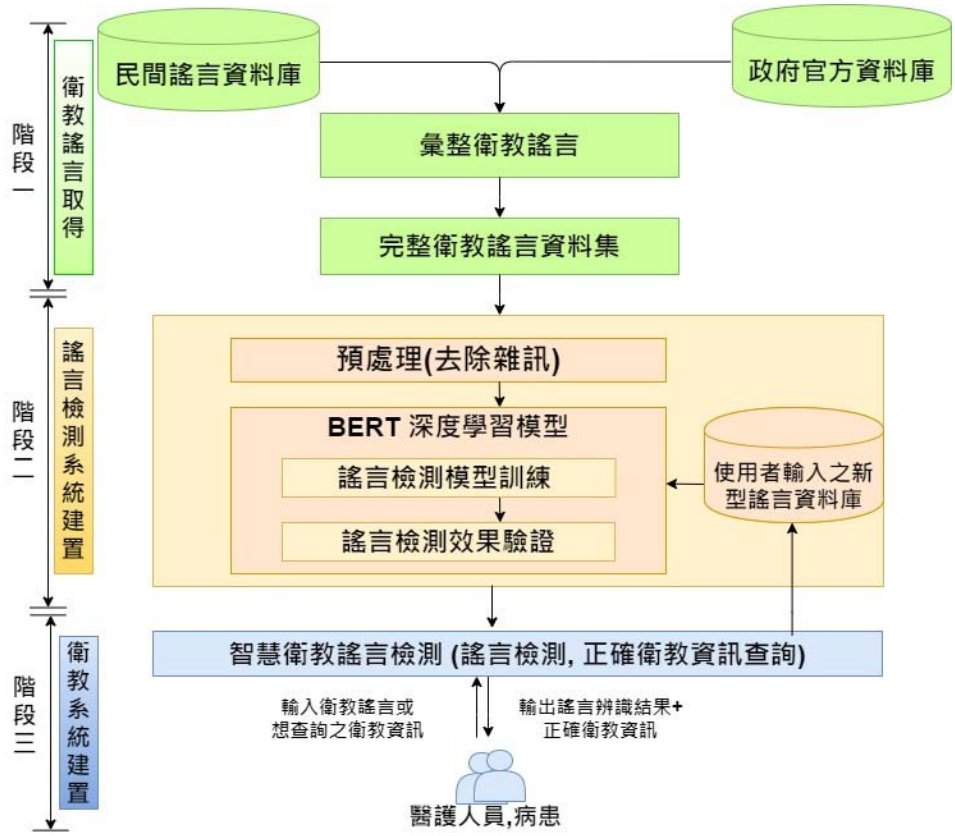

圖 1：研究流程

三個階段工作詳述如下：

(一) 階段一：取得衛教謠言資料

本研究由具公信力的政府官方與知名民間單位兩個管道蒐集衛教謠言。政府管道包含衛生福利部國民健康署(健康九九網站、保健闢謠專區)、食品藥物管理署（食藥闢謠專區、食藥好文網）、疾病管制署、行政院農業委員會農業試驗所以及行政院農業委員會漁業署等 7 個官方資訊平台。民間單位包括美的好朋友、MyGoPen、蘭姆酒吐司、不實消息部落格、食力 FoodNEXT、台灣事實查核中

心、丁香園醫療專業網站、康健等 8 個知名謠言資訊庫。上述資料來源具備公信力與相當的完整性，資料包括謠言與專家對謠言的看法。本研究針對上述來源資料進行人工篩選，篩選與衛教相關的流言並且去除「重複、不符合民情(例如資訊來源為中國大陸)與未提供謠言解答」的資料後，最後將資料集以 8：2 的方式隨機分為訓練用與驗證用資料作為後續使用。

(二) 階段二：謠言檢測模型訓練與驗證

此階段進行謠言檢測模型訓練與驗證，發展流程如圖 2 。本研究使用 Python 3.7 版搭配 TensorFlow 2.5 進行模型開發，此工具已廣泛被接受用於深度學習研究(Park et al. 2019; Sankar et al. 2019)，是一個可靠並具良好執行效率的深度學習模型開發工具。本研究謠言檢測模型的建立包括以下步驟：

1. 資料預處理：

由於網路謠言為吸引觀看與轉傳，內容設計相當口語化，經常包含表情符號、網址、標點符號等資料。為避免這些雜訊對下一階段的語義分析造成干擾，我們對謠言資料庫進行語料篩選、清理(Clean)、統一格式(Format)、缺失值(Missing value)處理與轉換(Transform)等工作，例如清除多餘標點符號、網址與表情等符號。此外，我們也過濾部分無意義且頻率高的詞彙以提升模型學習效率，例如「聽說、鄰居說、網路傳言」等詞，研究顯示資料預處理對深度學習模型的學習效率有明顯助益(Zaccone & Karim 2018)。

2. 謠言語義轉化、學習與分類訓練

此階段運用 BERT 轉換與學習謠言內容(Asghar et al. 2019) 。Google 提供 2 種不同規格的 BERT 模型：BERT-Base 與 BERT-Large。 其中 BERT-Base 有 12 個 Transformer 層， 768 個隱含單元，12 個自注意力層， 總共含有 1.1 億個參數，模型共包含約 2 萬的中文簡體字和繁體字。BERT-Large 有 24 個 Transformer 層，1024 個隱含單元， 16 個自注意力層，總共含有 3.4 億個參數 。囿於計算資源有限，一般學者多採用 BERT-Base， 本研究亦採用 BERT-Base 為研究訓練模型。圖 2 描繪了 BERT 模型的具體結構，模型輸入(Input)為謠言句子，輸出(Output)為分類標籤(Devlin et al. 2018)。另外，為了更好地在自然語言中處理上下文關係(Li et al. 2018)，本研究使用 Bidirectional LSTM(BiLSTM)作為 BERT 的下游分類器。將 BERT 最後一層的輸出作為輸入再進行訓練。

3. 謠言檢測效果驗證

此階段將驗證上一步驟訓練完成的謠言檢測模型分類效果。二元分類器常用的評價指標包括準確率(Accuracy)、精確率(Precision)、召回率(Recall) 與 F 值(F-score)。一般而言，為平衡 Precision 和

Recall之間的關係避免出現由於資料類別分佈不均衡導致兩個分數之間相差過大，無法充分反映分類器的效果，通常在分類器研究中會使用 F score (即上述兩者的調和平均值)作為分類的評價指標。因此，本研究以 Accuracy、Precision、Recall 以及 F score 為評價方式來驗證謠言檢測模型分類效果。

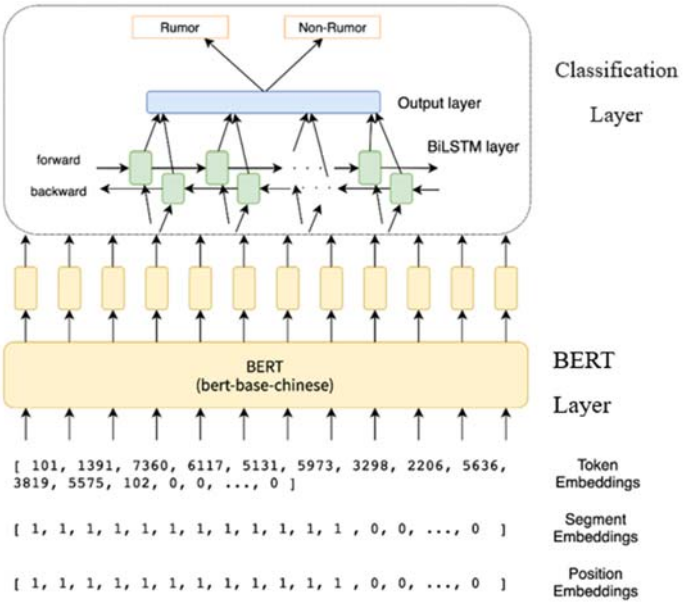

圖 2： BERT- BiLSTM 模型結構

(三) 階段三：衛教闢謠系統建置

本階段以訓練完成的 BERT- BiLSTM 模型為基礎建置提供醫護人員、病患使用之衛教闢謠系統。系統平台以台灣使用率最高的即時通訊軟體 Line 進行佈署，一方面減少使用者學習之困難，也減少系統開發之成本負擔。

肆、研究成果

一、模型建構環境

本研究使用 Google Colaboratory 線上機器學習服務(colab.research.google.com)來實現 BERT-BiLSTM 衛教謠言辨識模型，平台環境設定如表 1 所示，主要運用 Python 程式語言與 AI 發展系統 TensorFlow 數值計算函式庫為平台，搭配 Keras、Pandas 及 Numpy 深度學習函式庫來建構模型，硬體的部分，CPU 為 Intel(R) Xeon(R) 2.20GHz，GPU 為 Nvidia(R) Tesla T4，記憶體為 13G。

表 1：實驗環境(軟硬體)

Item	Model/ Specification/ Software version
CPU	Intel(R) Xeon(R) CPU @ 2.20GHz

GPU	Nvidia(R) Tesla T4
RAM	13 GB
Programming language	Python 3.7.10
Libraries	Tensor Flow 2.5.0、Keras 2.4.3、Numpy 1.19.5、Pandas 1.1.5、Openjdk-1.8、cuda 10.1、cudnn 7603、scikit-learn 0.22.2、torch 1.8.1+cu101、
IDE	Google colab

二、謠言資料庫

本研究於 2020 年 11 月收集來自 15 個具公信力資料庫的衛教謠言資料(如圖 3)，包括政府平台 1350 筆與民間平台 1017 筆(共計 2367 筆)，排除重複、不合宜以及未給予完整說明的資料後，最後獲得 1746 筆完整謠言與解答，包含食品、生活、性教育、醫療活動、化妝品、藥品、疾病、運動等八個類別，各分類之資料筆數與定義請參考表 2。

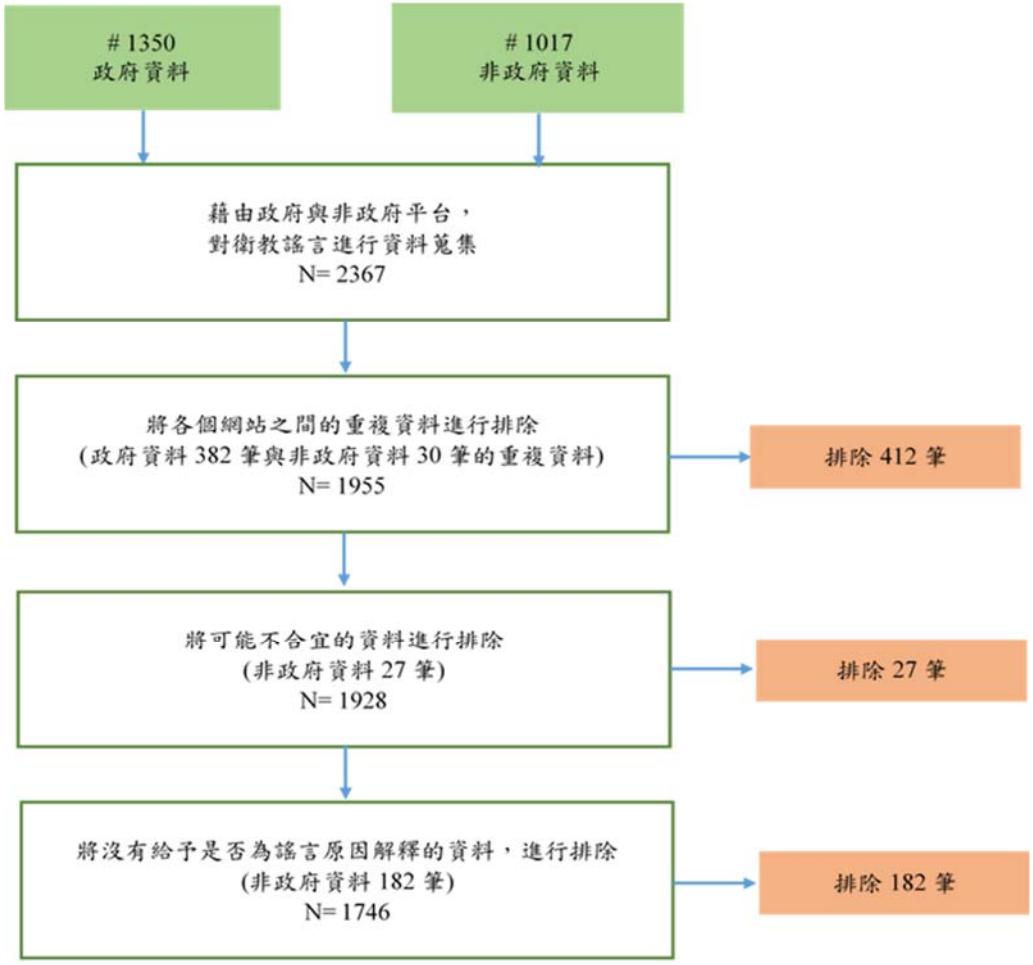

圖 3：資料蒐集與篩選

表 2：衛教謠言類型與數量

類型	定義	資料筆數
食品	食物相關謠言，例如：聽說多吃楊桃，可治療新型冠狀病毒，這是真的嗎？	786
生活	生活日常活動相關謠言，例如：菸癮、眼科、牙科、傷口處理、政治、安全、家電用品。	329
性教育	性相關謠言，例如：婦科、性知識、孕期等。	62
醫療活動	醫療活動相關謠言，包括疫苗、口罩、手術、健康檢查等，例如：乳房自我檢查可以降低乳癌死亡率嗎？	96
化粧品	身體外觀、皮膚等相關謠言。	49
藥品	藥品相關謠言，例如：避孕藥、抗組織胺等需要醫師處方簽之藥品、搖頭丸、K 他命等。	138
疾病	疾病相關謠言，例如：癌症、感冒等。	254
運動	運動相關謠言，例如：瑜珈、跑步等。	32
資料總筆數		1746

三、模型訓練與驗證

完成衛教謠言資料集預處理後，本研究將 1746 筆資訊依據 8：2 的比例隨機分為訓練用(1396 筆)與驗證用資料集(350 筆)，並將訓練資料導入本研究所提出之 BERT- BiLSTM 分析架構中，並使用驗證集進行謠言檢測效果驗證。具體學習步驟為：(a)首先將謠言輸入 BERT 模型(圖 3 標示 1)，輸入前先將高維度的謠言語義(例: 聽說感冒藥加咖啡效果更強)轉換成低維度的詞向量(word embedding, 例:000111…)，其中包含 position embedding(位置向量)、segment embedding(句子向量)、Token embedding (詞向量); (b)接下來每個 Transformer 利用 Multi-Head Attention 的機制建立詞與詞之間的聯結關係 (即權重值)(Vaswani et al., 2017)，並開始進行二項語義訓練(圖 3 標示 2)：(b1) Mask Language Model：運用類似克漏字的方式，隨機「遮蓋(Mask)」句子中的某些詞，讓 BERT 編碼器依照上下文來學習預測這些詞的正確語義，例如「聽說感冒藥加__[Mask]__效果更強」，BERT 會根據上下文的關係來預測被遮蓋的詞(Mask)；(b2) Next Sentence Prediction：上下句關係預測，BERT 會藉由預測兩個隨機句子能否組成上下句來學習句子間的關係。BERT 就在反覆練習的過程中學習謠言特徵，並了解正確與錯誤謠言的區別。研究顯示訓練後的 BERT 模型具有很強的句詞理解能力，無論是在字詞級別的自然語言分類任務(如命名識別)，還是問答類(QA)的句子級別的任務中都有卓越表現(Devlin et al. 2018)。

為了更好地在自然語言中處理上下文關係，本研究使用 Bidirectional LSTM (BiLSTM)作為 BERT 的下游分類器，即將 BERT 最後一層的輸出作為輸入再進

行訓練。表 3 顯示 BERT-BiLSTM 模型參數(各參數相關說明可參考 https://huggingface.co/transformers/model_doc/bert.html)，實驗採用不同參數進行，表 4 呈現實驗訓練與測試結果，因模式訓練過程成本函數(cost function)的誤差收斂狀況佳，模式準確度大多數達 90%以上，不同參數組合影響模式準確度不大(介於 88%-97%之間)，最後選定 hidden layers 數目設定為 2，Unit size 設定 512，Epochs 設定為 25。整體而言，BERT- BiLSTM 模型無論是 Accuracy (%)、Precision、Recall 或 F-score 均達到 90%以上的水準，顯示其良好的謠言檢測能力。

表 3：BERT-BiLSTM 模型參數

Parameter names	Setting
BERT vocabulary size (bert-base-chinese)	21128
BERT output dimension	768
BERT hidden layers dropout	0.1
BERT hidden layers	12
LSTM unit size	32, 64, 128, 256, 512
BiLSTM unit size	32, 64, 128, 256, 512
Number of hidden layers	1, 2, 3, 4, 5
Number of epochs	20, 25
Learning rate	1e-5
Batch size	16

表 4：BERT-BiLSTM 模型之驗證結果

Unit size	Epo-chs	Accuracy (%)				Precision				Recall				F-score			
		Number of hidden layers															
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
32	25	96.50	96.77	96.50	94.07	0.97	0.97	0.97	0.95	0.98	0.98	0.98	0.97	0.98	0.98	0.98	0.96
64	25	95.42	96.50	95.96	90.57	0.96	0.97	0.97	0.92	0.97	0.98	0.98	0.97	0.97	0.98	0.97	0.94
128	25	95.42	95.15	96.50	94.07	0.97	0.95	0.97	0.94	0.97	0.98	0.98	0.98	0.97	0.97	0.98	0.96
256	25	96.23	94.34	96.23	93.80	0.97	0.97	0.97	0.95	0.98	0.96	0.98	0.97	0.97	0.96	0.97	0.96
512	25	96.23	97.04	95.96	88.41	0.97	0.97	0.96	0.88	0.98	0.99	0.99	0.99	0.97	0.98	0.97	0.93

資料集：350 筆衛教謠言(來源：15 個台灣具公信力之政府與民間資料庫)

四、模型效能評估

為了客觀評估本研究提出之 BERT-BiLSTM 模型的效能，表 5 比較 BERT-BiLSTM 模型與最近 5 年常見的自動化謠言檢測模型於 Pheme 謠言資料集上的檢測結果。Pheme 是學術文獻中常用的知名謠言資料集(Asghar et al. 2019)，其內容為社交平台 Twitter 流傳的謠言和非謠言的集合，包含 9 個重要事件相關的謠言，並且每個謠言都以其真實性值(即 True，False 或 Unverified)標註。由表 4 可

得知，最近 5 年共有 5 位學者以深度學習技術進行 Pheme 謠言自動檢測，例如 Ma et al. (2016)提出 RNN 檢測法(F-score=0.8)、Ajao, Bhowmik, & Zargari. (2018)提出 LSTM-CNN 方法(F-score=0.83)、Asghar et al. (2019)提出 BiLSTM-CNN 法(F-score= 0.86)。相較於近 5 年內常見檢測法，本研究的 BERT-BiLSTM 模型最大不同之處，在於透過將 BERT 得到謠言上下文關係的詞嵌入(Contextualized Word Embedding)，而非一般的詞嵌入(Word Embedding)，它讓詞嵌入具有上下文關係，再透過擅長處理序列資料且能理解上下文關係的 BiLSTM 層做微調(Fine-tune)。實驗結果證實 BERT-BiLSTM 模型於 Pheme 謠言檢測任務中，其表現優於過去學者所提之檢測法，F-score 達 0.89、Precision 0.85、Recall 0.92，具有良好的謠言檢測效能。換言之，本研究所使用的 BERT- BiLSTM 架構具有良好檢測能力，無論在自行收集的台灣衛教謠言資料集或公用的 Pheme 資料集均有良好表現，因此能夠有效協助民眾進行衛教謠言檢測與幫助醫護人員快速給予病患適當衛教資訊。

表 5： BERT-BiLSTM 模型與其它謠言辨識深度學習模型之比較

Study	Deep Learning Model	Precision	Recall	F-score
Ma et al. (2016)	RNN	0.81	0.81	0.80
Yu et al. (2017)	1-Layer CNN	0.80	0.80	0.78
Ajao et al. (2018)	LSTM-CNN	0.83	0.84	0.83
Ma, Gao, & Wong. (2019)	GAN-GRU	0.78	0.78	0.78
Asghar et al. (2019)	BiLSTM-CNN	0.86	0.86	0.86
BERT-BiLSTM		**0.85**	**0.92**	**0.89**

Dataset：PHEME rumor dataset (330 conversational threads) (Zubiaga et al., 2016)

五、衛教闢謠系統

為了讓上述提出之 BERT- BiLSTM 模型能夠真正幫助病患闢謠，本研究將模型佈署於台灣民眾普遍使用的 Line 即時通訊軟體。如圖 4 所示，使用者可直接於 Line 輸入資訊，此資料經過 BERT- BiLSTM 模型處理後，由系統進行比對後於 output layer (參考圖 2)使用 Sigmoid 函數將結果映射為機率，其原理可理解為將 BERT- BiLSTM 模型所輸出之所有單詞、語意、前後文之分析結果轉換為介於 0 到 1 之間的邏輯函數(logistic function)，該值可視為判定某一條件是否為真的機率值，一般以 0.5 作為閾值(cut off value)，在該點以下將認定為 0，以上則認定為 1(以本研究為例，會將結果分為「虛假謠言(1)」與「真實資訊(0)」) (Jordan 1995; Nwankpa et al. 2018)，並將該結果回傳予使用者。然而，即使依據 BERT-BiLSTM 之分類結果以及 sigmoid 函數所轉換之機率值在一定程度上已經提供使用者判斷的資訊，但我們認為有必要將 sigmoid 函數所計算之結果直接提供使用者作為參考，使其了解系統分析結果並非斷言該輸入資訊為真或假，而是在一定

的機率下認定其真假。以圖 4 為例，使用者輸入「聽說吸二手煙對身體沒影響」後，系統分析的結果為該資訊為虛假謠言的機率為 79.86%，即是依據分析結果以及 sigmoid 函數轉換後認定之機率值所得。因此，本闢謠系統同時提供該結果相關的專家解釋，希望給予使用者更多相關衛教資訊作為實際判讀之參考，並達到教育民眾之目的。

圖 4： 闢謠機器人示意圖

伍、結論

台灣與各先進國家目前均面臨照護人力不足之問題，尤其新冠肺炎的流行已為公衛體系人力調度帶來極大影響，本研究提出以 BERT-BiLSTM 為架構之衛教謠言檢測系統具有約 90%準確率的良好謠言辨識能力，將有助於緩解醫護人員時間不足與不實謠言對病患的影響。整體而言，本研究具有學術面、衛教面以及實務面上的貢獻，分述如下

學術貢獻：自動化謠言檢測為目前熱門研究議題。然由於 Google BERT 為近 2 年之新技術，過往研究較少提出以 BERT 為基礎之檢測法，亦無針對衛教謠言進行自動檢測之研究，本研究成果可以彌補此知識缺口，增加自動化謠言學術研究之廣度，特別是衛生教育領域。

衛教貢獻：疾病治療僅是治標，健康促進(health promotion)方為治本之道，衛生教育是推動健康促進的重要方式，學者提出三級預防的概念(初級:預防疾病、次級:早期發現早期治療、三級:降低後遺症)(Sharma, 2016)。本研究建置之衛教謠言檢測系統可協助病患隨時評估接收到的衛教資訊，即早察覺與治療，也可以提供已患病者專業正確的衛教資訊防止病情惡化，對於目前衛生工作有實質幫助。

實務貢獻：目前主流的闢謠機器人多以使用者輸入之關鍵字作為查詢，並將闢謠結果提供給使用者參考或由真人闢謠。關鍵字型闢謠機器人，易發生該關鍵字控制的謠言沒有被事先建立、文意相同但文本不同的謠言無法被辨識等問題，而真人闢謠雖能給予專業回覆，但需花費較長時間。本研究提出之闢謠模型機器人可同時解決上述問題，由使用者輸入謠言後，即可馬上給予分類結果以及最相關的已知謠言並附有專家闢謠結果。

最後，本研究之構思、設計與模型建構雖已力求嚴謹與完整，然囿於時間與環境，仍存在以下限制：其一，衛教謠言資料收集不易，本研究僅收集到 2367 筆資料，最終使用 1746 筆完整具高品質之資料進行建模，然由於資料筆數較少，可能影響模型檢測效能。第二，Google BERT 最多可支援的文字長度為 512 個詞(token)，對於長度超出 512 個詞的序列資料則無法處理，侷限於處理較短序列資料，此為研究限制。最後的限制在於本研究提出之衛教謠言檢測模型僅實驗於繁體中文環境，不同語言差異可能會對最終檢測結果造成影響。

誌謝

本文感謝科技部提供計畫經費補助(MOST 109-2410-H-008-006 -)

參考文獻

汪志堅、駱少康 (2002)，「以內容分析法探討網路謠言之研究」，*資訊、科技與社會學報*，第一期，頁 131-149。

陳文信 (2010)。「醫師看診平均 1.4 分鐘 監院糾正」。*中國時報*，http://healthmedia.nownews.com/contents.aspx?cid=5,69&id=10404

曾芬郁、郭瑞祥 (2009)，「門診時段及醫師特質對就診流程品質之影響」，*台灣醫學*，第十三卷，第六期，頁 558-566。

Ajao, O., Bhowmik, D., & Zargari, S. (2018). Fake news identification on twitter with hybrid cnn and rnn models. *Proceedings of the 9th International Conference on Social Media and Society,* 226-230

Asghar, M. Z., Habib, A., Habib, A., Khan, A., Ali, R., & Khattak, A. (2019). Exploring deep neural networks for rumor detection. *Journal of Ambient Intelligence and Humanized Computing*, 12, 4315-4333.

Bastable, S. B. (2016). *Essentials of patient education*. Jones & Bartlett Publishers, MA.

Bondielli, A. & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, 497, 38-55.

Bordia, P. & DiFonzo, N. (2017). Psychological motivations in rumor spread. in Fine, G.A., Heath, C. & Campion-Vincent, V. (Eds.), *Rumor Mills: The Social Impact of Rumor and Legend*, Aldine Press, 87-101.

Cao, J., Guo, J., Li, X., Jin, Z., Guo, H., & Li, J. (2018). Automatic rumor detection on microblogs: A survey. *arXiv preprint arXiv:1807.03505*.

Castillo, C., Mendoza, M., & Poblete, B. (2011). Information credibility on twitter. *Proceedings of the 20th international conference on World wide web*, 675-684.

Devlin, J., Chang, M.W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DiFonzo, N. & Bordia, P. (2000). How top PR professionals handle hearsay: Corporate rumors, their effects, and strategies to manage them. *Public Relations Review*, 26(2), 173-190.

DiFonzo, N., Robinson, N. M., Suls, J. M., & Rini, C. (2012). Rumors about cancer: Content, sources, coping, transmission, and belief. *Journal of health communication*, 17(9), 1099-1115.

Goh, D. H.L., Chua, A. Y., Shi, H., Wei, W., Wang, H., & Lim, E. P. (2017). An analysis of rumor and counter-rumor messages in social media. *Proceedings of 19th International Conference on Asia-Pacifc Digital Libraries,* 256-266.

Huilai, Z., Fang, L., & Junjie, Z. (2016). What makes people resend healthy food messages online: the effects of message cues. *Proceedings of 2016 13th International Conference on Service Systems and Service Management (ICSSSM)*, 1-6.

Jawahar, G., Sagot, B., & Seddah, D. (2019). What does BERT learn about the structure of language? *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*.

Jordan, M. I. (1995). *Why the logistic function? A tutorial discussion on probabilities and neural networks.* (9503). MIT Computational Cognitive Science Report, ftp://che.mit.edu/pub/jordan/uai.ps

Kwon, S., Cha, M., Jung, K., Chen, W., & Wang, Y. (2013). Prominent features of rumor propagation in online social media. *Proceeding of 2013 IEEE 13th International Conference on Data Mining*, 1103-1108.

Li, B. & Chong, A. (2019). What Influences the Dissemination of Online Rumor Messages: Message Features and Topic-congruence. *Proceedings of the 40th International Conference on information systems*.

Li, J. (2019). Detecting False Information in Medical and Healthcare Domains: A Text Mining Approach. *Proceeding of International Conference on Smart Health*, 236-246.

Li, L., Cai, G., & Chen, N. (2018). A Rumor Events Detection Method Based on Deep Bidirectional GRU Neural Network. *Proceeding of 2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*, 755-759.

Liang, J. & Yang, M. (2015). On spreading and controlling of online rumors in we-media era. *Asian Culture and History*, 7(2), 42-46.

Lo, W.L. & Chiu, M.H.P. (2015). A Content Analysis of Internet Health Rumors. *Journal of Educational Media & Library Sciences*, 52(1), 1-24.

Ma, J., Gao, W., Mitra, P., Kwon, S., Jansen, B. J., Wong, K.F., & Cha, M. (2016). Detecting rumors from microblogs with recurrent neural networks. *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI 2016)*, 3818-3824.

Ma, J., Gao, W., & Wong, K.F. (2019). Detect Rumors on Twitter by Promoting Information Campaigns with Generative Adversarial Learning. *Proceedings of The World Wide Web Conference 2019*, 3049-3055.

Matthews, A. K., Sellergren, S. A., Manfredi, C., & Williams, M. (2002). Factors influencing medical information seeking among African American cancer patients. *Journal of health communication*, 7(3), 205-219.

Moein, S. (2014). *Medical diagnosis using artificial neural networks*. IGI Global.

Nguyen, T. N., Li, C., & Niederée, C. (2017). On early-stage debunking rumors on twitter: Leveraging the wisdom of weak learners. *Proceeding of International Conference on Social Informatics*,141-158.

Nutbeam, D. (2008). The evolving concept of health literacy. *Social Science & Medicine*, 67(12), 2072-2078.

Nwankpa, C., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation functions: Comparison of trends in practice and research for deep learning. *arXiv preprint arXiv:1811.03378*.

Oh, H. J. & Lee, H. (2019). When Do People Verify and Share Health Rumors on Social Media? The Effects of Message Importance, Health Anxiety, and Health Literacy. *Journal of health communication*, 24(11), 837-847.

Paek, H.-J. & Hove, T. (2019). Effective strategies for responding to rumors about risks: The case of radiation-contaminated food in South Korea. *Public Relations Review*, 45(3), 1-9.

Park, Y.-J., Bae, J. H., Shin, M. H., Hyun, S. H., Cho, Y. S., Choe, Y. S., Choi, J. Y., Lee, K.-H., Kim, B.-T., & Moon, S. H. (2019). Development of Predictive Models in Patients with Epiphora Using Lacrimal Scintigraphy and Machine Learning. *Nuclear medicine and molecular imaging*, 53(2), 125-135.

Pendleton, S. C. (1998). Rumor research revisited and expanded. *Language & Communication*, 18(1), 69-86.

Sankar, H., Subramaniyaswamy, V., Vijayakumar, V., Arun Kumar, S., Logesh, R., & Umamakeswari, A. (2019). Intelligent sentiment analysis approach using edge computing-based deep learning technique. *Software: Practice and Experience*, 50(5), 645-657.

Sharma, M. (2016). *Theoretical foundations of health education and health promotion*.

Jones & Bartlett Publishers.

Singh, J. P., Rana, N. P., & Dwivedi, Y. K. (2019). Rumour Veracity Estimation with Deep Learning for Twitter. P*roceeding of International Working Conference on Transfer and Diffusion of IT*, 351-363

Sjöström, A. E., Hörnsten, Å., Hajdarevic, S., Emmoth, A., & Isaksson, U. (2019). Primary Health Care Nurses' Experiences of Consultations With Internet-Informed Patients: Qualitative Study. *JMIR Nursing*, 2(1), e14194.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. P*roceeding of Advances in neural information processing systems,* 5998-6008.

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

Wang, Y., McKee, M., Torbica, A., & Stuckler, D. (2019). Systematic literature review on the spread of health-related misinformation on social media. *Social Science & Medicine*, 240, 112552.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., & Funtowicz, M. (2019). Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

Yang, F., Liu, Y., Yu, X., & Yang, M. (2012). Automatic detection of rumor on Sina Weibo. *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, 1-7.

Yu, F., Liu, Q., Wu, S., Wang, L., & Tan, T. (2017). A convolutional approach for misinformation identification. *Proceeding of IJCAI,* 3901-3907.

Zaccone, G. & Karim, M. R. (2018). *Deep Learning with TensorFlow: Explore neural networks and build intelligent systems with Python*. Packt Publishing Ltd.

Zhao, Z., Resnick, P., & Mei, Q. (2015). Enquiring minds: Early detection of rumors in social media from enquiry posts. *Proceedings of the 24th International Conference on World Wide Web*, 1395-1405.

Zhou, Z., Qi, Y., Liu, Z., Yu, C., & Wei, Z. (2018). A C-GRU Neural Network for Rumors Detection. Proceeding of 2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS), 704-708.

Zubiaga, A., Liakata, M., Procter, R., Wong Sak Hoi, G., & Tolmie, P. (2016). Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3), e0150989.