應用語言模型於線上招募詐欺偵測

An Application of Language Model on Online Recruitment Fraud Detection

洪英皓 [1]
國立高雄科技大學 企業管理系 研究生
F108157110@nkust.edu.tw
余銘忠 [2]
國立高雄科技大學 企業管理系碩士班 教授
yminchun@nkust.edu.tw

摘 要

近年來因科技網路快速進步，為生活帶來不少的效益。在企業招募人才方面，雖然透過網路招募能大幅減少成本，卻衍生出許多求職陷阱，不少民眾受害。而近年來人工智慧的興起，在不少應用領域中皆有突破的成果，但本研究發現過去較少學者研究網路招募詐欺偵測議題，並且較少研究採用深度學習的方法進行偵測。因此本研究基於語言模型(BERT, Word Embedding, Topic Model)來提出兩種模型架構(BERT-DNN,GRU-DNN)，藉此來探索適合應用在網路招募詐欺偵測的語言模型為何。本研究以EMSCAD、維基百科作為模型訓練與測試資料集，藉此驗證深度學習方法在招募詐欺偵測中的表現。在研究結果中，發現深度學習模型的表現相較於過往研究更為突出，其中GRU-DNN更為突出。因此未來若採用本研究所提出的研究架構，將能有效的降低民眾受害之機率。

關鍵詞：BERT,自然語言處理,主題模型

1. 緒論

1.1 研究背景與動機

隨著網路科技的發展，人們的生活型態也有大量的轉變。在資訊媒體傳播業也受到重大的變革，從傳統的平面廣告，報紙、雜誌、地方性徵才夾轉變到現今的自家公司招募網站、人力仲介平台、社交媒體平台等。而隨著網路的便利性、低成本、時程短特性，為企業雇者帶來了不少效益，但卻衍生出不少網路求職受到詐騙的案例。而常見的詐騙手法像是在求職網或社群媒體上提供高薪、高報酬、高福利的徵才廣告來吸引求職者上門等，或是利用學生暑期打工或是在家即可工作等宣稱來進行詐騙。

對於這些常見詐騙手法，政府、警方以及各大人力招募網站也不斷的宣導如何防範詐騙，像是7不3要的概念(台北求職防騙)以及像是留意詐騙的行為特徵(廖炳棋，2020)等防範作法。儘管人力銀行網站上皆有提醒如何防範網路招募詐騙行為(104人力銀行)，但是現今的詐騙手法日新月異的變遷加上警方及專家的工作量有限，因此還是有需多人受害。

雖然各國網路招募詐欺案件日漸增多(Australian Bureau of Stastics, 2016；Vidros et al., 2017)，但過去鮮少有人研究線上詐欺檢測此議題。一所希臘大學的學者們Vidros et al. (2017)從人力招募第三方平台Workable蒐集了2012-2014年的工作廣告並將它整理成資料集，隨後實驗室也開放存取權限給學術研究用。此資料逐漸開啟學術在此領域上的研究趨勢，但本研究發現過去學者在研究此議題時，採用傳統機器學習模型(Ensemble, SVM, Logistic Regression等)與實證分析的特徵工程居多，較少採用了以類神經網路為基礎的深度學習技術來測試(Vidros, 2017；Mahbub, 2018；Alghamdi, 2019；Reddy, 2018；Lal, 2019) 。由上述以上動機，本研究將使用上述的資料集來建立兩種不同的深度學習模型(BERT-DNN,GRU- DNN)，並將兩種模型進行績效的比較與評估。

1.2 研究目的

本研究與以往研究不同的方向為採用深度學習模型來進行詐欺偵測。除此之外，我們也比較預訓練模型及GRU的架構在此領域中的分類效果以及不同的詞向量模型(word2vec ,fastText)對於模型的影響力。另外，也增加了新的人工特徵提取(Manual Feature Extraction)方法來驗證是否能提升模型的表現。綜觀上述幾點，歸納了本研究在線上招募詐欺偵測之研究目的:

1.2.1 在此領域中提深度學習模型架構(BERT-DNN,GRU-DNN)並驗證其表現。

1.2.2. 調查不同的詞向量模型對於詐欺偵測模型之影響力。

1.2.3. 使用並驗證隱含狄利克雷分布(Latent Dirichlet Allocation)以及其他特徵提取(feature extract)方法來找出潛在的資訊。

2. 文獻探討

2.1 自然語言處理

自然語言處理(Nature Language Process, NLP)為語言學領域與人工智慧的結合，其目的為令電腦具有理解人類語言能力的技術，常見的應用有機器翻譯(Machine Translation)、機器問答(Question Answering)、詞性標註(POS Tagging)、情緒分析(Sentiment Analysis)、文本分類(Text Classification)、命名實體辨識(Named Entity Recognition)、資訊檢索(Information Retrieval)等應用。

自然語言處理(NLP)與資料探勘(Data Mining)最大的差異在於其資料型別為沒有既定形式的非結構化資料(陳昱儒，2019)。這些文字相較於傳統的數據資料處理難易度複雜許多，如何提取具有保存文字資訊的特徵以及處理高維且稀疏的資料成了現今熱門的研究議題。

基於自然語言處理的技術發展相當廣泛，主要可被分類為語言模型(Language Model)、主題模型 (Topic Model)、分類、分群等。

而本研究的議題被歸類於文本分類的領域中，因此採用了語言模型(Language Model)、主題模型 (Topic Modeling)、分類的技術，來對線上招募詐欺偵測進行分析。

2.2 主題模型

過去的文檔生成過程視為不斷從一個詞袋(Bag of words,Bow)中隨機提取詞彙而成，因此跟人類在真實寫作時的邏輯不太符合(陳昱儒，2019)。大部分人類寫作時，都會去訂定幾個主題，並且書寫的用詞都會圍繞在那個主題附近。這些主題為一篇文檔的核心概念，而一篇文檔可以有一至多個主題構成，從文檔與字彙的關聯中挖掘出潛在主題的過程稱為主題模型 (Topic Modeling) (陳昱儒，2019)。主題模型主要應用在挖掘出各文檔中的隱藏主題，通常是非監督式學習的技術。最常見且被廣泛使用的技術為潛在語義分析 (Latent Semantic Analysis),機率潛在語意分析(Probability Latent Sematic Analysis),隱含狄利克雷分布(Latent Dirichlet Aollcation)。這些技術往往被應用在主題分類,資訊檢索(Information Retrieve),資訊提取(Information extraction)等。

本研究採用的技術為隱含狄利克雷分布(LDA)。

2.2.1 隱含狄利克雷分布(Latent Dirichlet Allocation)

由於 PLSA 的算法認為每篇文檔具有相同的主題分布且每個主題文字只有一種機率分布，因此 Blei et al. (2003) 基於 PLSA 及貝式理論提出了潛在狄利克雷分布。加入了貝式統計的先驗機率-狄利克雷分布的特性，令每篇文檔都能有不同的主題分布，每個主題也有不同的文字分布。這個算法的概念就像是每篇應徵文關於工作內容的主題比例會不太一樣，而每個工作內容提到"應變"這個字的比例也不相同。

機率模型如圖 2-1 所示,其中 w 為文本的字詞,z 為尋找的隱藏主題,Θ 為文檔的主題分布，φ為主題的字詞分布，α,β為θ,φ的先驗分布參數。

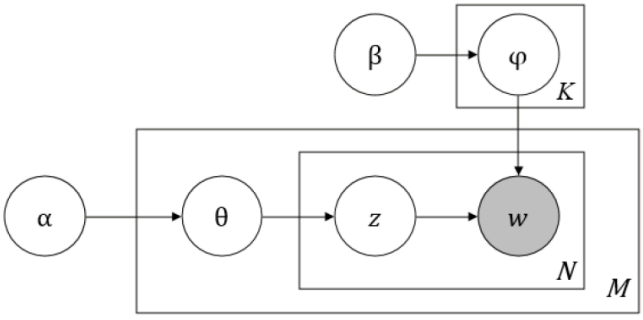

圖 2-1 隱含狄利克雷分布模型圖

資料來源:陳昱儒(2019)

常見的方法為使用吉布斯採樣(Gibbs Sampling)來實踐隱含狄利克雷分布。吉布斯採樣為依次估算條件機率的方法，為馬可夫鏈蒙地卡羅模擬的一個特例(陳昱儒，2019)。此抽樣法使用時機為當後驗機率難以直接進行觀察時，以近似抽取樣本的方式來估計其後驗分布。

而在吉布斯採樣(Gibbs Sampling)實踐隱含狄利克雷分布的過程為:

2.2.1.1 首先隨機對語料庫 D 中的每個字詞 W 指派一個初始的主題 $z^{(0)}$,然後計算各主題 Z 中的字詞分布φ，各文檔 d_i 中各主題的機率分布θ_i，得到初始的主題分布以及主題的字詞分布。

2.2.1.2 隨機抽取 D 中第 i 個文檔的第 j 個字詞 w_{ij} 出來，接著計算除了 w_{ij} 字詞以外所有字詞的主題機率分布，接著依照這個主題機率分布來計算出 w_{ij} 屬於各主題的機率:

$$P\left(z_{ij}\middle|z_{-(ij)},W,D\right) \tag{2-1}$$

2.2.1.3 然後即可得到 w_{ij} 的主題機率分布，並依照 w_{ij} 的機率分布來指派 w_{ij} 一個新主題 $z_{ij}^{(1)}$。

2.2.1.4 然後再重新計算新的字詞分布及主題分布。

2.2.1.5 不斷重複執行步驟 2,3,4，直每個文檔中的主題分布θ_i及每個主題的字詞分布φ收斂為止。

2.2.2 Topic 主題數量選擇

對於主題模型而言，決定合適的主題數量來維持高品質的主題是一個莫大的挑戰。過多的主題會導致主題的內容大同小異，而過少的主題則是讓主題的概念難以區分，因此本研究先訂定不同的主題來進行建模，隨後再依照衡量模型表現的指標進行選擇。本研究參考過去研究(陳昱儒，2019)採用的主題數量評估方式 Topic coherence(Röder & Hinneburg, 2015) 。

Topic Coherence 為計算主題內部兩兩字詞的相關性來進行衡量準則。其算法如以下公式(2-2):

$$\text{coherence(V)} = \sum_{(v_i,v_j)\in V} score\left(v_i,v_j,\epsilon\right) \tag{2-2}$$

其中 V 為該主題的字詞集合，ϵ為避免計算出的分數不為實數。而在計算兩字詞相關性的方法,採用的方式為 UMass coherence score，計算方式如下(2-3):

$$\text{score}\left(v_i,v_j,\epsilon\right) = \log\left(\frac{D\left(v_j,v_i\right)+\epsilon}{D\left(v_j\right)}\right) \tag{2-3}$$

其中$D\left(v_j,v_i\right)$為同時擁有 vj 及 vi 的文檔數量, $D\left(v_j\right)$為擁有 vj 的文檔數量,因此 UMass coherence score 計算所得的值為給定一個詞彙，計算出兩字詞同時出現在一個文檔中的機率，越靠近 1，代表兩字詞越相關，這樣的衡量方式較貼近人類的正常情況。

2.3 深度學習

現今的深度學習為基於人工神經網路的概念發展而來，同時深度學習也為機器學習的一個子領域。人工神經網路的架構為模仿生物神經網路結構和功能所產生的數學化模型。這種概念可源自於 Frank Rosenblatt(1957)的研究中所提出的模型，而此時的神經網路被稱為感知器(Perceptron)，此架構屬於線性模型，無法處理非線性的問題(徐雅玲，2018)，直到 David Rumelhart et al.(1986)提出的倒傳遞神經網路演算法(Back Propagation)才漸漸地能處理非線性問題。除了基礎的神經網路架構，現今也依照不同的情境發展了多種架構，其中最常應用在自然語言處理的架構為遞歸神經網路(Recurrent Neural Network,RNN),由以下部分逐一介紹。

2.3.1 長短期記憶(Long short-term memory, LSTM)

長短期記憶是遞歸神經網路(Recurrent Neural Network,RNN)家族中最具代表性的模型。相較於基礎的 RNN,此種模型能解決 RNN 的梯度消失(Gradient Vanish)問題(Bengio, 1994)。原因在於它改良了隱藏層神經元的機制，將隱藏層在各時間點傳遞資訊的部分使用一個新的架構 Memory-cell 來取代。其中 Memory cell 組成為三個閘道(Input Gate, Forget Gate, Output Gate)加上一個 cell 來控制資訊的傳遞。各 Gate 的運算公式如下:

Input Gate: $i_t = \sigma(W_i \cdot X_t + U_i \cdot h_{t-1} + V_i \cdot C_{t-1} + b_i)$ (2-2)

Forget Gate: $f_t = \sigma(W_f \cdot X_t + U_f \cdot h_{t-1} + V_f \cdot C_{t-1} + b_f)$ (2-3)

Output Gate: $O_t = \sigma(W_o \cdot X_t + U_o \cdot h_{t-1} + V_o \cdot C_{t-1} + b_o)$ (2-4)

Candidate state: $\tilde{C}_t = tanh(W_c \cdot X_t + U_c \cdot h_{t-1} + V_c \cdot C_{t-1} + b_c)$ (2-5)

Cell state: $C_t = i_t \cdot \tilde{C}_t + f_t \cdot C_{t-1}$ (2-6)

Hidden Output: $h_t = O_t \cdot tanh(C_t)$ (2-7)

其中 U,V,W 為權重矩陣,ht-1 為前一隱藏層的 output, Ct-1 為前一個儲存在 memory cell 的值,Xt 為當下的輸入資料,b 為偏差值，Candidate cell state 為當前時間點輸入資料所計算的資訊。透過此機制讓長短期記憶(LSTM)能捕捉時間間隔較長或延遲較久的序列資料關聯性。

由於 LSTM 能彌補這樣的缺失，因此在許多研究中非常的熱門，像是李政霖(2019)將 LSTM 應於即時預測空氣品質的議題中。Badjatiya (2017)則是比較不同的以神經網路模型在惡意評論中學習的文字表徵品質，實驗結果發現 LSTM 相較於其他模型具有最好的表現。Wang (2017)則是將 LSTM 應用在中文評論的情緒分析中，其實實驗結果也顯示加入了新優化器(Optimizer)能有效改善 LSTM 的表現。Liu & Lane (2016)使用具有注意力機制的 LSTM 架構模型來處理意圖分類(Intent classification)與槽填充(Slot filling)問題。

2.3.2 Gated Rucurrent Unit(GRU)

GRU(Cho et al., 2014)與長短期記憶(LSTM)一樣能夠處理遞歸神經網路為基礎的模型(RNN-based)遇到的梯度消失及無法捕捉處理時序較長的資料的問題。但相較於長短期記憶(LSTM)的運算機制，只有 Update Gate,Reset Gate 這兩個 Gate(Chung et al., 2014)，因此需要訓練的參數量比長短期記憶(LSTM)少了很多。而 GRU 的各 Gate 運算機制之公式(Chung et al., 2014)如下:

Update Gate: $z_t = \sigma(W_z x_t + U_z h_{t-1})$ (2-8)

Reset Gate: $r_t = \sigma(W_r x_t + U_r h_{t-1})$ (2-9)

Input: $\tilde{h}_t = tanh(W_t x_t + U_t(r_t \cdot h_{t-1}))$ (2-10)

Output: $h_t = (1 - z_t)h_{t-1} + z_t \tilde{h}_t$ (2-11)

其中 W、U 為權重矩陣，ht-1 為前一個時間點的隱藏狀態，$\tilde{h}_t$為當前時間點所計算的隱藏狀態。而 GRU 與長短期記憶(LSTM)的差異整理成以下幾點(Chung et al., 2014):

1.相較於長短期記憶(LSTM)，無法決定是否要讓隱藏層狀態輸出

2..相較於長短期記憶(LSTM)，隱藏層未具有 cell

3.在更新當前的隱藏狀態h_t時，無法獨立性的保存輸入的資訊($\tilde{h}_t, h_{t-1}$)

4.相較於長短期記憶(LSTM),在計算新的輸入資訊$\tilde{h}_t$時，能夠考慮是否要加入前一個時間點的隱藏狀態h_{t-1}

2.4 詞向量(Word Embedding)

在文字轉換成向量的表示法中，常見的作法為運用了獨熱編碼的詞袋模型(BOW)來轉換。但此種方法未考慮到每個詞彙的關聯性。為了處理此問題並能用於計算詞彙之間的關聯性，Hiton (1986)提出分散式表徵(distributed representation)的概念。其中最具代表性的模型為 Mikolov et al. (2013) 所提出的 word2vec(Skip-gram,CBOW)。Skip-gram 的目標函數為替當前詞彙找到合適的詞來 Maximize 附近詞彙的機率；而 CBOW 則反之，藉由此兩種做法能計算出涵蓋語意的詞向量。

而過了不久 Mikolov et al. 提出了負採樣(Negative Sampling)的方法，將上下文的詞彙視為正樣本(Positive sample),其餘的則是負樣本，然後只抽樣部份的負樣本以及正樣本詞彙來計算目標函數，就能大幅減少計算複雜度過大的問題。

雖然 Skip-gram, CBOW 應用的領域非常廣泛，但是其最大的缺點在於得到的詞彙表徵(Word Representation, Word Repr.)會因為語料庫(Corpus)的規模大小來改變字彙表徵品質。

2.4.1 fastText

當字彙是稀少字或是在訓練資料未擁有的詞彙時，word2vec 中學到的字彙表徵(Word reprenstation.)往往不是那麼的理想。相較於 word2vec, fastText 考慮到了字元層面的 n-gram 資訊，將詞彙視為由一個子詞集合(Subword Set)組合而成。故其作法為先透過 word2vec 的計算方法來得到這些子詞的表徵，再將屬於該詞彙的子詞合併進而得到詞彙表徵。

這樣的算法能解決稀少字、未出現的字以及某些語言型態變化(Morphology)豐富的問題。在實證結果中，此模型也證實了在文字相似性與 Word Analogy 的任務中具有更好的表現，且受到語料庫規模大小的影響較小(Bojanowski et al., 2017)。

2.5 BERT

對於 word2vec 這類型的詞向量來說，最大的問題在於無法產生多義詞，令相同的字依照上下文而產生不同的意思(楊竑昕，2019)。因此後續的語言模型研究而發展的模型，像是 ELMo (Peters, 2018) ,GPT (Radford, 2018)都能解決此問題。但是 ELMo,GPT 的缺點在於其運算規則只考慮單方向的語意，因此處理字彙層面的問題時效果較不穩定。

基於上述問題，Devlin et al. (2018)提出了 Bidirectional Encoder Representation from Transformers(BERT)。其模型架構採用了 Transformer 的 Encoder 部分。而 BERT 具有兩種階段，分為預訓練(pre-train)及微調(fine-tuning)階段，如圖 2-2 所示。

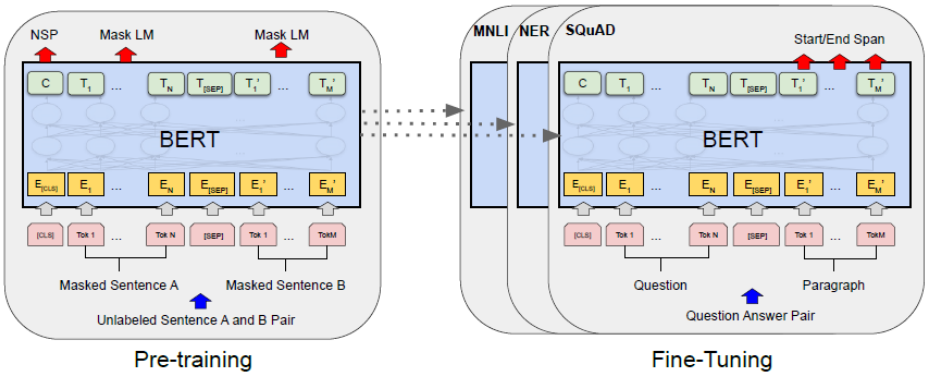

圖 2-2 BERT 不同階段示意圖

資料來源: Devlin et al. (2018)

在預訓練階段(pre-train stage)，BERT 會依照其目標函數 (MLM, NSP)來同時進行 training，在微調(fine-tuning)階段則是依照不同的下游任務來進行微調(Devlin et al. ,2018)。

由於 BERT 可依照不同的下游模型來進行微調，因此產生了多樣化的應用。吳承軒(2019)將 BERT 應用於機器問答的中文文章中，實驗發現相比於雙向 GRU 及檢索技術的方法，準確率提升 21.19%。Chen et al. (2019)將 BERT 應用在意圖分類(Intent classification)及槽填充(Slot Filling)的問題上，透過 Joint Learning 的方式大幅的改善了模型的範化(Generalize)能力。Adhikari et al. (2019)則是將 BERT 運用在文件分類上，實驗結果顯示相較於過去的模型，BERT 能夠得到最好的分類效果。而 Song et al. (2020)則是採用 Pooling 策略來整合 BERT 內部隱藏層的輸出，並將其應用在提升情緒分析以及自然語言推論(Nature Language Inference)的範化能力。

3 研究方法

3.1 研究架構

本研究提出之研究架構可分為三大子架構:分別為文字/資料預處理、模型訓練階段、模型測試階段。

首先，在文字/資料預處理的部分中，我們將資料分成四種類別資料進行資料處理，分別為含有 HTML 標籤的文字資料、文字資料、名目資料、類別尺度的資料。在模型訓練階段，先藉由 EDA 來查看資料分布找出可能潛在規則，再透過特徵工程/轉換來進行特徵提取，最後則是設並建立模型建構以及資料流來訓練模型。本研究建立的

模型為 BERT-DNN 以及 GRU-DNN。隨後將這兩個設計的模型採用評估指標來進行選擇。

在模型測試階段，將資料透過訓練階段的特徵工程流程來轉換資料，再將資料使用訓練好的模型來去預測該廣告是否為詐欺的文章。整體的研究流程如圖 3-1。

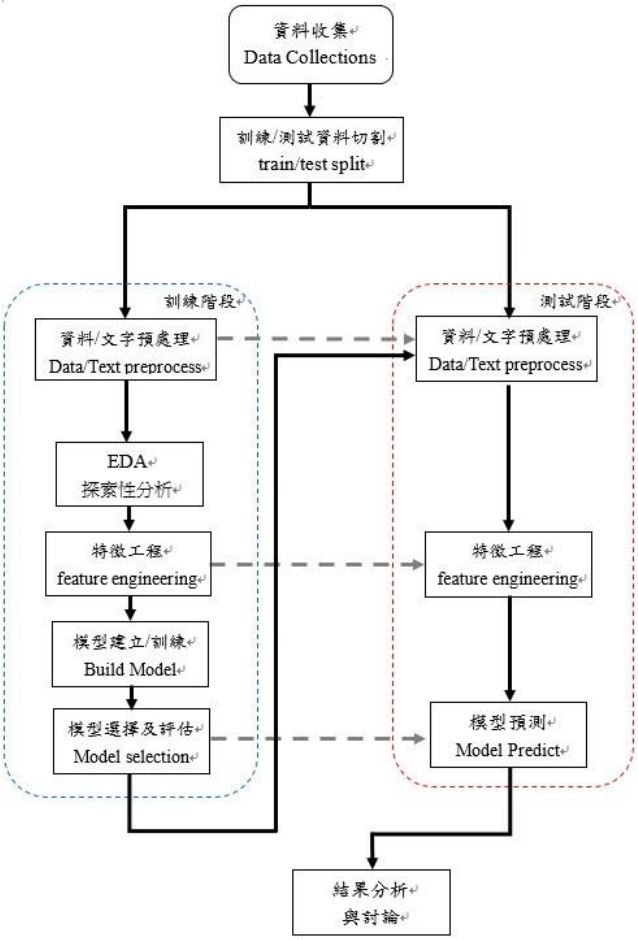

圖 3-1 研究架構流程

3.2 文字與資料預處理

為了提升模型判讀的精準度以及去除影響模型的雜訊，需要對不同資料型態的資料進行預處理。

在文字部分，首先進行文字清洗；在去除雜訊的部分，將會去除 HTML 標籤、表情符號、換行符號、不間斷空白字元、非斷句類以及連續型的標點符號、非 ASCII 的字元。在替換文字的部分，則是替換掉 URL 連結以及以"()" 標示的註解文字。 接下來預計採用 python spaCy 套件中的模型進行斷詞與斷句、詞性標註(POS tagging)、詞形還原(Lemmatization)以及命名實體識別(NER)的方法來處理文字。

在合併文字資訊合併與缺失值處理方面，我們基於提出的兩種模型架構來採用不同的策略。BERT-DNN 中使用" [SEP]" 特殊詞彙來合併文字欄位，若欄位有缺失值則是填補"[empty]"特殊詞彙來替代；GRU-DNN 中，各欄位句子間使用"[SEP]"特殊詞彙合併，而欄位之間則是直接合併，若為缺失值則使用"[PAD]"來填補。

在名目尺度的資料部分，則是將其類別轉換成數字並用 0 來填補缺失值。除此之外，我們也透過 Batch Normalization 的方法來進行正則化。

3.3 詞向量訓練

本研究將採用 fastText,word2vec 詞向量模型進行預訓練，藉此調查詐欺偵測模型影響力。此外，為了確保詞向量模型的品質，本研究將採用具有龐大詞彙量之語料庫-2017 年 1 月 1 日的英文維基百科資料存檔來進行預訓練，以便後續比較不同詞向量維度對模型影響力。

詞向量模型以 gensim 開源函式庫進行實作，分別建立 word2vec 與 fastText 之 CBOW 與 Skip-gram 模型，除了模型維度的選擇外，其餘參數接為固定如下:

視窗大小為 5，最小辭頻數量為 5，最大詞彙量 100000，訓練回數 6 回。

3.4 模型架構

3.4.1 BERT-DNN

首先在 BERT 子模型部分，將涵蓋 HTML 標籤的文字欄位兩兩合併成一個欄位(Company profile 與 Benefits、Job Description 與 Requirements)並分別輸入至同個 BERT，藉此得到各時間點的涵蓋上下文資訊之詞向量。接著採用不同的全連接層與 Pooling 策略來合併各時間點的詞向量，藉此整合成文本的表徵。而最後本研究採用連接(concat)與轉換的方法來合併欄位的文本表徵。

而在工作標題 (Title) 部分，由於其資料特性不具有語法結構，所以我們只採用詞向量的方法來得到詞彙表徵，隨後使用 Mean pooling 的方法來得到該標題的表徵。並將其表徵與其餘資料進行連接與轉換。

最後我們將 HTML 文本表徵、文字與其餘資料的資訊一同輸入至 2 層全連接層在透過輸出層(Output layer)來進行詐欺偵測判斷。此模型架構如圖 3-2。

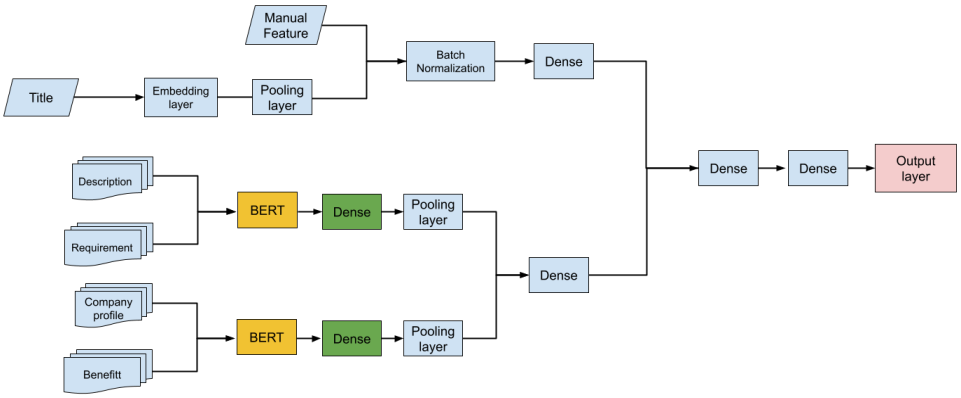

圖 3-2 BERT-DNN 模型架構圖

3.4.2 GRU-DNN

此模型架構與 BERT-DNN 相似，主要不同的點在於將序列特徵部分改為採用預訓練的詞向量模型與各自的 GRU 來提取文本表徵，在標題部分 (Title)則是先採用預訓練的詞向量模型以及 Mean pooling 來提取表徵，隨後將未經過 pooling 策略的表徵與 Description 的表徵送至 Attention block 進行計算。模型架構如圖 3-3。

在 Attention block 部分，其運算方法採用 Scaled Dot-Product Attention (Vaswani et al. ,2018) 的算法，query 為 Title 的表徵，而 key 與 value 則是 Description 的表徵。而計算後的輸出將與工作內容的表徵進行連接與轉換來做為 Attention block 的輸出。模型架構如圖 3-4。

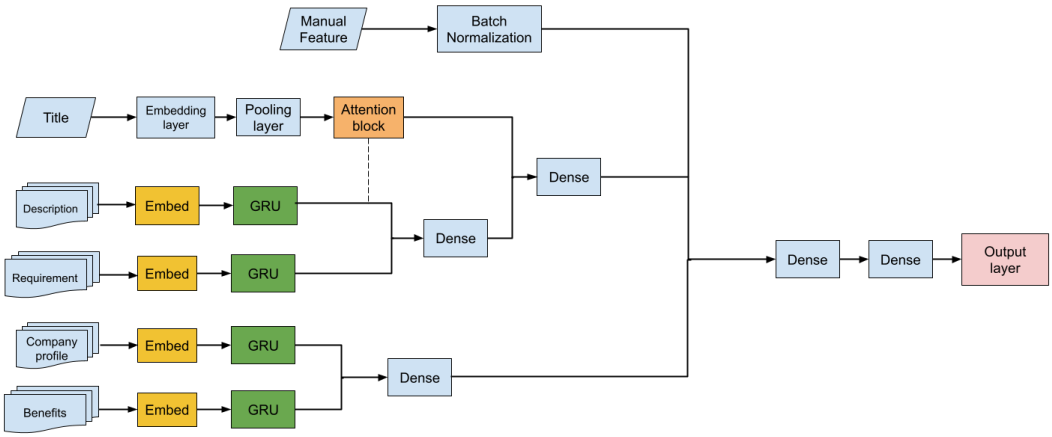

圖 3-3GRU-DNN 架構圖

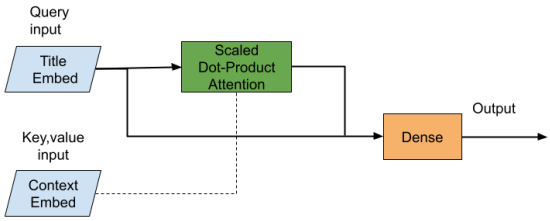

圖 3-4 Attention Block

3.5 模型評估指標

由於我們資料的標籤類別分佈呈現極度的不平衡，為偏態分佈。過去研究也驗證了準確率在不平衡資料的情境為不是有效指標，且對資料類別分佈的變化非常敏感 (He & Garcia, 2009)。因此我們採用了 TNR,Precision, Recall, F-measure, Balanced Accuracy 的指標來衡量模型的成效。

假設我們有一個二元分類的混淆矩陣(Confusion Matrix)來表示模型分類結果與真實資料類別的對應數量 (表 3-2)，其中若欄位名稱加入 True，則代表模型分類為正確的，而 False 則反之。

表 3-1 混淆矩陣(Confusion Matrix)

	Actual Positive	Actual Negative
Predicted Positive	True positive(TP)	False positive(FP)
Predicted Negative	False negative(NG)	True negative(TN)

資料來源: Johnson & Khoshgoftaar (2019)

而 Precision 為衡量模型在判定正類別(Positive)中有多少比例是正確的(式 3-1)，Recall 則是衡量對於真實資料的正類別中，模型預測正確的有多少比例(式 3-2)。F-measure 則是考慮了 Precision 與 Recall 的調和平均數，透過β係數來調整 Recall 的重要性(式 3-3)。而 TNR 則是衡量對於真實資料的負類別中，模型預測正確的有多少比例(式 3-4)最後，Balanced Accuracy(3-5)則是將 TPR 與 TNR 進行平均而得(Johnson & Khoshgoftaar, 2019)。

上述的指標可由如下公式來表示(3-1~3-5):

$$Precision = \frac{TP}{TP + FP} \tag{3-1}$$

$$TPR = Recall = \frac{TP}{TP + FN} \tag{3-2}$$

$$F - measure = \frac{(1 + \beta^2) \times Recall \times Precision}{\beta^2 \times Recall + Precision} \tag{3-3}$$

$$TNR = \frac{TN}{TN + FP} \tag{3-4}$$

$$Balanced\ \ Accuracy = \frac{1}{2} \times (TPR + TNR) \tag{3-5}$$

4 實驗與討論

4.1 實驗資料集

本研究所採用的資料集為 EMSCAD (Vidros et al., 2017)。其資料從網路招募平台 Workable 中所收集，收集年限為 2012-2014 年的工作招募廣告(Job Ad.)，其中合法的廣告具有 17014 筆，違法的 886 筆，共計 17880 筆資料。在標記資料的部分，原作者藉由委外來標註，其判定準則為求職者的投訴數量、未有聯絡及公司資訊等。此資料集共

計有 17 個結構化與非結構化欄位，如表 4-1。

表 4-1 資料欄位說明

資料類別	欄位名稱	欄位概述
字串	標題(Title)	工作廣告標題
	地理位置(Location)	工作時所在的地理位置
	部門(Department)	公司內部部門單位(行銷部)
	薪水範圍(Salary Range)	工作薪資範圍(ex:$50000-60000)
具有 html 標籤的字串	公司概要(Company Profile)	公司簡介(Company description)
	工作闡述(Job Description)	關於職缺內容的細節說明
	條件要求(Requirements)	工作條件要求(ex:學經歷)
	福利(Benefits)	雇主提供給職缺的福利
名目(Nominal)	雇用種類(Employment type)	Ex:全職、兼職、約聘
	經驗要求(Required experience)	Ex:實習生、
	學歷要求(Required education)	Ex:博士、碩士、學士學位等
	產業別(Industry)	Ex:自動化、健康照護、IT
	職缺類別(Function)	Ex:顧問、工程師、研究員
二元類別	遠端工作(Telecommuting)	若有聯絡方式，則為 1(True)
	公司 logo (Company Logo)	若有公司 logo，則為 1(True)
	職缺問題 (Questions)	若問題欄有紀錄，則為 1(True)
	詐欺 (Fraudulent)	資料標籤(label)

4.2 資料與文字前處理

4.2.1 資料前處理

本研究所採用的資料集具有 2,943 筆重複資料，為了避免造成訓練模型的偏誤，將其去除後剩下 14,937 筆。而原資料集因為沒有測試資料集來測試模型的範化能力，因此我們以 0.8 與 0.2 比例的方法來切割成訓練與測試資料集。

除此之外，為了完整的驗證訓練後模型的表現，本研究採用 3 Fold 隨機分層抽樣(StratifiedShuffleSplit)的方法來評估，每次 Fold 分別將訓練資料以 0.9 與 0.1 比例進行切割，再將其驗證分數進行平均作為模型的驗證分數。最後再利用先前切割的測試資料集來評估模型的範化能力，各資料集的筆數如表 4-2。而訓練階段，為了處理不平衡資料集的問題，我們採用 Random Under Sampling 的方法(He & Garcia, 2009)。將負類別(Negative) 的資料筆數抽樣至正類別資料筆數與其損失權重之相乘結果相同，解此來減少不平衡資料集對模型的影響。

表 4-2 各 Fold 資料集筆數

	Fraud(Positive)	Legit(Negative)	Total
Train(Each Fold)	431	10323	10754
Val(Each Fold)	48	1147	1195
Test	120	2868	2988

4.2.1 文字前處理

由於在文字欄位中包含了許多雜訊，為了讓模型能獲得正確的資訊，本研究採用 3.2 所提及之文字清洗與預處理方法來進行處理與轉換。轉換與清洗後的結果如表 4-3。

表 4-3 文字前處理範例

文字清洗前-以條件要求欄位為例
Required Skills: High level knowledge of: Red Hat Enterprise Linux and IBM AIX operating systemsRocket (U2) UniVerse database technologyJBoss (WildFly) server technology. VMware and related infrastructure technologiesStorage Area Networks and storage devices (preferably IBM)Basic understanding of: Windows Server platformNetworking concepts – Layers 2 and 3 – preferred experience w/CiscoStorage transport protocols – iSCSI, Fiber, FCoECloud computing – SaaS, IaaS, etc.
文字清洗後-以條件要求欄位為例
Required Skills High level knowledge of Red Hat Enterprise Linux and IBM AIX operating systems Rocket Uni Verse database technology JBoss server technology. VMware and related infrastructure technologies Storage Area Networks and storage devices Basic understanding of Windows Server platform Networking concepts Layers 2 and 3 preferred experience w/Cisco Storage transport protocols i SCSI, Fiber, FCo ECloud computing SaaS, IaaS, etc.
預處理與斷句斷詞後-以條件要求欄位為例
[['required', 'skills', 'high', 'level', 'knowledge', 'of', '[org]', 'and', '[org]', 'aix', 'operating', 'system', '[product]', 'database', 'technology', '[norp]', 'server', 'technology'], ['vmware', 'and', 'related', 'infrastructure', 'technology', '[org]', 'and', 'storage', 'device', 'basic', 'understanding', 'of', 'windows', 'server', 'platform', 'networking', 'concept', 'layers', 'and', '[cardinal]', 'preferred', 'experience', 'w', '[org]', 'transport', 'protocol', 'i', '[org]', '[org]', 'fco', 'ecloud', 'compute', 'saas', 'iaas', 'etc']]

4.3 實證分析與特徵工程

本研究藉由統計方法來挖掘出潛在的文字特徵，首先計數各欄位的文字數量來進行觀察，在結果中(圖 4-1)，發現在公司、工作簡述欄位中，詐欺者較少去填寫資訊，這與 Vidros et al. (2017)所提出的觀點一致。除此之外，本研究亦採用魏克生等級和檢定(Wilcoxon Rank Sum Test)來調查詐欺與非詐欺類別的差異性，其檢定結果如表 4-4。

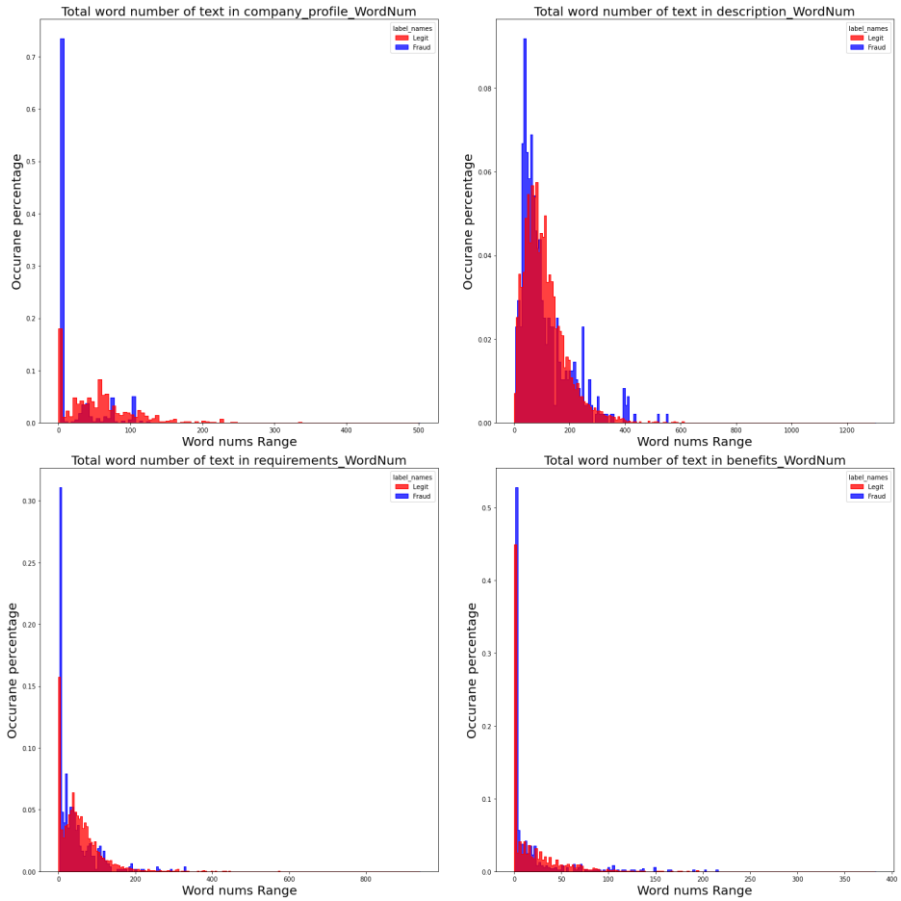

圖 4-1 文字數量分布圖

表 4-4 魏克生等級和檢定

欄位名稱		Kolmogorov Smirnov Test		Wilcoxon Rank Sum Test	
		統計量	p-value	統計量	p-value
Company profile	Legit	0.8202	0.0	-21.08	1.17e-98
	Fraud	0.5	2.94e-111		
Description	Legit	0.999	0	-4.44	8.93e-6
	Fraud	0.9882	0		
Requirement	Legit	0.8566	7.43e-285	-9.309	1.28e-20
	Fraud	0.7559	0		
Benefits	Legit	0.552	2.94e-111	74.272	1.93e-5
	Fraud	0.5	0		

除了從文字中挖掘潛在的特徵，本研究也藉由觀察兩類別在各欄位中的比例差異，來挑選出適合的特徵作為模型的額外輸入。依照 Vidros et al. (2017)的實證分析，發現在表公司 Logo 欄位中詐欺廠商較少張貼公司 Logo 且與合法的廠商比例相差了 54.3%，證實了詐欺公司因為無法向政府登記營利，因此基本上不會有公司的商標。而在遠端工作方面，則是發現詐欺公司可接受遠端工作的數量較多，也符合現今詐欺廠商喜好用在家工作便有高薪收入的情境。最後在工作詢問部份，本研究也發現合法的公司徵才貼文相較於非法的多了 6%的人會詢問公司職缺內容的資訊。

表 4-5 數值欄位比例差異

	公司 Logo		遠端工作		工作詢問	
	Legit(%)	Fraud(%)	Legit(%)	Fraud(%)	Legit(%)	Fraud(%)
有	82	27.7	4.2	7.5	44.86	38.6
無	17.9	72.23	95.7	92.4	55.13	61.37

在工作與學歷要求方面，從表 4-6 與 4-7 的兩類別比例中可以得知，相較於合法的徵才公司，大多數的詐欺公司對於學歷與工作要求門檻非常低，大多都要求高中職與基層員工的工作經驗即可，這與許多詐騙案例中的標榜口號「免經驗即可工作」的傾向符合，因此本研究依照其階級順序轉為數值，藉此作為模型的輸入來調查對模型判斷詐欺貼文時的影響力。

表 4-6 學歷要求比例

類別	unspecified	Some high school course work	High school	vocational HS diploma
Legit	15.13%	0.08%	19.3%	1.6%
Fraud	15.69%	6.72%	37.21%	6.27%
類別	vocational	Vocational-degree	Associate	Some college course work
Legit	0.66%	0.09%	3.01%	1.2%
Fraud	0%	0%	2.24%	1.34%
類別	Bachelor	Master	Doctorate	Profressional
Legit	52.83%	4.68%	0.33%	0.85%
Fraud	20.6%	8.07%	0.44%	1.34%

表 4-7 工作經驗要求比例

類別	Not Applicable	Intership	Associate	Entry-level
Legit	10%	3.9%	22.21%	21.3%
Fraud	13.8%	4.06%	7.31%	43.4%
類別	Mid-senior	Director	Executive	
Legit	37.46%	3.55%	1.3%	
Fraud	26.42%	3.65%	1.2%	

4.3.3 主題模型分析

除了採用現有的數值型特徵，本研究也想藉由運用主題模型從工作闡述中挖掘出潛在的主題分布，作為額外的特徵。而為了建立一個良好的主題模型，本研究採用 topic coherence score(U_MASS)來選擇最適當的主題數量。

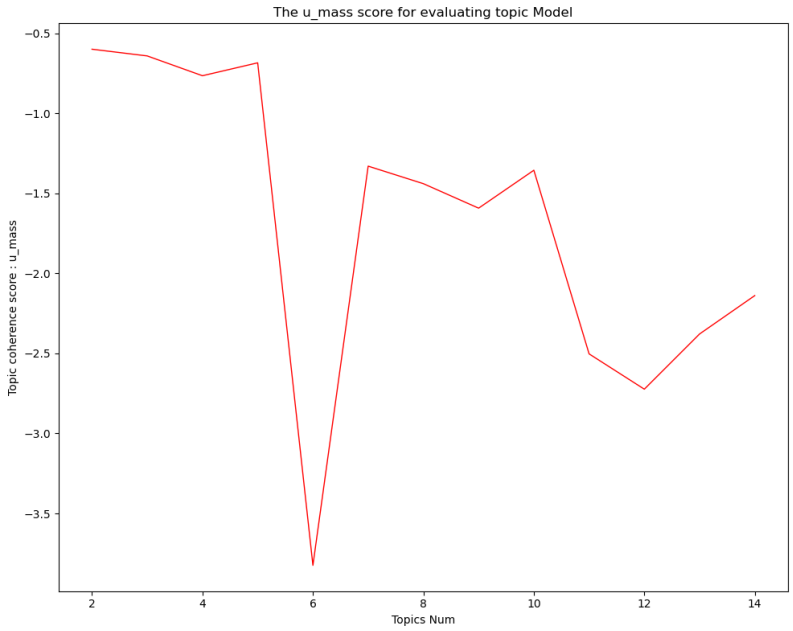

圖 4-3 主題模型之 Topic coherence score

圖 4-3 為依照不同主題數量建立的主題模型所得的 Topic coherence score.從圖 4-3 可得知當主題數量為 2,3,5 時具有較良好的主題分佈。由於主題數量過少可能導致每個主題的概念模糊不清；過多時又可能導致多個主題內容大同小異，因此我們選擇主題數量 3 作為最佳主題數量。

4.4 訓練模型最佳參數設置與評估分析

在深度學習領域中，如何找出一組能讓模型表現最佳的參數為重要的因素。故本研究設計了一系列的實驗來進行尋找，在每次的實驗中皆變動一個參數值，藉此找出各個最佳參數值，並且在每次實驗中皆採用前述所提到的 3Fold 策略來降低最小誤差。而本研究在各模型的初始參數設定為採用 256 維度的隨機初始化詞向量模型、512 隱藏層為度、Dropout 為 0.1、最大文本長度 128，採用停用詞與數值資料。在經過實驗後發現，對於詞向量模型的選擇，BERT-DNN 與 GRU-DNN 採用 128 與 256 維的 CBOW，在隱藏層則是 256 維與 512 維。而為了避免文本過長導致資訊丟失、複雜度過大，兩者模型皆設置最大長度為 64。另外，為了確保模型有一定的範化能力，兩者 Dropout 皆設置 0.3。在文字與數值資料方面則是採用移除停用詞與加入數值資料的策略，藉此來給予模型更多的資訊與去除噪音。最後，本研究將最佳參數訓練的模型與本研究之初始預設參數模型進行比較。各模型結果如表 4-8 所示。

表 4-8 GRU-DNN 與 BERT-DNN 比較

Test set(3 Fold StratifiedShuffleSplit)					
	Accuracy	F1-score	precision	Recall	TNR
GRU-DNN(初始)	0.8621	0.5822	0.4786	0.7694	0.99
GRU-DNN(最佳)	0.8253	0.6105	0.523	0.733	0.988
BERT-DNN(初始)	0.7874	0.435	0.3334	0.6277	0.9837
BERT-DNN(最佳)	0.7876	0.4197	0.3141	0.6333	0.9839

從實驗結果中可看到 GRU-DNN 在採用最佳參數設置後，模型在範化能力上有明顯的改善，尤其在精確率方面改善了約為 5%的水準，相較於 BERT-DNN 則是沒有明顯的改善。除了評估最佳參數的影響力，本研究也與先前研究(Vidros et al., 2017)所訓練的機器學習模型(Random Forest)比較。由於先前模型採用整個資料集進行評估，故我們也採用整個資料集來驗證最佳參數模型之表現，藉此保持比較的一致性，三種模型結果如表 4-9。

表 4-9 過往研究模型表現比較

All DatSet					
	Accuracy	F-score	precision	Recall	TNR
GRU-DNN	0.9733	0.7713	0.649	0.9686	0.9986.
BERT-DNN	0.9436	0.5976	0.4384	0.9394	0.9973
Random Forest	0.8269	0.4098	0.282	0.751	0.9028

在表 4-9 中可以發現不管是 BERT-DNN 或是 GRU-DNN 都比 Vidros et al. (2017)採用的方法之表現還要好。本研究也發現在各實驗中 GRU-DNN 的表現遠比 BERT-DNN 還要好，故推論可能與輸入格式不同有關，由於 BERT based 方法在 pre-trained 階段時未看過具有空值欄位的資料格式，導致後續在 fine-tuning 模型時較難去找出能判讀此資料格式的參數。因此本研究建議若為來在此資料集上發展深度學習模型時，須考慮到如何訓練出一個能處理空值資料格式的模型。

5.結論與建議

5.1 結論

由於現今科技快速的發展，公司招募人才的管道也逐漸從平面媒體轉換成網路媒體。雖然此轉變大幅改善了公司的成本，卻衍生出了許多網路求職陷阱。儘管政府不斷宣導防範求職的理念，還是有不少民眾受害。近年來隨著大數據與人工智慧的崛起，其技術在不少研究議題上皆有突破性的成果。然而在自然語言處理的文本分類任務中，較少學者研究網路招募詐欺偵測的相關議題，大多數皆為假新聞辨識的研究。直到 Vidros 等人於 2017 年貢獻了網路招募詐欺偵測的資料集 EMSCAD ,才逐漸開啟學術上對此議題的研究趨勢。

但本研究發現在招募詐欺偵測領域中，過往研究只採用機器學習的方法進行偵測，尚未採用深度學習模型進行辨識。故本研究運用深度學習方法提出兩種模型架構 BERT-DNN 與 GRU-DNN 來處理，並採用 EMSCAD 資料集來與過往模型進行驗證與比較差異。除了提出不同的模型架構，本研究亦採用與過往研究不同的文字預處理方法，採用了主題模型 LDA,NER 與特殊 token 填補缺失值的方法來進行處理，並將詞彙轉換成詞向量來提升模型績效。本研究也以一系列的實驗設計探討了各參數設置對於模型的影響力，在最終的研究結果中，發現了兩種提出的模型皆比過往研究採用的機器學習模型表現還好。但也發現由於具有缺失值的資料格式與預訓練階段(pre-trained)時的資料格式稍微不同，導致 BERT-DNN 在詐欺偵測的能力上不及 GRU-DNN。因此未來研究方向可從如何採用或設計一個能有效處理缺失值資料格式的預訓練模型進行探討。

5.2 研究限制

本研究提出的模型雖能改善招募詐欺偵測的問題，但由於 EMSCAD 資料集的語系為英文，導致無法得知若模型應用在中文資料集中的表現如何。另外，由於資料集中涵蓋了些許錯字，可能產生模型會誤判文章中的語意資訊之問題。本研究亦發現許多資料中文字屬性的部分皆為空值、具有重複的資料等問題，若按照最原始的方法直接刪

除未符合完整性的資料，將會導致資料量過小的問題，容易導致模型在此領域中具有過擬合的現象產生，目前也較少人收集並標註此領域的資料，因此暫無能有效增加資料量的方法。

此外，由於本研究提出的 BERT-DNN 模型參數量過大，因此在訓練模型時對於硬體設備的規格要求較高。

5.3 未來研究建議

未來研究在進行 EMSCAD 資料集的文字預處理時，可採用更細緻的斷詞方法(BPE, WordPiece 等)來處理錯字、詞之間無空白符號等問題。而雖然本研究採用填補特殊詞彙的方法來處理文字屬性缺失值的問題，但容易造成未能符合預訓練模型輸入資料之格式，進而影響模型表現。因此如何讓預訓練模型也能應對缺失值欄位或是適當的填補方法，將是未來研究可探討的方向之一。

在挑選資料欄位方面，本研究參考於過去研究(Vidros et al. ,2017)提出的實證與相關性分析結果來進行選擇，因此未來研究能再進一步探索各屬性對模型的影響力，甚至亦可採用其他的學習方法(Joint learn)來改善模型在招募詐欺偵測領域中的範化能力。

而本研究提出的模型架構中，GRU-DNN 會因為文本長度過長而導致運算時間過久，BERT-DNN 則是因為參數量過大而導致硬體規格要求較高。故未來研究可進一步探討不同種類的深度學習方法在網路招募領域中的表現，例如採用可支援平行化運算之模型 CNN, Transformer 來解決時間複雜度過大；採用 Network compression 此類技術來減少模型參數量等。未來可依循此方向來建立一個更完善的線上招募詐欺偵測系統。

6.參考文獻

中文部分：

104 人力銀行(無日期)。104 職場安全網。無日期，取自: https://www.104.com.tw/area/104safety/p1.cfm

台北求職防騙 (無日期)。無日期，取自:台北市政府勞動局台北求職防騙網頁: https://job7n3y.bola.taipei/#

吳承軒(2019)。基於 BERT 語言模型之多國語言機器閱讀理解研究。國立台北科技大學資訊工程系碩士論文。

李政霖(2019)。及時空氣品質動態監測系統結合 LSTM 模型預測 PM2.5 濃度之應用。國立台北科技大學電機工程學系碩士論文。

徐雅玲(2018)。利用多模態模型混和 CNN 和 LSTM 影音特徵已自動化偵測急診病患疼痛程度。國立清華大學電機工程學系碩士論文。

陳昱儒(2019)。基於隱含狄利克雷分布進行開放式問卷之主題導向文字探勘。國立中央大學資訊工程學系碩士論文。

廖炳棋(民 109 年 7 月 12 日)。暑期學生求職潮，應徵財務助理倫帳戶遭詐團洗贓款。聯合新聞網。民 109 年 7 月 12 日，取自: https://udn.com/news/story/7320/4695705

英文部分：

Adhikari, A., Ram, A., Tang, R., & Lin, J. (2019). Docbert: Bert for document classification. arXiv preprint arXiv:1904.08398.

Alghamdi, B., & Alharby, F. (2019). An Intelligent Model for Online Recruitment Fraud Detection. Journal of Information Security, 10(3), 155-176.

Australian Bureau of Stastics. (2016, April 20). Personal Fraud. Retrieved November 3, 2020, from https://www.abs.gov.au/statistics/people/crime-and-justice/personal-fraud/latest-release#scam-fraud

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2), 157-166.

Bengio, Y., Simard, P., & Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks, 5(2), 157-166.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146.

Chen, Q., Zhuo, Z., & Wang, W. (2019). Bert for joint intent classification and slot filling. arXiv preprint arXiv:1902.10909.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555.

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.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on knowledge and data engineering, 21(9), 1263-1284.

Hinton, G. E. (1986, August). Learning distributed representations of concepts. In Proceedings of the eighth annual conference of the cognitive science society (Vol. 1, p. 12).

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 27.

Lal, S., Jiaswal, R., Sardana, N., Verma, A., Kaur, A., & Mourya, R. (2019, August). ORFDetector: Ensemble Learning Based Online Recruitment Fraud Detection. In 2019 Twelfth International Conference on Contemporary Computing (IC3) (pp. 1-5). IEEE.

Liu, B., & Lane, I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. arXiv preprint arXiv:1609.01454.

Mahbub, S., & Pardede, E. (2018). Using contextual features for online recruitment fraud detection.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Reddy, M. N., Mamatha, T., & Balaram, A. (2018, January). Analysis of e-recruitment systems and detecting e-recruitment fraud. In International Conference on Communications and Cyber Physical Engineering 2018 (pp. 411-417). Springer, Singapore.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65(6), 386.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. nature, 323(6088), 533-536.

Song, Y., Wang, J., Liang, Z., Liu, Z., & Jiang, T. (2020). Utilizing BERT intermediate layers for aspect based sentiment analysis and natural language inference. arXiv preprint arXiv:2002.04815.

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

Vidros Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017, April). Deep learning for hate speech detection in tweets. In Proceedings of the 26th International Conference on World Wide Web Companion (pp. 759-760).

Wang, J., & Cao, Z. (2017, October). Chinese text sentiment analysis using LSTM network based on L2 and Nadam. In 2017 IEEE 17th International Conference on Communication Technology (ICCT) (pp. 1891-1895). IEEE.