Communications of the CCISA
Vol. 24 No. 4 Oct. 2018

以中文文本分析為主的線上社交訊息作者辨識

柯冠廷[1]、葉國暉[2]、駱立軒[3]
[1, 2, 3]國立東華大學資訊管理學系
[1]nchureborn@gmail.com、[2]khyeh@gms.ndhu.edu.tw、[3]610639012@gms.ndhu.edu.tw

摘要

本研究主要探討基於社交聊天訊息文本之身份鑑別，近年來，線上社交詐騙行為頻傳，大多情況為利用社交工程手法進行個人帳號之盜用，鑑於此，本研究以此現象為研究標的，希望能建立一套有效率的身分鑑別系統以辨別文本訊息之來源使用者的真實性與合法性。研究方法中將以使用者的社交文本訊息作為使用者鑑別資料來源，並利用語意分析模型(Semantic Analysis Model)、多層感知器(Multilayer Perceptron, MLP)與支援向量機(Support Vector Machine, SVM)做為主要的資料分析演算法，進行使用者鑑別符元的產生與鑑別準確率檢測。研究成果顯示，在語意模型分析實驗中，有 65%的檢測案例之相似度皆低於 70%，而多層感知器分析與支援向量機分析則分別可達到 80%與 88%的鑑別準確率。

關鍵詞：身份鑑別、社群網路、語意模型、支援向量機、多層感知器

Toward to a stylometric analysis model for the authorship verification of online social message

Guan-Ting Ke[1], Kuo-Hui Yeh[2], Li-Hsuan Lo[3]
[1, 2, 3] Department of Information Management, National Dong Hwa University
[1]nchureborn@gmail.com, [2]khyeh@gms.ndhu.edu.tw, [3]610639012@gms.ndhu.edu.tw

Abstract

Recently, cases of scamming on social media keep pouring in. Most cases are related to hacked social media accounts, which belong to those who suffered from identity stealing by social engineering. In this research, we focus on how users' instant messages can be exploited to defeat identity thieves. We proposed an authentication system based on stylometry of users' instant messages, which is able to tell whether the current user of the account having both of its representation and perpetuity. We collect users' instant message as the raw data for training process, create the classifiers through Latent Semantic Analysis (LSA), Multilayer Perceptron (MLP) and Support Vector Machine (SVM). The research result pointed out that, with only LSA model equipped, 65% of test cases reach lower than 70% of similarity, while utilizing

MLP and SVM can reach 80% and 88% of accuracy, respectively.

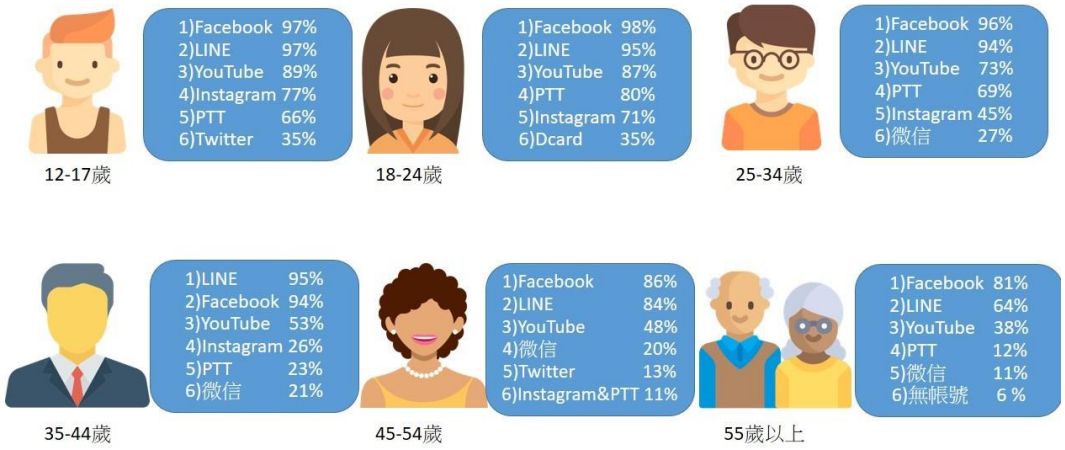
Keywords: Authentication, Social Media, Semantic Analysis Model, Support Vector Machine, Multilayer Perceptron

壹、前言

1.1 研究背景與動機

隨著資訊科技的發展，網路逐漸普及化，人們常常利用電腦及手機來使用網路服務。網路服務提供了許多生活的便利性，人們可以藉由網路購物、享受數位娛樂、資訊查詢等，網路改變了人們的生活，也使人們的生活圈不僅僅限於日常生活中，而產生了一群以相同興趣活動建立的線上社群的社群網路，在社群網路上，人們有許多互動的方式，如聊天、寄信、影音、檔案分享、部落格、新聞群組等，多樣的活動使得社群網路在近年來蓬勃的發展。

近幾年，隨著社群網路爆炸性的發展，出現了許多不同的社群平台，每個平台都有大量的使用人數。在台灣，從小孩到老人都會使用社群軟體，各個軟體的使用比率也不同，如 Facebook 擁有極高的使用率，2016 年台灣地區 Facebook 每月活躍使用人數就高達 1800 萬人[24]，圖一顯示了台灣各年齡層擁有社群網站帳號比例。因此，Facebook 平台每天都會產生大量的資料，這些資料形式為張貼文章與影音。此外，人們常藉由手機使用社交聊天軟體，導致每個時間點都會產生出大量的即時通訊資料，故資料處理時的效率與安全性常會成為使用者使用該社群服務時的主要評估標準之一。再者，在社群網路的匿名性下，社群網路常被利用於不良的使用，像是假新聞、網路詐騙等[26]，也讓文章(與其作者)的動機性、真實性與合法性逐漸被人們所重視，更顯示了網路使用者身分鑑別(Authentication)的重要性。



圖一：台灣各年齡層擁有社群網站帳號比例[23]

身份鑑別是個重要的研究領域，生活中大大小小的事都應用到身份鑑別，如常見的證件證明便是一個例子。在資訊科技發達的現今，每個人也都能擁有一個從個人生物訊號衍生而來的個人鑑別符元，這方面的研究也很多元，像是藉由觀察腦波、瞳孔、指紋等生物特徵或是分析網路上點擊紀錄、鍵盤輸入按壓力道等行為特徵做身份的合法性與真實性驗證。近年來，部分研究探討了網路假新聞及詐騙行為的動機與防護方式，大多以文章驗證作為分析主軸，然而，在網路的匿名特性下，使得該問題一直無法完美地被解決。一般來說，文章的驗證可以藉由文章的差異性去做分類，但在各種語言特性的不同下，導致文章驗證的分類效果與驗證準確率一直無法提升，且大多的研究以英語為主，其他語言相關研究則較少。在中文方面，因其字詞多義，在自然語言處理上總比英文來的難處理，因此繁體中文方面的文本驗證研究一直是被學術社群所重視的重要研究議題之一。

1.2 研究目的

隨著社群網路的快速發展下，網路聊天成為人們最常用的溝通方式，但網路聊天中常變隨著身分假冒和詐騙等問題，此外，網路匿名特性亦使得問題變得更為棘手。儘管網路聊天的訊息資料大幅增加，已提供了足夠的資料量以用來支援身分鑑別時的資料分析，但目前繁體中文做自然語言分析與身分鑑別的研究並不多。鑑於此，本研究將以繁體中文文本分析為基礎的身分鑑別系統開發作為研究主要方向，比較現有的文本分析方法結果並以此為基礎，希望藉由此研究產生出一套以中文訊息作為驗證標準且鑑別率高的系統，並將此系統用於 Facebook 上的文章作者身份鑑別以及其他社群軟體上身份鑑別的使用，進一步對網路詐騙防範產生正面效應的幫助。

本研究將收集 Facebook Messenger 社交聊天文本訊息資料並以 Python 語言處理訊息資料特徵，透過矩陣運算產生 Term by document 資料矩陣，再以語意模型和機器學習作為實驗方法，在機器學習方法中我們使用其中的支援向量機和多層感知器，我們將使用特徵資料矩陣去訓練模型，最後依訓練好的模型進行實驗測試得出身分鑑別結果。本文架構為四個章節，第一章主要說明本論文的研究背景、動機與目的。第二章文獻探討將針對社群網路、即時通訊、身分鑑別概念和定義，以及語意模型與機器學習進行介紹，了解其發展與相關應用。第三章研究方法依據本篇論文之概念提出身分鑑別系統的架構，進行研究實驗規劃與流程的解說。第四章將進行研究實作並記錄流程以及探討研究結果背後的意義，第五章則為本文的結論。

貳、文獻探討

2.1 社群網站

社群網站是一種新的人與人互動的社交模式，隨著資訊科技的迅速發展，人們的社交範圍不再受到地理環境的限制，在網路上，人們可以藉由遊戲、聊天、直播、影音、新聞分享等互動成為一個社群。社群網站的定義如下：(一)允許個人在有界限的系統上建立半公開的個人資料，(二)允許個人清楚表達哪些使用者與其有連結關係，(三)允許查看系統中其他人所建立的連結關係[4]，且從上個世紀以來社交一直被認為人類行為的關鍵因素，因此有許多計算機科學家對於社群網路使用者的行為進行了基礎研究分析，社群網路的定義為一種有向圖，其中每個節點代表一位使用者，而每個邊為節點之間的關係[3]。

目前有許多不同的社群網站如 Facebook、Twitter、Instagram、YouTube、微博等，每個社群網站的服務內容不同，吸引的人群也不同。因此，在各個地區的使用率也不同。以台灣為例，Facebook 在各年齡層的使用人數最多，其次為 YouTube，再者為 Instagram。由於使用人數眾多，Facebook 產生了大量且足夠的資料，因此本研究以 Facebook Messenger 資料作為實驗資料集，望能藉由分析這些資料了解人們的行為模式。

圖二：社群網站世界使用分佈圖[6]

2.2 即時通訊

即時通訊是一種透過計算機與網路達成幾乎同步的一對一通訊系統，使用者不需處於同個空間，且同一時間能進行多個一對一通訊，也能多人同時通訊，但需雙方同時在線且允許通訊[16]。隨著資訊科技的發展，透過即時通訊已能夠傳遞文字訊息、檔案、語音、視訊交流等，即時通訊越來越受大眾歡迎，目前台灣即時通訊軟體的使用率已位居手機 app 排名中的第一名[27]。

台灣常見的即時通訊軟體種類非常多，從早期的 Yahoo! Messenger、Windows Live Messenger 到近年來的 Skype、Facebook Messenger、LINE，即時通訊軟體已經存在了十幾年，隨著人們的需求變化，這些通訊軟體的服務不僅僅提供傳遞文字訊息的基礎功能，還包括各種影音檔案、情境貼圖以及視訊電話功能，目前以 LINE 及 Facebook Messenger[25]最受我國民眾所青睞。隨著網路的蓬勃發展，未來可能出現更多元的服務內容，使即時通訊軟體的使用狀況更加興盛。

2.3 語意模型

潛在語意分析(Latent Semantic Analysis; LSA)最早在 1990 年被提出，是以數學統計為基礎的語意提取模型，為向量空間模型的一種延伸，主要是用來改善圖書館的索引和

搜尋引擎多查詢性能(Manning)[14]。潛在語意分析為一種資訊檢索的技術，主要透過奇異值分解(Singular Value Decomposition; SVD) (Golub & Reinsch)[9]和維度縮減(Dimensionality Reduction) (Deerwester)[7]為核心之語意推導模型，分析文件字詞矩陣中所存在的潛在語意關係，潛在語意分析將文件從稀疏的高維語句空間映射到一個低維的向量空間，稱之為潛在語意空間，而其中映射的方法使用 SVD 和維度縮減來達成，接著以向量空間模型(Vector Space Model; VSM)計算文件語意相似性。

潛在語意分析的優點是對字詞與文件之間的關係進行維度縮減過濾噪聲，減少儲存空間，經過奇異值分解後得到的對角矩陣為整份文件的語意空間，而語意空間就是文件中每個字詞的定義空間，每個字詞可以透過語意空間的定位來找到其代表意義，將能夠發現字詞間的相似性、文件和文件的相似性，文件和字詞間的語意關聯，對同義詞有良好的處理結果;由於潛在語意分析基於 SVD 運算處理，因此迭代計算次數很多，在處理大量資料時，文件與字詞的維度將會迅速成長，使得 SVD 的計算複雜度非常高，潛在語意分析在一詞多義問題上處理結果不佳，此外潛在語意分析並沒有明顯表達出字詞出現次數的機率模型。

潛在語意分析廣泛應用於各種資訊檢索(Information Retrieval)領域(Landauer)[13]，其中 Evangelopoulos[8]討論許多潛在語意分析應用，例如語言學、心理學、認知科學、教育學等、Kuo 等人[12]針對影音資料使用 Multiple-type Latent Semantic Analysis 來分析背景音樂和視頻的潛在關聯性、Klein 等人[11]使用潛在語意分析針對考試命題中的簡答題進行評估，判斷其答案是否正確，而 Ozsoy 等人[17]則利用潛在語意分析進行文件整理，達成文件彙總的目的。

2.4 機器學習

機器學習(Machine Learning)是人工智慧學習的一個分支，從 1940 年代開始陸陸續續被提出討論，Samuel [18]設計出一個自我學習的西洋棋跳棋程式，能依據棋盤上棋子移動的情況來判斷結果，經過大量的資料訓練後已勝過人類棋手，Rosenblatt[19]提出感知器(Perceptron)的概念，抽象描述一種如人類神經元運作的系統，藉由輸入特徵向量，透過感知器輸出結果，為最簡單的類神經網路，Vapnik[20]等人提出支援向量機(Support Vector Machine; SVM)，能處理高維度資料，將高維度資料以向量型態映射到高維度空間，用以分類目標，屬於一種高效能的分類演算法。機器學習是為了使電腦能夠自行從資料中找出規律並得出解決問題的方法，藉此發展出人工智慧，幫助人類解決各種不同的難題。近年來由於硬體設備的發展與大量資料的產生，使得機器學習再次被廣泛討論研究，並解決過去無法解決的問題。

機器學習可以由其訓練模式主要分為兩類，監督式學習和非監督式學習。監督式學習是一種需要提供已標記好結果的資料，讓電腦從資料特徵中學習規律並且比較標記好結果，進而得出辨別機制，常見的演算法如支援向量機(SVM)、最近鄰居法(K Nearest

Neighbors)、決策樹(Decision Tree)等。非監督則不提供標記好的資料，讓電腦自行觀察學習資料特徵中的不同，常見的演算法如迴歸分析(Regression Analysis)、k-平均演算法(K Means Clustering)等。此外還有演算法是同屬於兩者的，如類神經網路演算法(Neural Network)，類神經網路為近年來發展最興盛的研究領域，提升了人工智慧研究方面的進展。

機器學習廣泛應用於各種領域，從早期的棋盤遊戲到現今的金融、醫療、生物科技、文本分析等等，Wu 等人[21]使用支援向量機來偵測社群網路詐欺事件，藉由分析用戶操作行為模式來判斷，此研究在 2 分鐘內可以達到 80%以上準確率，觀察 7 分鐘後可以達到 90%以上準確率; Mantyjarvi 等人[15]使用最近鄰居法、多層感知器(Multilayer Perceptron, MLP)和線性辨別函數(Linear Discriminant Function, LDF)做情感識別，藉由生物傳感器蒐集回來的物理訊號分析，三種分類方法都能達到約 80%的準確率。

2.5 身份鑑別

身份鑑別是一種用來辨識使用者身份的方法如證件證明以及生物特徵證明，依據辨識特徵的不同有不同的方法，主要有生物特徵和金鑰密碼等[28]，生物特徵方面有大家熟悉的指紋、虹膜、腦波等，而金鑰密碼方面最常見的是數位簽章的應用，隨著資訊科技的發展身份鑑別在日常生活中被廣泛使用如手機解鎖、門鎖、電子檔案確認等等，身份鑑別的研究在不同的領域中都有問題需要被解決。而本研究關注的是在社群網路上的身份鑑別，以社群網路上的聊天訊息做為研究資料，以辨認出使用者身份。

近幾年，社群網路上詐騙事件頻傳，由於社群網路具有匿名性，網路犯罪往往成為棘手的問題，因此網路上身份鑑別的問題勢必需要一套解決的方法。恰好人們在網路上常留下許多文字資料，因此可以藉由這些資料做作者身份鑑別，Abbasi 等人[1]為了解決網路文本匿名問題提出了利用詞法、句法、結構和內容特異性做為分析識別的特徵，並採用支援向量機(Support Vector Machine)與主成分分析(Principal Components Analysis)等作為檢測方法，其結果準確率達到 94%，Brocardo 等人[5]提出作者身份鑑別研究分為三個不同的領域，作者身份識別、作者特徵描述、作者驗證，而作者身份識別和作者特徵描述已有許多研究，但在作者驗證的方面研究較少，為了解決作者驗證問題提出以詞法、句法做為特徵以及從 n 元語法(n-gram)分析中提取新特徵，並以資訊獲得(Information Gain)做為特徵選擇方法且採用支援向量機做為檢測方法，其檢測結果由樣本大小不同等誤差率(Equal Error Rate) 從 9.98%變化到 21.45%。

在作者驗證方面我們將以作者寫作型態作為研究方向，每種語言都有其獨特的句法詞語，每個人的對文字的表達方式也有所不同，在相同的語言學習環境下，人們各自發展出獨特的寫作風格，從這些不同的寫作風格可以觀察及測量其特徵並進一步量化，再藉由這些特徵進行身份驗證，Zheng 等人[22]為了解決網路上留言訊息其作者身份的匿名性，提出了以詞語、句法、結構和特定內容做為特徵，並採取以決策樹(Decision trees)、

反向傳播神經網路(Backpropagation neural networks)和支援向量機為檢測方法，其驗證結果達到 70%到 95%的辨別率。

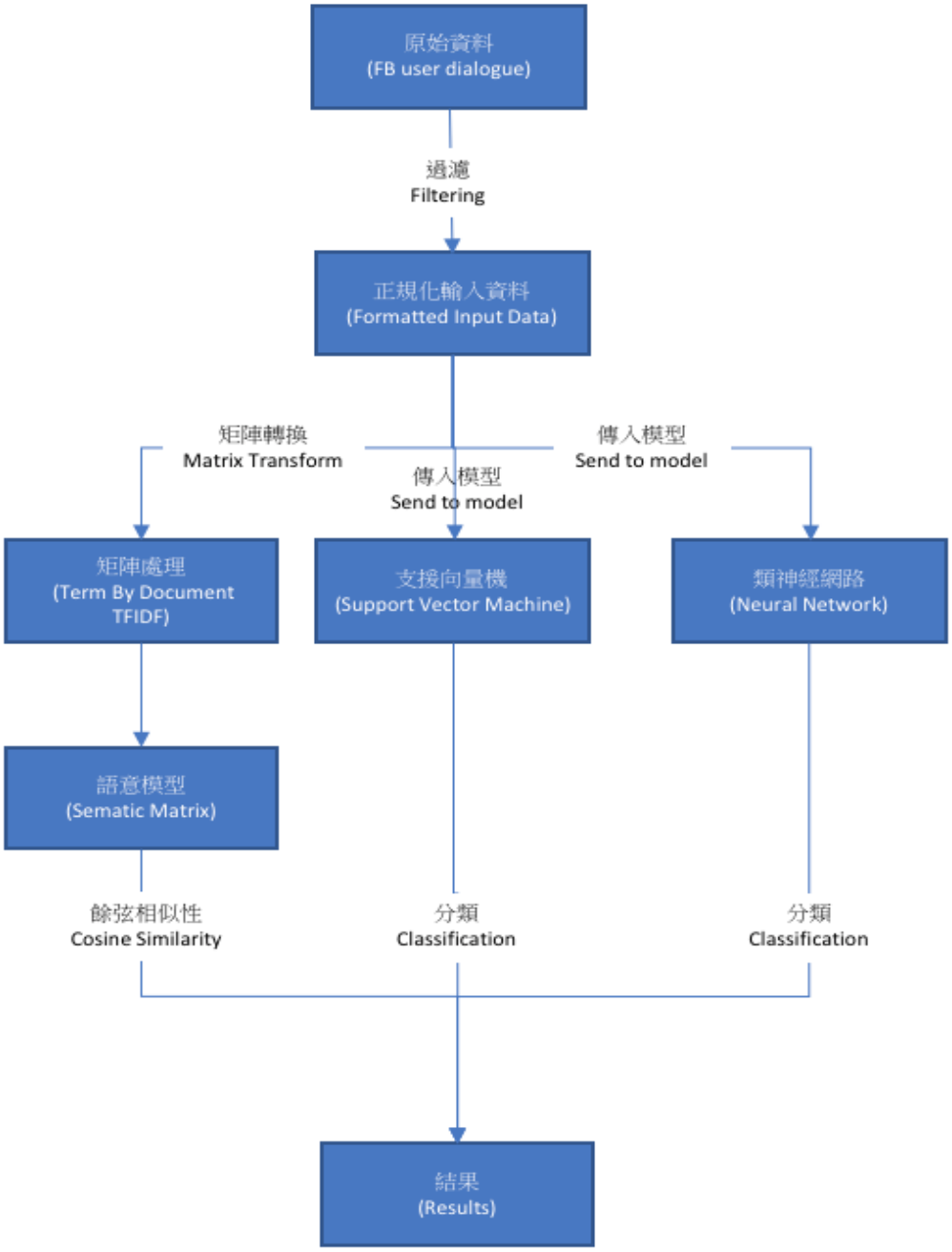
目前大多文獻都以英文文本做研究，其他語言作者身份驗證的研究較少，Albadarneh 等人[2]為了解決阿拉伯語言方面作者驗證問題，以阿拉伯語推特作為文本資料，提出以詞袋模型(Bag-of-words model)及 TFIDF 來計算特徵向量，接著將資料放置 Hadoop 分部是框架上進行樸素貝葉斯分類器(Naïve Bayes classifier)，其準確率為 61.6%。本研究以中文文本做為研究資料，世界上中文使用人口不斷地增加，因此對於中文的研究有其必要性，中文相較其他語言複雜，且大多自然語言處理(Natural Language Processing)的方法不適用於中文，因此本論文將以句法架構做為特徵與使用自然語言處理捨去之內容作為特徵，如字詞長度頻率、語助詞、表情符號、標點符號、斷句格式。

參、研究架構

第三章針本論文的研究架構與流程進行描述。首先第一節說明本研究流程，第二節說明資料來源與處理，第三節說明本研究的三種方法，第四節說明本研究的評估指標。

3.1 研究流程

本研究流程如圖三所示，第一步將整理研究資料集，研究資料來源為 Facebook Messenger 使用者的訊息對話資料，第二步將擷取資料，將有用的資料整理為格式化矩陣輸入資料，接者將資料輸入三種方法。首先第一種為 TFIDF 矩陣處理和語意模型，接者矩陣進行餘弦相似性計算產生結果。第二種方法為支援向量機模型。第三種為多層感知器神經網路，最後比較各方法結果效能。



圖三：研究實驗流程架構

3.2 資料蒐集與處理

本階段第一步為蒐集實驗資料，實驗資料來源為 Facebook Messenger 使用者對話訊息。本研究以一個使用者帳號與其他使用者對話記錄為資料集，將所有資料分類成每位使用者的輸入內容。在正規化輸入資料步驟將原始文字資料轉換為特徵矩陣。特徵矩陣

可分為五個類型，分別為句對話長度、語助詞、標點符號、表情符號、空白符號，而每個特徵類型中之細項將於第四章說明。

3.3 實驗方法

第一種方法為 TFIDF 和語意模型 LSA，首先使用 TFIDF 將原始特徵矩陣進行處理，增加資料顆粒度，並藉餘弦相似性產生結果，再將進行 TFIDF 後的矩陣結果輸入 LSA 模型，降低資料維度產生矩陣結果，最後輸入餘弦相似性比較其相關性，餘弦相似度。第二種方法為支援向量機，實驗將原始特徵矩陣輸入支援向量機訓練，接者以測試資料輸入產生預測結果。第三種方法為多層感知器，首先輸入原始特徵矩陣訓練神經網路，接者以測試資料產生預測結果。

3.4 評估指標

本研究資料經過矩陣和語意模型處理分析後，需要使用評估指標來檢驗研究結果成效。本研究採用的語意模型產出之結果皆以向量矩陣形式呈現，在向量矩陣評估中經常採用餘弦相似性(Cosine similarity)，其特點為計算簡單，可以同時對多個矩陣中向量進行運算，所以計算速度快，常用於資訊檢索領域。餘弦相似性為計算兩個向量夾角，兩向量越相似其夾角越小，餘弦值也越近於 1，其公式如(1)。支援向量機和多層感知器則以預測結果準確率做評估

$$similarity = \cos\theta = \frac{A \cdot B}{|A| \times |B|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}} \quad (1)$$

肆、實驗結果

4.1 研究環境

研究實作中將依規劃的研究流程與模型架設研究環境，硬體設備，中央處理器為 Intel®CoreTM i7-4790(3.40GHz)，隨機存取記憶體規格為 DDR3 SDRAM，容量為 16GB。軟體方面，作業系統為 Windows 10 專業版 64-bit 版，程式開發工具有 Python3.6.1、Anaconda 1.6.9、spyder3.2.4。

4.2 研究資料

本研究採用 Facebook Messenger 上聊天訊息做為實驗資料。如圖四所示，在 Facebook Messenger 上聊天內容資料有文字、檔案、貼圖、影音連結等，而本研究只採用文字資料做分析。

圖四：Facebook Messenger 介面圖

本研究以使用者與十位用戶的聊天資料進行分類，將十位用戶聊天記錄取出並整理後如圖五所示，研究資料以每 50 句為一組輸入資料，為增加資料量進行重疊資料，每次重疊 25 句對話，由於人們對話內容上下句之間大多有關連性，因此重疊不至於大幅影響實驗結果，接著以本研究觀察常用特徵進行特徵比對統計產出矩陣，特徵內容如表一所示，其中語助詞方面參考[29]定義和資料經由 n-gram 後次數高的詞語，而表情符號參考[10]定義。

快來吧
對的
我還在多容
對呀
嗯嗯
等等12在多容
懶得出門Gg
哈哈哈
我要腳架
？
應該會
好好考慮
恩...
幾點？？
還沒拍完咧....
還久
努力中
麻煩˙啊啊啊
哈哈還久??
愛情吧(？)
有關
哈哈
就
我在剪片
沒ㄟ
不會呀
XD
XD
相信有人比我更爛
應該不會
加油XD
考慮XD
XD
XD
有看一下下
XD
~
安抓

圖五：聊天訊息資料

表一：比對特徵資料

單句對話長度	語助詞	標點符號	表情符號	空白符號
加總計算平均長度	還好	~	XD	計算空白出現頻率
	真假	！	厂厂	
	嘿啊	？	QQ	
	欸欸	？！	= =	
	阿災	,	:）	
	嗯		. . .	
	嘿啊		☺	

4.3 資料處理方式與流程

本研究將原始資料整理後經由特徵比對產出 Term By Document 矩陣，接著採用三種方法進行使用者身分鑑別。第一種將矩陣透過 TFIDF 與 LSA 分析後產出向量矩陣，最後將向量矩陣做餘弦相似性後產出辨別結果。第二種將矩陣資料輸入支援向量機進行訓練及預測結果。第三種將矩陣資料輸入多層感知器訓練神經網路模型且預測結果。

本研究首先透過表一所列之比對特徵資料將原始資料轉換成 Term By Document 與 TFIDF 矩陣，再經由 LSA 模型進行資料分析。研究成果顯示經由 LSA 模型分析過的樣本，其相互之間的餘弦相似性大多介於 0 到 0.5 之間，由此判斷本系統在相似度 0.5 的門檻值之假設時，系統將可成功辨別每個人的輸入模式，進行身份驗證。此外，本研究亦採用支援向量機進行模型訓練及預測，將輸入特徵矩陣進行訓練，再用測試資料產生出預測結果，其結果目前達到 88%，擁有較高程度的準確率。在多層感知器方面，本研究採用 2 個隱藏層，隱藏層神經元各為 1000 個，實驗結果也可達到 80%的準確率。

伍、結論

本研究透過語意模型 LSA 與 TFIDF 的搭配、支援向量機和多層感知器等三種資料分析流程，實作了社群網路文本訊息之身份鑑別系統。主要以 Facebook Messenger 做為研究樣本資料之擷取來源，針對訊息內容提取特徵，再透過語意模型和 TFIDF 進行資料處理，並搭配餘弦相似性，藉此得知使用者間的相似程度，得到初步的社群網路身份鑑別結果，而支援向量機和多層感知器則能進一步地提升身份鑑別的預測結果。

由本研究結果顯示原 TFIDF 和經過 LSA 處理後的結果差異性並不大，其原因可能是本研究所採用特徵與語意關聯性不大，因此語意模型 LSA 無法完全發揮其作用。在支援向量機和多層感知器的結果中，我們可看出多層感知器因資料量不足，所訓練出的

結果較不盡理想，而支援向量機結果則較佳，但其擷取之特徵構面仍過於複雜，未來將選取較重要之特徵以降低支援向量機判斷錯誤之可能性。

參考文獻

[1] A. Abbasi and H. C. Chen, "Writeprints: A Stylometric Approach to Identity-Level Identification and Similarity Detection in Cyberspace," *Journal of ACM Transactions on Information Systems*, Vol. 26, No. 7, 2008, DOI: 10.1145/1344411.1344413.

[2] J. Albadarneh, B. Talafha, M. Al-Ayyoub, B. Zaqaibeh, M. Al-Smadi,Y. Jararweh and E. Benkhelifa, "Using Big Data Analytics For Authorship Authentication of Arabic Tweets," *2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC)*, Limassol, Cyprus, Dec 7-10 2015.

[3] J. Botelho and C. Antunes, "Combining Social Network Analysis with Semi-supervised Clustering: a case study on fraud detection," *Mining Data Semantics (MDS'2011) in conjuction with SIGKDD*, 2011.

[4] D. M. Boyd and N. B. Ellison, "Social Network Sites: Definition, History, and Scholarship," *Journal of Computer-Mediated Communication*, Dec 2007, DOI: 10.1111/j.1083-6101.2007.00393.x.

[5] M. L. Brocardo, I. Traore and I. Woungang, "Authorship verification of e-mail and tweet messages applied for continuous authentication," *ACM Journal of Computer and System Sciences*, Vol. 81, pp. 1429-1440, 2015.

[6] V. Cosenza, VINCOS BLOG, http://vincos.it/world-map-of-social-networks/2017 (accessed on 15th April 2018)

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society for Information Science*, Vol. 41, No. 6, pp. 391-407, 1990.

[8] N. E. Evangelopoulos, "Latent semantic analysis," *Journal of the Wiley Interdisciplinary Reviews: Cognitive Science*, Vol. 4, No. 6, pp. 683-692, 2013.

[9] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Journal of the Numerische Mathematik*, Vol. 14, No. 5, pp. 403-420, 1970.

[10] P. Gonçalves, M. Araújo, F. Benevenuto and M. Cha, "Comparing and Combining Sentiment Analysis Methods," *in Proceedings of the first ACM conference on Online social networks*, pp. 27-38, 2013, DOI: 10.1145/2512938.2512951.

[11] R. Klein, A. Kyrilov and M. Tokman, "Automated assessment of short free-text responses in computer science using latent semantic analysis," *Proceedings of the Sixteenth Annual*

Joint Conference on Innovation and Technology in Computer Science Education (ITiCES 2011), Darmstadt, Germany, June 27-29, pp. 158-162, 2011.

[12] F. F. Kuo, M. K. Shan and S. Y. Lee, "Background music recommendation for video based on multimodal latent semantic analysis," *2013 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, Jul. 15-19, pp. 1-6, 2013.

[13] T. K. Landauer, D. S. McNamara, S. Dennis and W. Kintsch, "Handbook of Latent Semantic Analysis," *Psychology Press*, London, UK, 2013.

[14] C. D. Manning, P. Raghavan and H. Schütze, "Introduction to information retrieval," *Cambridge University Press Cambridge*, 2008.

[15] J. Mantyjarvi, J. Himberg and T. Seppanen "Recognizing human motion with multiple acceleration sensors," *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat.No.01CH37236)*, IEEE, Oct. 2001, DOI: 10.1109/ICSMC.2001.973004.

[16] B. A. Nardi, S. Whittaker and E. Bradner, "Interaction and Outeraction: Instant Messaging in Action," *CSCW '00 Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pp.79-88, 2000.

[17] M. G. Ozsoy, F. N. Alpaslan and I. Cicekli, "Text summarization using latent semantic analysis," *Journal of Information Science*, Vol. 37, No. 4, pp. 405-417, 2011.

[18] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, Vol. 3, pp. 210-219, July. 1959.

[19] F. Rosenblatt, "The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain," *Psychological Review*, pp. 386-408, Nov. 1958.

[20] V. Vapnik and C. Cortes, "Support-Vector Networks," *Journal of Machine Learning*, Vol.20, pp. 273-297, Sept. 1995.

[21] S. H. Wu, M. J. Chou and C. H. Tseng, "Detecting In Situ Identity Fraud on Social Network Services: A Case Study With Facebook," *IEEE Systems Journal*, Vol. 11, pp. 2432-2443, Dec. 2017, DOI: 10.1109/JSYST.2015.2504102.

[22] R. Zheng, J. Li, H. Chen and Z. Huang, "A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques," *Journal of the American Society for Information Science and Technology*, Vol.57, No. 3, pp. 378-393, Feb. 2006.

[23] 八成以上台灣人愛用 Facebook、Line 坐穩社群網站龍頭 1 人平均擁 4 個社群帳號 年輕人更愛 YouTube 和 IG https://www.iii.org.tw/Press/NewsDtl.aspx?nsp_sqno=1934&fm_sqno=14 (accessed on 15th Apr. 2018)

[24] 台灣活躍用戶破 1800 萬人，Facebook 鎖定電商發力 https://www.bnext.com.tw/article/40252/BN-2016-07-19-174028-223 (accessed on 15th Apr. 2018)

[25] 社群新寵兒:即時通訊軟體全球使用率上升 12%，更多網路使用者選擇非開放的社群平台 http://www.cna.com.tw/postwrite/Detail/179665.aspx#.WkDY9t-WZhF (accessed on 15th Apr. 2018)

[26] 盜用 LINE 帳號誆稱借錢 中老年族群最易被騙 http://www.chinatimes.com/realtimenews/20170415004271-260402 (accessed on 15th Apr. 2018)

[27] 資策會 FIND/經濟部技術處「資策會 FIND(2016)/ 服務系統體系驅動新興事業研發計畫（2/4）」，https://www.iii.org.tw/Press/NewsDtl.aspx?fm_sqno=14&nsp_sqno=1952 (accessed on 15th Apr. 2018)

[28] 維基百科，https://zh.wikipedia.org/wiki/身份验证 (accessed on 15th Apr. 2018)

[29] 維基百科，https://zh.wikipedia.org/wiki/助詞 (accessed on 15th Apr. 2018)