運用文字探勘技術預測 MBTI 人格特質

邱鈺雯	許毓庭	楊子瑩	陳柏言	楊佳真
獸醫所	資訊管理系	資訊管理所	資訊管理所	資訊管理所
國立台灣大學	國立台灣大學	國立台灣大學	國立台灣大學	國立台灣大學
R10629018	B10705035	R11725001	R11725004	R11725040

1 THE PURPOSE OF OUR PROJECT

人格測驗可以應用在許多領域上，例如臨床心理、諮詢、人力資源管理、職業選擇等，然而人格測驗較為主觀，當受試者不夠了解自己或是被外在因素影響，人格測驗的結果可能會出現偏誤。自動化人格預測(Automated Personality Prediction)可以用較客觀科學的方式解決這個潛在的問題，而利用社群媒體資料進行自動化人格預測也正逐漸受到自然語言處理和社會科學領域的重視。[1]

不少求職者或學生，在選擇職業或科系後，發現自己並不適合所選的領域，為此各式的人格測驗出現，而測驗結果可作為職涯或科系選擇的參考依據。其中，MBTI 測驗又稱為「16 型人格測驗」，近年來成為全球最知名且廣為使用的人格測驗分析。

另外，國內相關以 MBTI 作為研究內容的文獻數量較少，且絕大部分都是針對管理面進行研究，較難找到使用機器學習預測的相關文獻。國內利用文字探勘或機器學習技術預測人格的研究主要是針對五大人格或是九型人格[2][3]，故本專案欲嘗試使用文字探勘技術進行 MBTI 人格特質自動化預測，後續進行相關分析，並對新鮮人職涯發展問題進行應用，為實務應用領域做出貢獻。

2 OUR SOLUTION

2.1 資料收集

本專案採用 Kaggle 公開資料集 -(MBTI) Myers-Briggs Personality Type Dataset[4]，該資料集由論壇 Personality Cafe 上的貼文收集而來，網站特色為每個使用者可以在個人資料設定顯示自己的 MBTI，論壇上也為 16 種 MBTI 人格分別開設討論板，供使用者集中於各板討論。

資料集共有 8675 筆資料，每一筆代表一個人，包含此人 MBTI 類型及 50 篇此人在論壇上的 po 文。另外，資料集有 2 個欄位，分別為 type 及 posts。

(MBTI) Myers-Briggs Personality Type Dataset			
	英文變數名稱	中文變數名稱	型態
1	type	16型人格	Nominal
2	posts	貼文內容	String

表一、(MBTI) Myers-Briggs Personality Type Dataset 資料集原始欄位

2.2 資料前處理

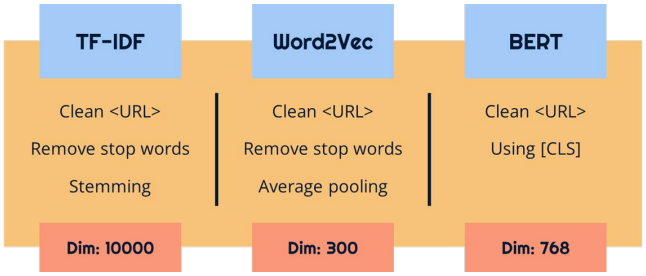
資料集共有 8,600 筆資料，每筆資料皆為一使用者的 MBTI(type)及最近期的 50 篇發文(posts)，每篇發文會以特殊符號|||隔開。文本前處理部分會先移除分隔符號|||、網址，本專案將 http 開頭的網址都取代成<URL>，並以三種方式作前處理。

1. TFTIDF 會將停用字去除，並進行 stemming 處理，在生成 TFIDF vector 的過程中也會進行 lowercase 和移除標點符號的處理。我們將 TFIDF vector 限定維度為 10000 維。

2. word2vec 使用 google pre-trained word2vec 將每個 token 轉成 300 維向量，沒

有在 word2vec pretrain model 的字會直接忽略(標點符號並沒有在 word2vec 字典中)，大小寫因為在 pretrain model 中是有分別的，所以保留大小寫差別。最後將 word vector 做 average pooling 平均當作 document vector。

3. BERT 部分，使用 pre-trained BERT base model，標點符號有在 BERT pre-train model 字典裡，所以會保留，而每個字在 BERT 中則會轉成 lowercase。最後取 CLS 維度為 768 代表 document vector。



圖一、資料前處理方法

2.3 分類預測

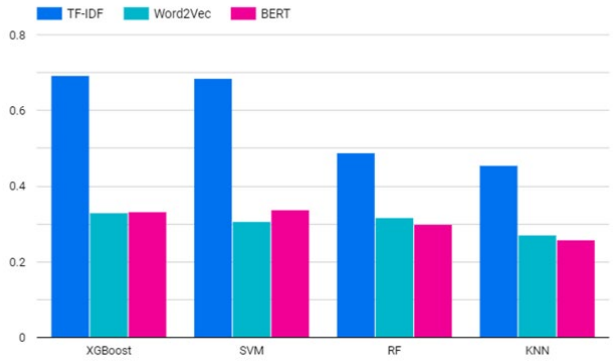
在分類預測的部分使用前一小節前處理過的資料，並以 8:2 的比例切分為訓練及測試資料集。模型挑選參考了 Cui 及 Qi 於 2017 年所做的研究，該研究使用了 Logistic Regression, Naive Bayes 以及 SVM 等方法透過社交媒體上的貼文對使用者的 MBTI 進行預測，結果指出在這幾項不同的作法中由 SVM 帶來最好的表現。[5] 本專案參考選用了 SVM 及 Naive Bayes，並另外使用 KNN, Random Forest 及 XGBoost 來進行訓練，最後針對五個模型的結果加以比較，試圖找出表現最好的結果以利進行後續的應用。

KNN 會先在 1~15 之間的奇數 k 下找出哪一個 k 值可以可以讓模型有最高的準確率，經測試不論是 TFIDF, Word2Vec 或是 BERT embeddings 都是在將 k 設為 13 時可以達到最好的表現。SVM 在嘗試不同的 kernel function 後最後採用準確率最高的 linear kernel。五個模型都是以 default 的模型下去訓練，沒有對模型額外調參。

表二為訓練後的五種分類模型對測試資料進行預測的 accuracy 結果，可以看到 TFIDF 在所有模型都有較佳的表現，最好的是 TFIDF+XGBoost 的組合，準確率為 0.69。

	TFIDF	Word2Vec	BERT
KNN	0.4369	0.2697	0.2571
Naïve Bayes	0.2824	-	-
SVM	0.6847	0.3061	0.3378
Random Forest	0.4876	0.3159	0.2980
XGBoost	0.6916	0.3308	0.3314

表二、16 型分類結果



圖二、各項分類預測結果分布

Hernandez, R. 等人[6]在其研究中提到，MBTI 主要以四個二分類組成，分別為 E/I, N/S, F/T, J/P，故該研究建立了四個二元分類器，針對四個二分類進行預測，而非以一個多類別分類器預測 16 型人格。本專案也參考該研究作法，除了針對 16 型人格，也嘗試用四個分類器分別對 E/I, N/S, F/T, J/P 進行預測，整體的準確率由四個 classifier 預測的值組起來跟原本的 label 比對來進行計算，要剛好四個 classifier 的輸出都是正確的才有辦法對整體準確率有貢獻。本專案嘗試過後，發現大部分都是直接以 16 分類表現較佳。

TFIDF	E/I	N/S	F/T	J/P	對照原本label		原16型分類結果
XGBoost	0.8664	0.9240	0.8606	0.7961	0.5979	<	0.6916
SVM	0.8629	0.9217	0.8698	0.7915	0.5910	<	0.6847
KNN	0.8376	0.8940	0.6947	0.7304	0.4366	<	0.4548
Random Forest	0.7995	0.8790	0.7972	0.7166	0.4078	<	0.4876
NB	0.7984	0.8790	0.7903	0.6429	0.3399	>	0.2824

表三、4 個二分類分類器 / TFIDF

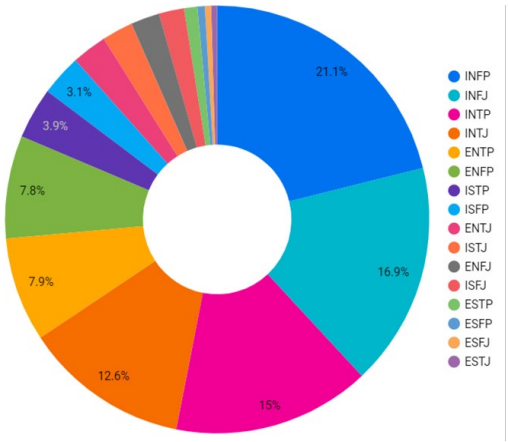
Word2Vec	E/I	N/S	F/T	J/P	對照原本label		原16型分類結果
SVM	0.7972	0.8779	0.7592	0.6106	0.3041	<	0.3061
XGBoost	0.7938	0.8767	0.7465	0.6221	0.3007	<	0.3308
Random Forest	0.7903	0.8744	0.7304	0.7166	0.2972	<	0.3159
KNN	0.7915	0.8767	0.6947	0.5922	0.2788	>	0.2697

表四、4 個二分類分類器 / Word2Vec

BERT	E/I	N/S	F/T	J/P	對照原本label		原16型分類結果
SVM	0.7984	0.8744	0.7465	0.6774	0.3641	>	0.3378
XGBoost	0.8099	0.8767	0.7074	0.6636	0.3237	<	0.3314
Random Forest	0.8007	0.8767	0.6993	0.6452	0.3180	>	0.2980
KNN	0.7926	0.8779	0.6613	0.6164	0.2995	>	0.2571

表五、4 個二分類分類器 / BERT

2.4 嘗試提升分類準確率



圖三、各類型人格的資料不平衡

由於各類型人格的資料不平衡，所以在本專案中，針對不平衡資料問題進行處理，嘗試提升準確率時。

在本專案中，首先是調整模型的 sample weight 參數，讓模型依照平衡的資料進行訓練，嘗試提升模型準確率。調整完參數後，各模型準確率結果如下：

	TFIDF		Word2Vec		BERT	
	SW 前	SW 後	SW 前	SW 後	SW 前	SW 後
KNN	0.4369	0.4369	0.2697	0.2697	0.2571	0.2571
Naïve Bayes	0.2824	0.5107 ↑	-	-	-	-
SVM	0.6847	0.6767 ↓	0.3061	0.2017 ↓	0.3378	0.3130 ↓
Random Forest	0.4876	0.6104 ↑	0.3159	0.2893 ↓	0.2980	0.2571 ↓
XGBoost	0.6916	0.6611 ↓	0.3308	0.2553 ↓	0.3314	0.2697 ↓

表六、針對 16 型分類，調整模型 Sample Weight 參數

根據表六結果，可以發現在使用 TFIDF、Word2Vec 和 BERT 表示法的情況下，各個模型準確率都沒有超過調整 sample weight 參數前的最高準確率 69 %。

接著是 4 個二分類分類器進行預測的部分，在使用 TFIDF 表示法情況下，同樣調整各模型的 sample weight 參數，測試後準確率結果如下：

TFIDF		E/I	N/S	F/T	J/P	對照原本 label		原16型分類結果
XGBoost	SW 前	0.8664	0.9240	0.8606	0.7961	0.5979	<	0.6916
	SW 後	0.8341	0.8698	0.8456	0.7892	0.5438		
SVM	SW 前	0.8629	0.9217	0.8698	0.7915	0.5910	<	0.6847
	SW 後	0.8537	0.8882	0.8583	0.7615	0.5484		
KNN	SW 前	0.8376	0.8940	0.6947	0.7304	0.4366	<	0.4548
	SW 後	0.8364	0.8952	0.6993	0.7281	0.4286		
Random Forest	SW 前	0.7995	0.8790	0.7972	0.7166	0.4078	<	0.4876
	SW 後	0.8065	0.8790	0.8180	0.7293	0.4343		
NB	SW 前	0.7984	0.8790	0.7903	0.6429	0.3399	>	0.2824
	SW 後	0.7569	0.8237	0.8341	0.7085	0.3986		

表七、針對 4 個二分類分類器，
調整模型 Sample Weight 參數 (TFIDF 表示法)

如表七所示，在使用 TFIDF 向表示法的情況下，Random Forest 和 Naive Bayes 準確率有顯著的提升，但大部分模型的分類準確率仍然比原先直接分類 16 型人格的結果還來的低。

接著在使用 Word2Vec 和 BERT 表示法的情況下，同樣調整分類模型的 sample weight 參數，測試後準確率結果如下：

Word2Vec		E/I	N/S	F/T	J/P	對照原本 label		原16型 分類結果
SVM	SW 前	0.7972	0.8779	0.7592	0.6106	0.3041	<	0.3061
	SW 後	0.6521	0.6578	0.7753	0.6118	0.2350		
XGBoost	SW 前	0.7938	0.8767	0.7465	0.6221	0.3007	<	0.3308
	SW 後	0.6889	0.7339	0.7500	0.6060	0.2431		
Random Forest	SW 前	0.7903	0.8744	0.7304	0.7166	0.2972	<	0.3159
	SW 後	0.7938	0.8779	0.7408	0.6187	0.3018		
KNN	SW 前	0.7915	0.8767	0.6947	0.5922	0.2788	>	0.2697
	SW 後	0.7915	0.8767	0.6947	0.5922	0.2788		

表八、針對 4 個二分類分類器，
調整模型 Sample Weight 參數（Word2Vec 表示法）

BERT		E/I	N/S	F/T	J/P	對照原本 label		原16型 分類結果
SVM	SW 前	0.7984	0.8744	0.7465	0.6774	0.3641	<	0.3378
	SW 後	0.6959	0.7085	0.7535	0.6717	0.2650		
XGBoost	SW 前	0.8099	0.8767	0.7074	0.6636	0.3237	<	0.3314
	SW 後	0.7984	0.8779	0.6912	0.6498	0.3053		
Random Forest	SW 前	0.8007	0.8767	0.6993	0.6452	0.3180	>	0.2980
	SW 後	0.7131	0.7512	0.7062	0.6302	0.2477	<	
KNN	SW 前	0.7926	0.8779	0.6613	0.6164	0.2995	>	0.2571
	SW 後	0.7926	0.8779	0.6613	0.6164	0.2995		

表九、針對 4 個二分類分類器，
調整模型 Sample Weight 參數（BERT 表示法）

根據表八、表九，可以發現大部份的分類模型準確率有下降的情況，準確率並未提升。

除了調整模型的 sample weight 參數，本專案也嘗試使用 chi-square 進行特徵選擇，分別挑選 2000 個特徵和 5000 個特徵進行測試，各模型進行 16 型人格預測的準確率結果整理如下：

TFIDF	原始	Sample Weight	Feature Selection		Sample Weight + Feature Selection	
			2000 特徵	5000 特徵	2000 特徵	5000 特徵
KNN	0.4369	0.4369	0.3787	0.3919	0.3787	0.3919
Naïve Bayes	0.2824	0.5107	0.3614	0.3487	0.5516	0.5401
SVM	0.6847	0.6767	0.6663	0.6680	0.6761	0.6795
Random Forest	0.4876	0.6104	0.6190	0.5781	0.6473	0.6421
XGBoost	0.6916	0.6611	0.6888	0.6865	0.6646	0.6559

表十、特徵選擇

從表十結果來看，與未特徵選擇相比，進行特徵選擇後，SVM 和 XGBoost 模型表現沒有太大差異，而 Naive Bayes 和 Random Forest 模型準確率則有些微提升，KNN 模型準確率則有所下降。

除了上述作法，本專案也參考了 Noureen Fatima 等人的研究[7]，該研究在資料處理階段使用了 SMOTE 進行資料平衡處理，以預測 MBTI 人格，因此我們也嘗試了 SMOTE 資料平衡方法。不過，經過 SMOTE 資料處理後，同樣發現大部分模型的準確率有所下降。另外，本專案也採用集成模型的方式進行測試，使用了 XGBoost、Random Forest、SVM 三個模型進行訓練並預測，準確率落在 64 % 左右。

在嘗試以上不同方式後，本專案發現在使用 TFIDF 表示法和 XGBoost 模型的情況下，分類表現最好，準確率為 69 %，所以本專案在後續會使用這個模型進行相關的應用分析。

3 SYSTEM OUTCOMES

本專案採用 kaggle—LinkedIn Influencers' Data[8]，作為模型後續應用。LinkedIn Influencers' Data 是一份 linkedin 上排名前 69 位影響力人物的資料集，這份資料中有許多不同領域的知名人物，像是維珍集團主席、劍橋大學院長、紐約市長候選人等。資料集共有 34520 筆資料、18 個欄位，經過初步刪減後，保留 3 個欄位，分別為 Name、headline、content。

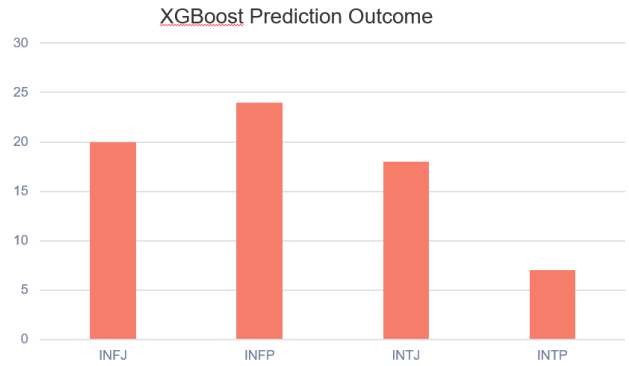
LinkedIn Influencers' Data 保留欄位			
	英文變數名稱	中文變數名稱	型態
1	Name	使用者姓名	String
2	headline	Linkedin 個人資料標題（內含其職稱＆企業單位）	String
3	content	貼文內容向量	String

表十一、LinkedIn Influencers' Data 資料集保留之欄位

我們從他們的個人資料中擷取職涯關鍵字，之後對這些人在 LinkedIn 的貼文進行前處理，方法和前面訓練資料相同，之後將貼文以 TFIDF 向量表示法丟入 XGBoost 模型，以預測他們的 MBTI 人格類型。

預測結果僅得出四種 MBTI 類型，且皆以 IN 作為開頭，於是本專案進一步對 IN 個性本質進行探討。研究後發現，I 指的是內向型人格，這類型的人往往以思想為導向，比較享

受深度有意義的社交，並在後續的自我獨處獲得能量、傾向於三思而後行，且興趣廣泛；N 指的是直覺型人格，這類型的人在思考上較為抽象，注重未來的可能性、喜歡尋求有創意的想法，能夠放眼大局。[13]



圖四、預測結果分布

對於大眾印象而言，能夠成為有影響力的人，並且於社群媒體上活躍，或許應為 E 型(外向型人格)較多[14]，與此專案結果之 I 型相反。然，此資料為 69 名 LinkedIn 影響力排名人物，這些人通常是在各領域有所成就的人士，並且在業界曉諭名門，若有在經營 LinkedIn，便容易登上影響力排行榜。因此這些人的共同特質是在業界具有領導力與前瞻力，並不一定是印象中外向的人。由對 IN 型人格的分析，我們可以初步了解甚麼樣的人可以成為各界巨擘。

Hernandez, R.等人[6]在其研究中，使用四個二元分類器，預測貼文的四個二分類(E/I, N/S, F/T, J/P)後，依據所預測出的四個二分類將貼文進行整理，使用貼文中的文字建立文字雲，以觀看八個類別的貼文文字出現頻率。本專案參考了該研究，並採取不同的文字雲分析作法。

本專案預測完 69 位影響力人物的 MBTI 人格（共四個，INFJ、INFP、INTJ、INTP）後，依照這四個 MBTI 人格將 69 位人物的職涯關鍵字進行彙整，並建立文字雲，以觀察這四個 MBTI 類型的影響力人物職涯關鍵字分布狀況，並且本專案蒐集網路上對於這四個 MBTI 類型的人適合從事哪些職業的資料進行比對。經過比對後發現，約有四成的影響力人物，其職涯關鍵字與其 MBTI 人格合適職業吻合，若再加上如 writer、author 這樣相同職位、用字不同的職業及相同領域的各種管理位階，相信比對吻合的程度會更高。

此外，從肉眼比對這四種 MBTI 類型的人的職業關鍵字文字雲，本專案也發現有許多職業出現多次，表示相同類型的人確實對職業選擇有共通點，以下將對這四種 MBTI 類型人格的個性及職業做解析。

INFJ 類型[9][15]

INFJ 類型的人被稱為諮詢師或提倡者。他們喜歡在幕後施展其影響力，可對應到文字雲中的「Director」。他們擅長於理解複雜的問題，並且有高度洞察力，在現實生活中解決問題常常是他們的人生目標，可對應文字雲中的「Entrepreneur(創業家)、Marketing(行銷)」。對 INFJ 最有價值的工作是讓他們在幫助他人的同時成長為一個人，因此也可看到文字雲出現「Mental(心理)、Managing(管理)」。任何妨礙這些價值觀的事情，像繁文縟節、毫無意義的規則、官僚體制，和不道德的同事都會嚴重削弱倡導者的積極性，下方文字雲中呈現職業確實較無辦公室的工作。

圖五、INFJ 類型之影響力人物職業關鍵字文字雲

INFP 類型[10][16]

INFP 又稱為調停者，對於調停者來說，理想的職業道路應該像是一種召喚，而不僅僅是一份工作。他們具有創造力，通常在語言上有天賦，喜歡通過寫作來表達自己，因此在文字雲中的「Author(作家)、Media(媒體)」占比特別大。維珍集團主席 Richard Branson 於我們的分類結果中即屬此類，我們認為十分符合他的個性，喜歡更多的變化、做很多事情，有強烈的自我認知和對企業的認知。

圖六、INFP 類型之影響力人物職業關鍵字文字雲

INTJ 類型[11][17]

INTJ 又稱為建築師，在工作場所，建築師通常以能力和效率著稱，將理性分析和辛勤工作相結合視為成功的基礎，是天生的領導者，可對應文字雲中的「Founder、CoFounder」。他們也具備分析和簡潔表達複雜概念的才華，可對應文字雲中的「Speaker(演說家)、Consultant(顧問)」，此類型工作需要有化繁為簡，讓人輕易抓到重點的口說能力。

圖七、INTJ 類型之影響力人物職業關鍵字文字雲

INTP 類型[12][18]

INTP 又稱為邏輯學家，具有這種性格類型的人渴望智力刺激、追求自己想法的自由以及解決具有挑戰性難題的機會。他們通常是最有邏輯思維的人格類型，擅長解決問題，可對應文字雲中的「Technology(科技)、Educator(教育家)」。他們也擅長計劃和製定戰略，可對應文字雲中的「Business(商業)、Investment(投資)」，在金融業的挑戰，正好符合他們喜歡突破自我的個性。 另外，此類型的人不擅長辦公室中重複的工作內容，雖然分於此類者較少，但於他們的就業內容中並未見基層工作。

圖八、INTP 類型之影響力人物職業關鍵字文字雲

本專案所選資料中的人物因為各界巨擘，因此常同時有多種職業，例如身兼公司執行人與相關書籍作家，因此在職業的分類上較模糊，造成文字雲中有許多職業重複出現的現象。然而這也顯示了在現實生活中職業的選擇並非單一的，一個人可以身兼許多副業，在各種情況下都能找到符合自身個性且擅長的事，展現人格的多元面貌。此外，此資料及缺乏普通人的就業情況，因此也不能代表所有相同人格類型的人的就業情況與領域，例如 INFJ 也可能受其他誘因或外力而從事辦公室的職業，

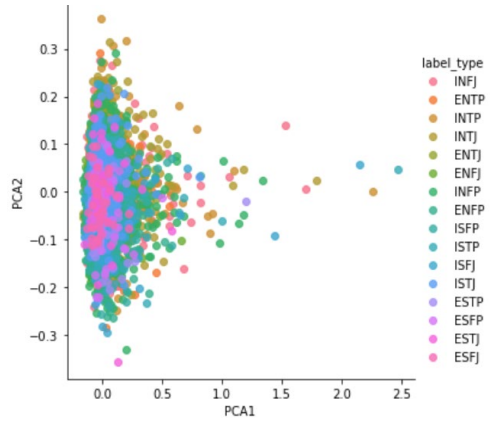
就算不符合自己的個性，也可能在各種情況下找到自己的容身之處。

4 CONCLUSIONS

本專案中 Word2Vec 和 BERT 表現差的原因可能為：

1. 採用的 pre-trained model 皆是基於新聞或是較正式的文本所訓練的，因此套用在社群論壇貼文上，可能有很多詞不在這些 pre-trained model 的字典內，導致不足以產生具有代表性的文本向量。
2. Word2Vec 為詞向量，而採用所有詞向量平均作為 document representation 的方法有可能造成最後的文本向量喪失很多資訊。
3. BERT base model 最多只能處理 512 個 token，此資料集的貼文每則約 2000-4000 個 token，可能過程中丟失過多訊息。

若資料本身乾淨及群和群之間分布明確，我們預期 PCA 降維結果會是相近人格應互相靠近，不過 PCA 結果看到所有人格特質混雜在一起，這可能是因為資料本身的品質問題或是人格特質使用單純貼文來代表較難有明確分別。



圖九、PCA 降維結果

總結來說，本專案在使用五種不同分類模型下，預測準確率最高為 XGBoost 的 69 %。從結果來看，僅使用基本模型而沒有對模型進行調參或許是導致結果不太理想的原因，未來可以考慮使用像是 GridSearch 等方法幫助調整參數，提供幾組模型及參數的候選組合進行嘗試，找出有最好表現的組合來進行訓練。

藉由本專案的研究與相關分析，能建立出分類表現較好的人格特質預測模型，補足過往文獻較少探討的研究內容，為相關領域的研究發展帶來貢獻價值。透過後續取用 LinkedIn 影響力知名人物資料，可發現有相似職涯及成就之人物，以模型預測之人格類型相似，由此可驗證人格及職業之關聯性，以提供學生及求職者作為參考。

5 REFERENCES

[1] Reddit: A Gold Mine for Personality Prediction

[2] 利用語言特徵和表情符號提升九型人格分類效率

[3] 運用文字探勘技術在社群行為上之人格預測

[4] (MBTI) Myers-Briggs Personality Type Dataset | Kaggle

[5] Cui, B.; Qi, C. Survey Analysis of Machine Learning Methods for Natural Language Processing for MBTI Personality Type Prediction. (2017)

[6] Hernandez, R.; Knight, I.S. Predicting Myers-Bridge Type Indicator with text classification. In Proceedings of the 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017

[7] Fatima, Noureen et al., A rule-based machine learning model for career selection through MBTI personality, Research Journal Of Engineering & Technology, Vol 4, No 2, page 185-196, 2022

[8] LinkedIn Influencers' Data | Kaggle

[9] The Best Careers for INFJ Personality Types | Truity

[10] The Best Careers for INFP Personality Types | Truity

[11] The Best Careers for INTJ Personality Types | Truity

[12] The Best Careers for INTP Personality Types | Truity

[13]「MBTI」是在測什麼？4 種維度 8 面向，16 型人格測試看出你的人格特質 | 職場熱議 | 104 人力銀行

[14] Which Personality Types Make Great Influencers? | 16 Personalities

[15] INFJ Workplace Habits | 16 Personalities

[16] INFP Workplace Habits | 16 Personalities

[17] INTJ Workplace Habits | 16 Personalities

[18] INTP Workplace Habits | 16 Personalities