

分析社群貼文內容與發文者的 MBTI 相關性

張書明
B08208010

藍知生
B08701163

吳郁心
R12725043

林姝延
R12723006

黃崢霖
B08701103

| 報告大綱：

分析貼文內容，判斷貼文者之 MBTI。

| 研究動機：

MBTI 是一項性格評量，每位受測者會從 4 個指標中，得到其中一種傾向，組成特定性格。指標分別為：

1. 外向型 (Extraversion)、內向型 (Introversion)
 2. 實感型 (Sensing)、直覺型 (Intuition)
 3. 理性型 (Thinking)、感性型 (Feeling)
 4. 系統型 (Judging)、彈性型 (Perceiving)
- ，共有 16 種組合。

近年 MBTI 逐漸受到重視，在社群掀起風潮。其用途也變得多元，除個人想了解自己的性格外，目前多家企業也開始在選才過程將 MBTI 納入考量，注重適才與多元性。本研究認為僅透過測驗方出題或受測方填答均可能出現偏誤，若能從日常行徑歸納出受測方性格將提高準確度及可參考性。因此希望先以資料集作為樣本嘗試歸納出 MBTI 不同性格，在社群留言擁有的表現是否具有鑑別度。

| 資料來源及前處理：

為豐富本研究的樣本多元性，本組自 Kaggle 及 Zenodo 兩資料庫中分別下載 PersonalityCafe forum、Twitter、Reddit 三種社群平台的 MBTI 資料集。以下將分述此三種資料集的形式。

1. PersonalityCafe forum (以下簡稱 PCF)
下載自 [Kaggle](#)，共兩個欄位，含有 8,675 筆資料。Type 欄位為此貼文者的 MBTI；Post 欄位則為此貼文者發布之多個貼文，每筆貼文以 ||| 區分。

Type	Post
INFJ	'http://www.youtube.com/watch?v=qsXHewe3krw http://41.media.tumblr.com/tumblr_lfouy03PMA1qa1roo0l...'
ENTP	T'm finding the lack of me in these posts very alarming. Sex can be boring if it's in the same po...

表一、PCF 資料集範例

2. Twitter

下載自 [Kaggle](#)，共三個欄位，含有 7,581 筆資料。Index 欄位為每筆資料的編號；Type 欄位

為此貼文者的 MBTI；Text 欄位則為貼文者發布之多個貼文。每筆貼文以 ||| 區分。

index	text	label
0	@Pericles216 @HierBeforeTheAC @Sachinettiyil The Pope is infallible, this is a catholic dogma It doesn't, Â mean the, Â https://t.co/qmt0ezk0Ey 	intj
1	@Hispanthicckk Being you makes you look cute @ThiccWhiteDuke_ On, because then I can have the fun of peeling it off... @whedonesque "Bored now." 	intj

表二、Twitter 資料集範例

3. Reddit

下載自 [Zenodo](#)，共三個欄位，含有 1,128,420 筆資料。author_flair_text 欄位反映的是貼文者的 MBTI；body 欄位則為此貼文者發布之多個貼文，每筆貼文以 ”” 區分；subreddit 欄位則顯示貼文所在的特定 Reddit 子版面或討論主題。

author_flair_text	body	subreddit
ENTJ	"Well consider this, Freud is considered the father of psychoanalysis. His ""findings"" were not exactly reliable. But here we are on an MBTI subreddit, something slightly more reliable that followed something less reliable. So learning from Reddit could simply just be a source of inspiration to put together a slightly more reliable picture of the people around us."	intj
INFJ/4w3	"Wow, i might be using a few of these starting this week....thanks!"	infj

表三、Reddit 資料集範例

清理資料時，本組會將各貼文者的多筆貼文拆分，並以相同欄位模式儲存在新的一列，再將各個平台內貼文者的 MBTI 拆分成 4 個面向，並新增 4 個欄位以儲存 E/I、S/N、T/F、J/P 的類型，方便後續針對各個指標進行分析。另外，為避免判斷結果受貼文內的無意義訊息干擾，將貼文內容中的網址、短於三個詞的貼文刪除，並將各貼文的內容還原詞形

(Lemmatization)。最後合併三個社群平台的資料集，再新增一個欄位以標示該筆資料出自何種社群平台，再將資料向量化，以便後續使用在資料輪廓描述及模型預測。因此資料集在清理及合併後的外觀如下，共有七個欄位，隨機挑選 3% 的資料分析，共 63,498 筆。

type	posts	E/I	S/N	T/F	J/P	media
INFP	@ punyulada ITS RGB GAMING LIGHTS	I	N	F	P	twitter
INFJ	If you 're not really go for a programmer job ...	I	N	F	J	PCF
ISFP	I wan na be a mom too . Sometimes , when Im on...	I	S	F	P	PCF
INTJ	Ditto ! I always imagine this mind 's eye thin...	I	N	T	J	reddit
ENFP	DID I MISS THE POST WEAR SOMEONE REFERENCED RU...	E	N	F	P	PCF

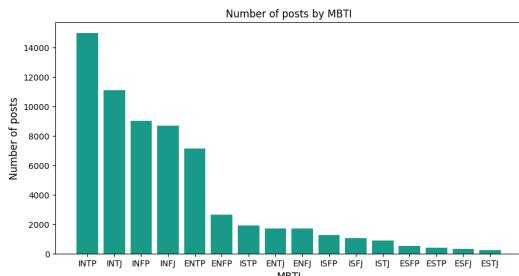
表四、隨機挑選後的資料集範例

| 資料輪廓描述：

以下將描述隨機挑選 3% 後的資料輪廓。

1. MBTI 在整體資料集的分佈

在整體資料集中，INTP（23.6%）的貼文數最多，其人格特質為安靜內斂下蘊含著豐富的想法與洞察，傾向於獨處思考、深入探究新的知識和理論以及分析事物的運作方式和解決辦法等，因此推論其在網路上抒發己見的可能性較高，導致資料集中 INTP 的貼文數最多。



圖一、資料集中各 MBTI 類型的貼文數

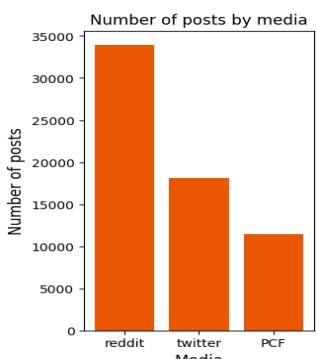
而從圖一可看到貼文數前四多的 MBTI 均為 IN 開頭的人格（69%），與圖二顯示的狀況相符，由 I、N 型佔資料集的大宗（分別佔比為 76.9% 及 89.6%）。



圖二、資料集中 MBTI 四大維度的佔比

2. 三大社群平台的差異

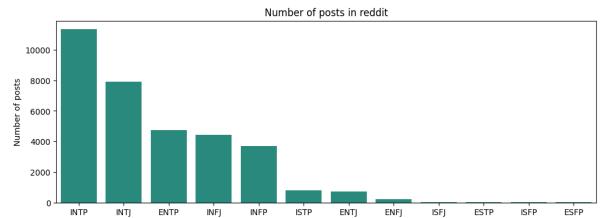
隨機挑選後的資料集中，來自三個社群平台的貼文佔比分別為 Reddit（53.4%）、Twitter（28.6%）、PCF（18%）。



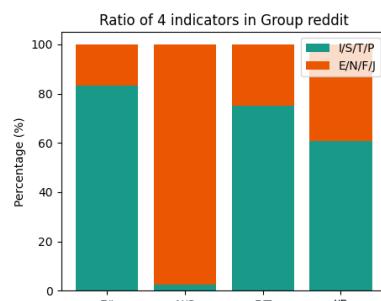
圖三、資料集中各社群平台的貼文數

a. Reddit

因數據來源主要來自 Reddit，整體數據受 Reddit 的貼文影響較大。與另兩種社群相比，可發現 Reddit 的貼文中 MBTI 性格分佈較為集中，貼文來自 INTP、INTJ 等人格的比例較高，與整體資料集的表現相似。



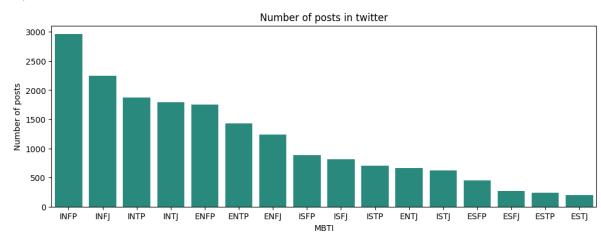
圖四、Reddit 中各 MBTI 的發文數



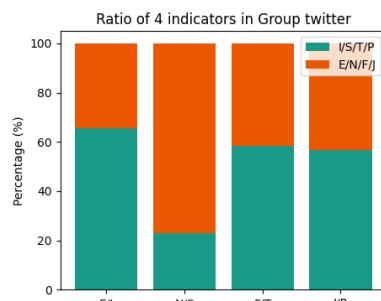
圖五、Reddit 中 MBTI 四大維度的佔比

b. Twitter

在 Twitter 上發布貼文者的 MBTI 組成則較多元，貼文數最多的為 INFP、INFJ。其人格特質為重視人際和諧、情感交流，擁有自己價值觀且不輕易妥協、情緒豐沛而敏感，特別著迷於音樂、藝術、文學等富含創造力與情感的事物。



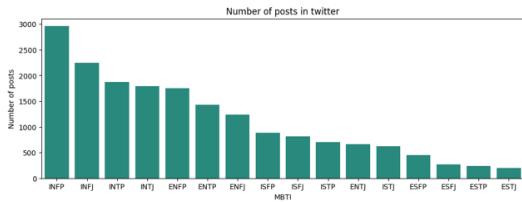
圖六、Twitter 中各 MBTI 的發文數



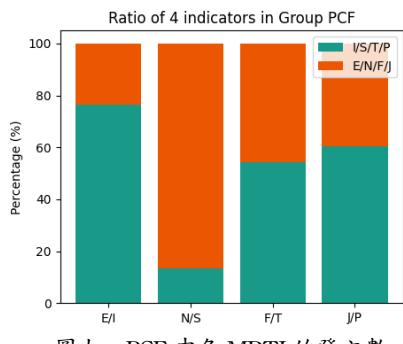
圖七、Twitter 中各 MBTI 的發文數

c. PCF

PCF 的貼文者則與 Twitter 的貼文者的 MBTI 分佈較相近，最高者同樣為重視人際和諧、情感交流的 INFP 以及 INFJ 人格。



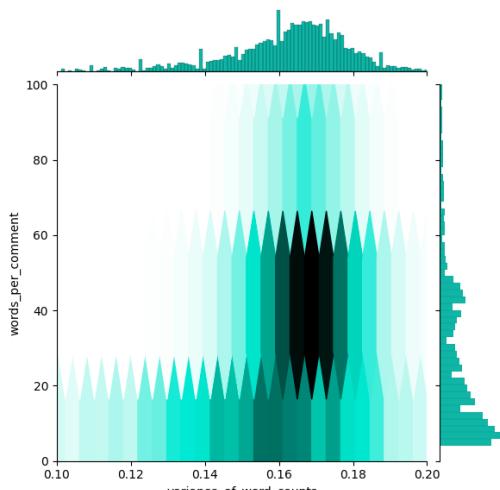
圖八、PCF 中各 MBTI 的發文數



圖九、PCF 中各 MBTI 的發文數

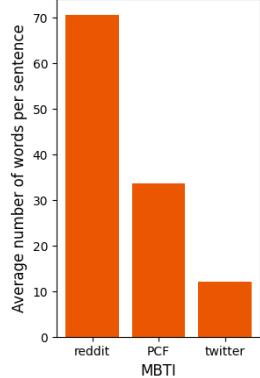
3. 用字習慣

從下圖可發現貼文的字數與變異數較無相關性。而 Reddit 具有最高的平均句長（近 70 字）。INTJ 平均用字數量則最多（近 60 字），可推測這與 INTJ 富有想像力、喜歡獨立思考、目標導向等性格特徵有關，其是 MBTI 16 型人格中極稀有且公認最具才華的類型，佔全球男性人口 1.2%、女性人口更僅有 0.8%。INTJ 名人包括 Tesla 創辦人 Elon Musk、Facebook 創辦人 Mark Zuckerberg。

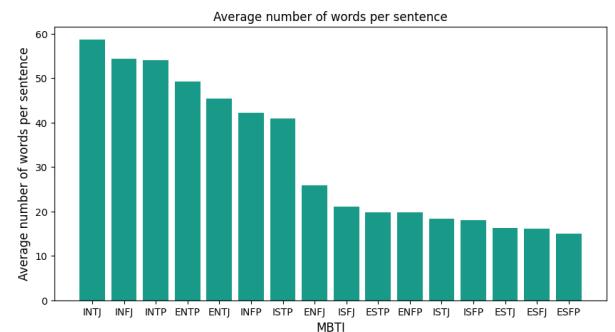


圖十、字數與變異數的相關性

Average number of words per sentence



圖十一、各社群平台的貼文平均句長



圖十二、各社群平台的貼文平均句長

4. 文字雲

以下將從三個面向觀察不同人格以及社群平台中貼文者使用頻率較高的字眼。為避免停用詞 (stopwords) 千擾文字雲的繪製，本組定義 stopwords 為 NLTK 套件中的 stopwords、資料集中出現頻率前五十大的字詞以及'ENFJ', 'ENFP', 'ENTJ', 'ENTP', 'ESFJ', 'ESFP', 'ESTJ', 'ESTP', 'INFJ', 'INFP', 'INTJ', 'INTP', 'ISFJ', 'ISFP', 'ISTJ', 'ISTP', 'one', 'u', 're', 'know', 'thing', 'r', 'e', 'would', 'im', 'make' 作為清理資料集的 stopwords。

```
# Top 50 Frequency words
words = list(df['posts'].apply(lambda x: x.split()))
words = [x for y in words for x in y]
common_words = Counter(words).most_common(50)

# NLTK stopwords
nltk.download('stopwords')
stopwords_set = set(stopwords.words('english'))

# Expand stopwords
for word, _ in common_words:
    stopwords_set.add(word.lower())

morewords = ['ENFJ', 'ENFP', 'ENTJ', 'ENTP', 'ESFJ', 'ESFP', 'ESTJ', 'ESTP', 'INFJ', 'INFP', 'INTJ', 'INTP', 'ISFJ', 'ISFP', 'ISTJ', 'ISTP', 'one', 'u', 're', 'know', 'thing', 'r', 'e', 'would', 'im', 'make']

for word in morewords:
    stopwords_set.add(word.lower())

# Function to remove stopwords
def remove_stopwords(text):
    return ' '.join([word for word in text.split() if word.lower() not in stopwords_set])

# Clean data
df['posts_cleaned'] = df['posts'].apply(remove_stopwords)
df['posts_cleaned'] = df['posts_cleaned'].head()
```

a. 16 型人格

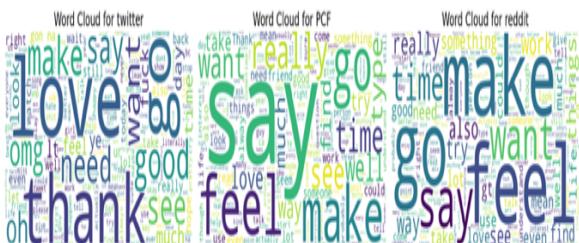
觀察以下文字雲可發現，MBTI 為 ENFP 和 ENFJ 的貼文者使用 “love” 和 “feel” 的頻率較高，這可能表示這些類型的人在情感表達方面特別外顯。

而在 ISTJ 和 INTJ 的文字雲中，“work”和“need”的頻率較高，這可能表示這些類型的人較為務實。另外 INFP 和 ISFP 的發文者則較常使用“feel”和“time”等字詞，這可能代表這些類型的人在情感表達較為內斂。



b. 三大社群平台

從左到右分別為 Twitter、PCF、Reddit 等三大社群平台的文字雲。在 Twitter 的文字雲中，“love”、“go”、“thank”的出現頻率較高，這可能表示 Twitter 用戶在表達情感方面特別活躍。而 PCF 的文字雲，“say”、“feel”、“want”的頻率較高，這可能意味著 PCF 的用戶對於情感表達較為內斂。另外 Reddit 的文字雲中，“feel”、“time”、“try”的頻率較高，這可能反映了 Reddit 用戶的行動力較高。



c. 16 型人格的四個維度

E 型人格的文字雲中，動詞如“go”、“say”和“make”的頻率較高，這可能表明外向型人格傾向於實際行動和直接溝通的方式。I 型人格的文字雲中，“real”、“learn”、“say”的頻率較高，可能顯示內向型人格更偏向於思考和學習。

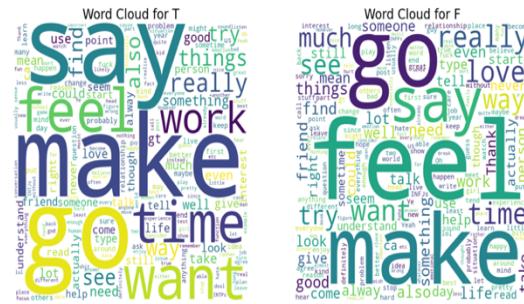


S 型人格的文字雲中，“love”、“good”、“make”突出，這可能意味著感覺型人格重視實際的人際關係。而 N 型人格的文字雲中，“say”、“really”和“feel”較大，直覺型人格可能更偏好抽象的概念和情感表達。



T 型人格的文字雲中，詞彙如“make”、

“say”和“want”的頻率較高，這可能表示思考型的人在溝通時較為直接且著重於行動和目標。F 型人格的文字雲中，“feel”、“love”和“really”很突出，這可能顯示感覺型的人在表達情感和個人價值觀方面較為豐富。



J 型人格的文字雲中，“want”、“make”和“way”相對較大，這可能意味著判斷型的人傾向於有計劃和組織性的溝通。P 型人格的文字雲中，“want”、“say”和“things”較顯著，這可能代表感知型的人在探索和開放性的對話中較為活躍。



| 研究方法：

1. 分析簡介

進行 MBTI 預測：依情緒分類、特徵詞彙、特徵語氣建立模型，預測貼文者的 MBTI，分為以下兩大部分：

a. 第一部分：文章十六型人格分類

- i. 貼文 MBTI 分類：根據特徵詞彙、語氣，以模型將文章區分為 MBTI 的 16 種類別。
- ii. 模型評估：使用 Precision, Recall 等指標評估模型。
- b. 第二部分：文章 Binary 分類
- iii. 貼文 MBTI 分類：根據特徵詞彙、語氣，以模型區分 I-E, N-S, T-F, J-P。
- iv. 模型評估：使用 Precision, Recall 等指標評估模型。

2. 模型及訓練與測試資料集

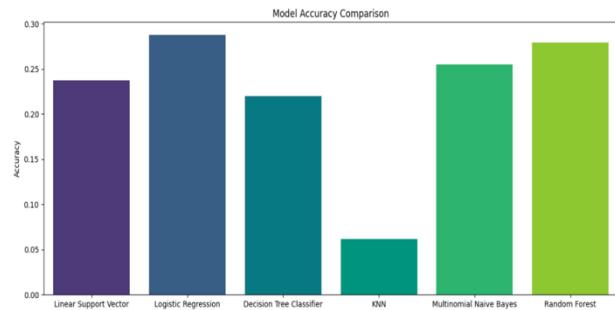
本組使用以下模型進行訓練及預測，並將原始資料 90% 設定為訓練資料，剩餘 10% 則為預測資料集。

- a. Linear Support Vector Classifier
- b. Logistic Regression
- c. Decision Tree Classifier
- d. KNN
- e. Multinomial Naive Bayes
- f. Random Forest Classifier

| 結果呈現：

1. 以 Accuracy 衡量模型對於 16 型人格的預測結果，各模型結果如下：

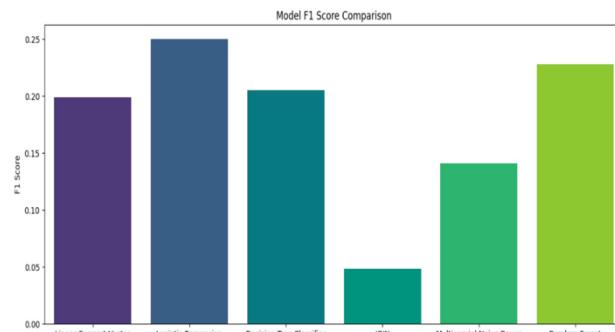
- a. Linear Support Vector Classifier: 0.237
- b. Logistic Regression: 0.287
- c. Decision Tree Classifier: 0.219
- d. KNN: 0.062
- e. Multinomial Naive Bayes: 0.255
- f. Random Forest Classifier : 0.279



Regression 為表現最好的模型，其準確率不及 0.3，而 KNN 模型表現最差，準確率僅 0.0618。

2. 以 F1 Score 衡量模型對於 16 型人格的預測結果，各模型結果如下：

- a. Linear Support Vector Classifier: 0.199
- b. Logistic Regression: 0.250
- c. Decision Tree Classifier: 0.205
- d. KNN: 0.048
- e. Multinomial Naive Bayes: 0.141
- f. Random Forest Classifier : 0.228

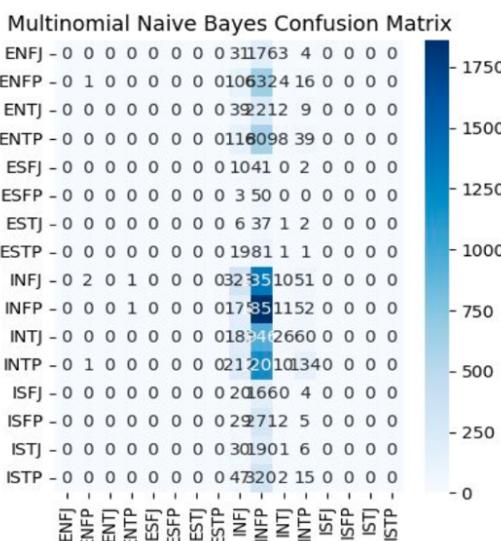
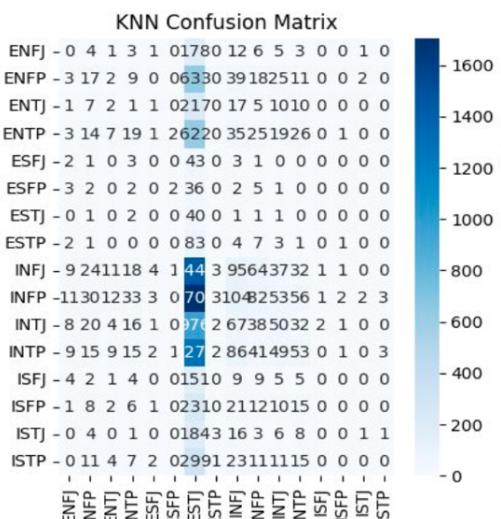
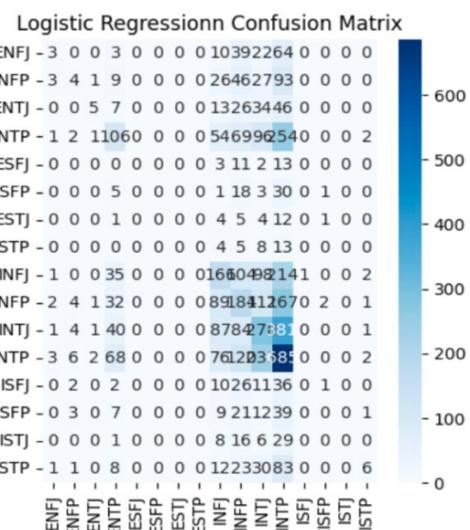
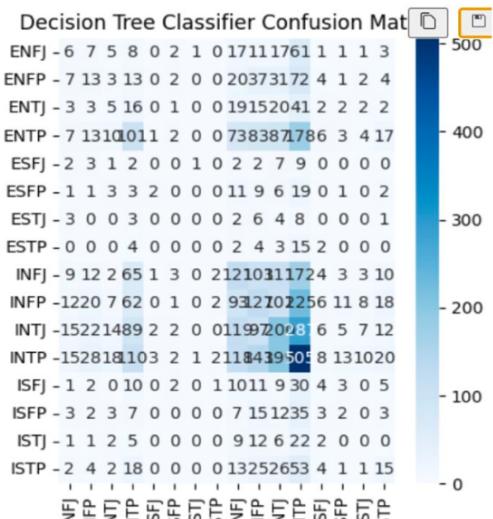
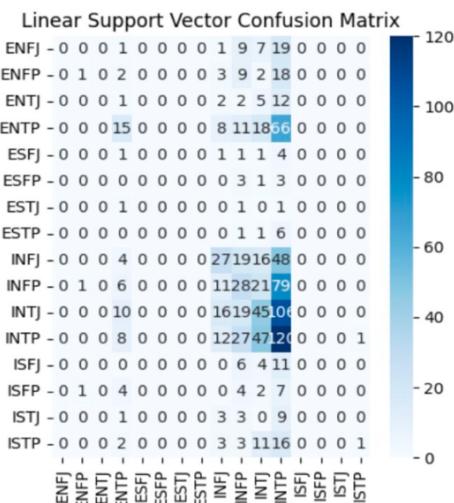


同樣可發現 KNN 模型表現最差，Logistic Regression 表現最佳，而所有模型 F1 Score 均沒有超過 0.25。

推測可能原因如下：

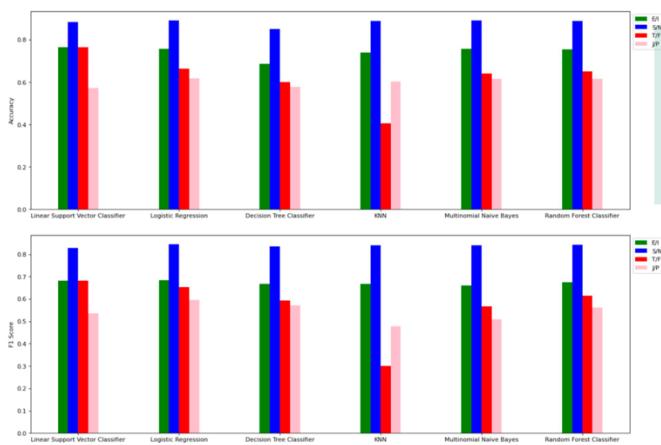
- a. 分類過細：總共區分 16 種分類，且每種分類間無非常明顯差異，導致模型較難進行準確預測。
- b. 主觀因素：MBTI 分類來自個人測驗結果，受主觀因素和個體解釋影響。
- c. 樣本不平均：模型可能會偏向於預測樣本數較多的類型，而樣本較少的類型預測效果較差。

- d. 文本多樣性：社群文本涵蓋各主題，有些用戶的言論風格可能不夠典型，使模型難以捕捉 MBTI 特徵。
- e. 數據噪音：社群媒體文本數據可能含大量噪音，雖進行過前處理，仍可能包含簡寫、拼寫錯誤、非正式用語等，使模型難準確捕捉 MBTI 特徵。
3. 觀察各模型混淆矩陣，可發現多數模型將大量樣本歸類於 INTP，或 IN 相關類別，即原始數據集中樣本數量佔比最高的族群，因此模型此預測傾向合理，唯獨 KNN 模型的預測結果中多數為 ESTJ 類型，使得其預測效能與其他模型相比略遜一籌。



Random Forest Classifier Confusion Matrix															
ENFJ	-12	4	3	2	2	0	0	0	40	109	1521	0	3	2	1
ENFP	-5	79	2	16	1	0	0	0	13	36	15787	3	3	2	6
ENTJ	-1	3	9	7	0	0	0	0	45	120	8943	2	0	0	2
ENTP	-4	23	5	78	1	0	0	0	1382	296	1133	2	5	7	
ESFJ	-0	1	0	3	1	0	0	0	17	18	3	9	0	0	1
ESFP	-0	0	0	4	0	1	0	0	14	27	3	3	0	1	0
ESTJ	-0	3	0	3	0	0	1	0	13	20	3	2	0	0	1
ESTP	-1	4	1	5	0	0	0	1	23	45	9	9	0	1	2
INFJ	-2	28	4	30	0	0	0	25	0	31	0	9083	4	2	7
INFP	-5	43	1043	0	0	0	43	9	19	42	352	8	4	19	
INTJ	-1	20	2	27	2	0	1	0	24	80	21	9832	3	4	8
INTP	-3	16	3	43	3	0	0	1	30	56	148661	2	1	8	
ISFJ	-1	2	2	5	0	0	0	0	39	8914	2211	0	2	3	
ISFP	-0	8	0	10	0	0	0	0	55	16	2033	3	5	3	2
ISTJ	-0	7	0	4	0	0	0	0	45	10	2032	0	2	9	1
ISTP	-1	6	1	9	0	0	0	0	70	17	2865	1	2	3	25
ENFJ	-1	2	2	5	0	0	0	0	39	8914	2211	0	2	3	
ENFP	-5	43	1043	0	0	0	43	9	19	42	352	8	4	19	
ENTJ	-1	3	0	3	0	0	1	0	13	20	3	2	0	0	1
ENTP	-4	23	5	78	1	0	0	0	1382	296	1133	2	5	7	
ESFJ	-0	1	0	3	1	0	0	0	17	18	3	9	0	0	1
ESFP	-0	0	0	4	0	1	0	0	14	27	3	3	0	1	0
ESTJ	-0	3	0	3	0	0	1	0	13	20	3	2	0	0	1
ESTP	-1	4	1	5	0	0	0	1	23	45	9	9	0	1	2
INFJ	-2	28	4	30	0	0	0	25	0	31	0	9083	4	2	7
INFP	-5	43	1043	0	0	0	43	9	19	42	352	8	4	19	
INTJ	-1	20	2	27	2	0	1	0	24	80	21	9832	3	4	8
INTP	-3	16	3	43	3	0	0	1	30	56	148661	2	1	8	
ISFJ	-1	2	2	5	0	0	0	0	39	8914	2211	0	2	3	
ISFP	-0	8	0	10	0	0	0	0	55	16	2033	3	5	3	2
ISTJ	-0	7	0	4	0	0	0	0	45	10	2032	0	2	9	1
ISTP	-1	6	1	9	0	0	0	0	70	17	2865	1	2	3	25

4. 本組嘗試進行 MBTI 的 Binary 分類，以模型區分 I-E, N-S, T-F, J-P，並使用 Accuracy 及 F1 Score 衡量預測效果，結果如下：



由上圖可得，相較於 16 種性格分類，Binary 分類效果改善許多，多數預測準確率皆高於 0.5，其中 S/N 及 E/I 分類效果最佳，推測這兩種分群中，兩類別間典型特徵差異較明顯，使模型更容易進行區別，且樣本當中 I, N 類別樣本佔比高，也使模型更容易完整學習這兩類文本的特徵並進行更準確的預測。

| 結論

經過模型訓練及發現我們使用的六種模型所提取的特徵對預測 MBTI 類型並不具代表性，因此導致模型無法準確對發文者的 MBTI 進行分類。可能原因是我們礙於運算資源有限，僅使

用 TF-IDF vector 作為文章的特徵向量，這樣的局限性可能無法像 BERT 充分反映文本複雜結構、內容和文字之間的相關性。

此外，我們也觀察到資料集不平衡的現象，這也可能讓模型有預測偏誤，傾向於預測占比較高類別的 MBTI 類型，而對樣本數較少的類別預測效果較差。

儘管存在一些不足和限制，本組認為以貼文進行 MBTI 各類別的預測仍具有參考價值，未來可以針對不同 MBTI 類型的原始樣本資料進行平衡，提高模型的整體性能和準確性，另外嘗試使用 BERT，使模型更全面理解文章中每個字間的相關性，也更精確捕捉字的特性，提高預測性能。

| 參考文獻

- Shipankar, S., Sawale, G., Shelke, R., & Khairkar, A. (2022). Personality Prediction Using Social Media Platform. Khairkar, Personality Prediction Using Social Media Platform.
- Al-Fallooji, A. S., & Al-Azawei, A. (2022). Predicting Users' Personality on Social Media: A Comparative Study of Different Machine Learning Techniques. Karbala International Journal of Modern Science, 8(4), 617-630.
- Escobido, M., & Stevens, G. (2013). Can personality type be predicted by social media network structures?. In The Asian Conference on Psychology & the Behavioral Sciences 2013 (pp. 159-173).
- Mushtaq, Z., Ashraf, S., & Sabahat, N. (2020, November). Predicting MBTI personality type with K-means clustering and gradient boosting. In 2020 IEEE 23rd International Multitopic Conference (INMIC) (pp. 1-5). IEEE.