分析社群貼文內容與發文者的MBTI相關性

組員：林姝延、吳郁心、黃崢霖、張書明、藍知生

Outline

1. 簡介與動機
2. 資料介紹
3. 資料清理
4. 資料分布
5. 模型與結果
6. 結論

MBTI

概覽	MBTI 反映人們態度與認知功能偏好，幫助人們了解自己，在職場與生活中做出更好的選擇
動機	透過社群媒體中的留言預測留言者的MBTI，預期未來職場得個客觀精準判斷人們真實性格。
簡介	

資料介紹

	text	label
Twitter	I'm like entp but idiotic|||Hey boy, do you want to watch twitch with me?|||I kin Simon from Alvin A...	infp

	body	subreddit
Reddit	The scores on an individual level mean very little and will vary according to a number of conditions. On some days you may score higher, and others lower. Don't sweat it.	intj

	posts	type
Personality Cafe Forum	'Good one _____ https://www.youtube.com/watch?v=fHiGbolFFGw|||Of course, to which I say I know; t...	INTP

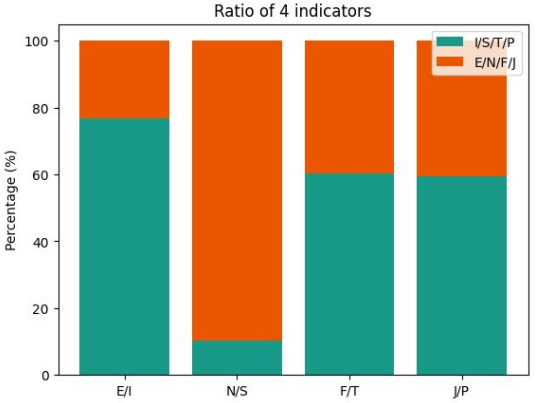
資料清理

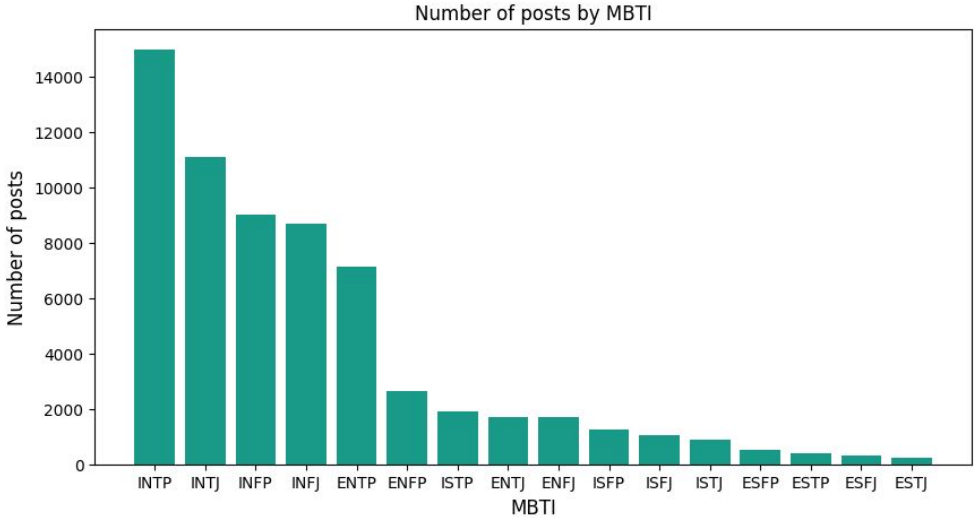
1	移除貼文中的連結	連結無法直接傳達語意，容易干擾模型預測結果
2	詞型態還原lemmatize	將詞語正規化，提升模型預測效果
3	排除短於三個詞的貼文	過短的貼文無法表達個人特質，予以排除
4	隨機挑選3%資料分析	減少資料大小，以提升模型運算效率

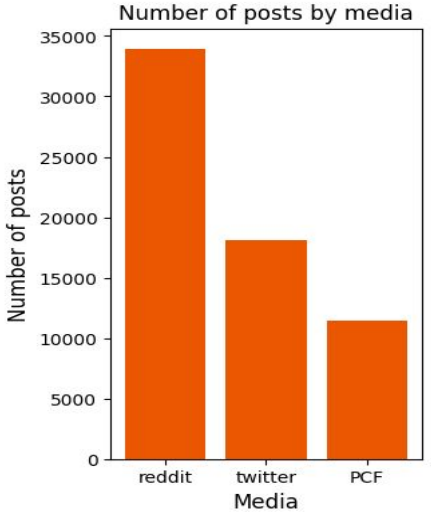
最終資料共 63,498 筆，包含每篇貼文內容及發文者MBTI

MBTI 分佈

- 主要數據來自Reddit（53.4%）、Twitter（28.6%）、PCF（18%）。
- 社群使用中最高數為INTP（23.6%），安靜內斂下蘊含著豐富的想法與洞察，傾向獨處思考、深入探究新的知識和理論、分析事物的運作方式和解決辦法。
- 前四高均為IN 人格（69%），右圖可見I、N 型佔大宗（分別佔76.9%、89.6%）。

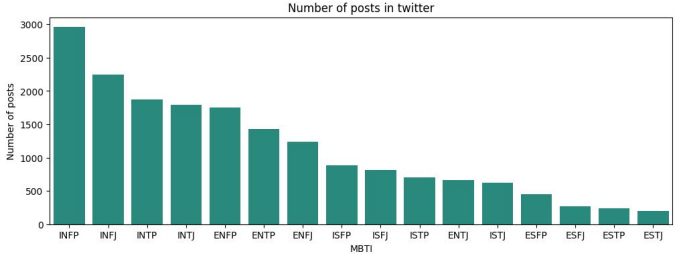


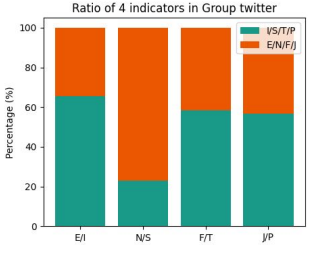


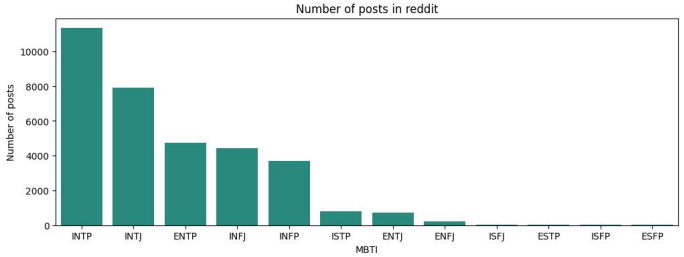


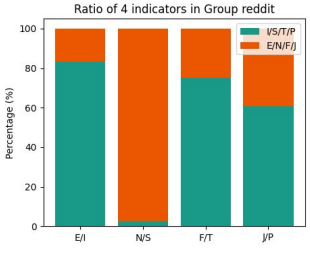
社群平台分佈擁有些微差異

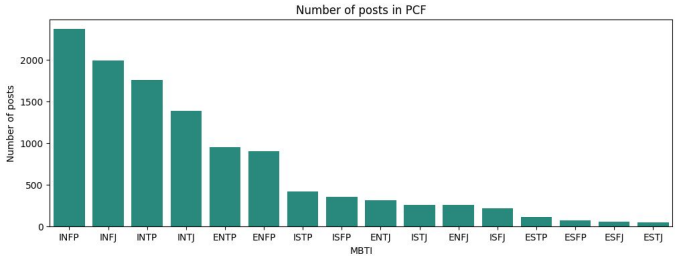
- 因數據來源主要來自 Reddit，整體數據受 Reddit 影響。若分別探討三者社群可發現 Reddit 具較集中的 MBTI 性格，最高者為 INTP、INTJ，與整體接近。
- 而 Twitter & PCF 使用者性格較多元，最高者為 INFP、INFJ。重視人際和諧、情感交流，擁有自己價 值觀且不輕易妥協、情緒豐沛而敏感，特別著迷於音樂、藝術、文學等富含創造力與情感的事物。
- 分佈受資料切割偏差、選擇媒體的偏差，也受使用此類 App 的國家、社經等背景效果。

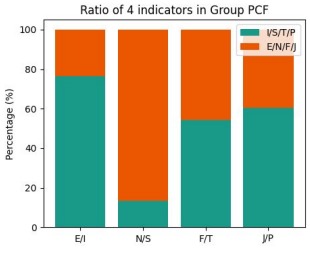






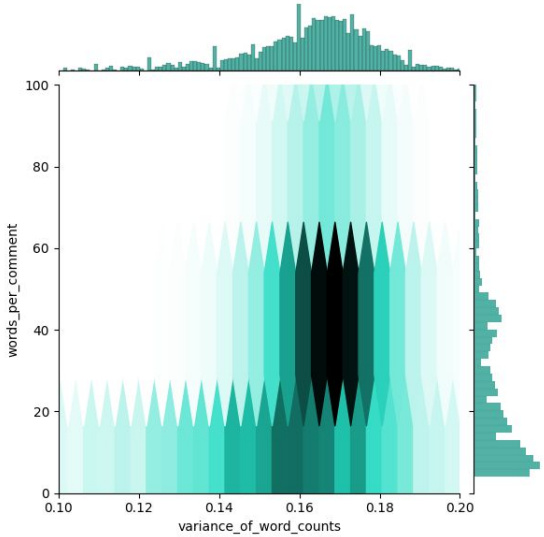


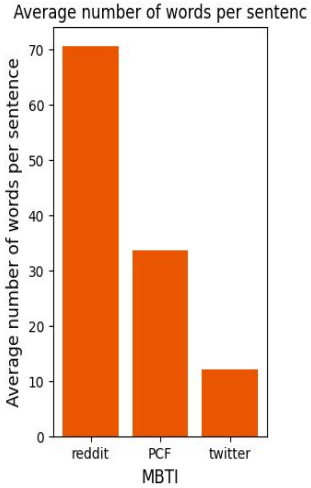


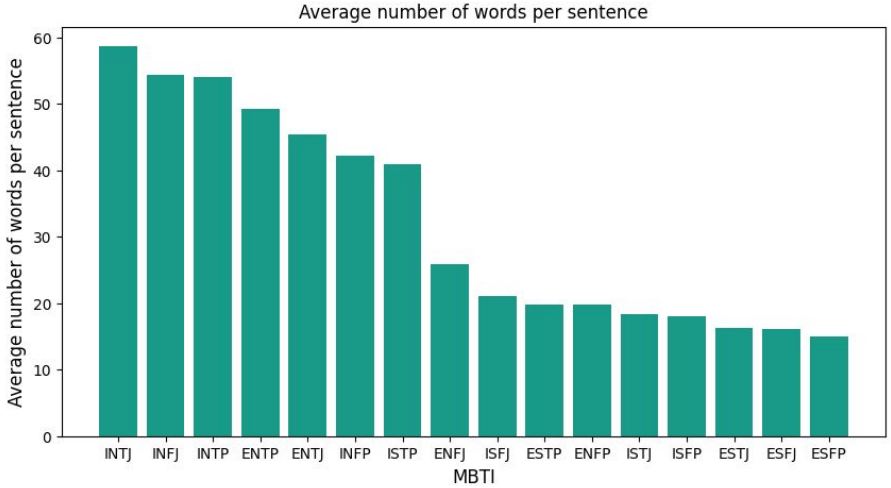


用字習慣

- 字數與變異數較無相關性,Reddit 平均句長最高(近70 字)。
- INTJ 平均用字最高(近60 字), INTJ 富有想像力、喜歡獨立思考、目標導向, 對自己與他人都抱持高標準的完美主義者, 偏好探尋大量資訊與理論, 並熱愛解構與結構, 在豐沛的內在思考中獲得能量, 是 MBTI 16 型人格中極稀有且公認最強的類型, 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。







文字雲--16型人格

- ENFP和ENFJ的文字雲中，“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。
- 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。
- INFP和ISFP的文字雲中，“feel”和“time”的頻率較高，這可能代表這些類型的人在情感表達較為內斂。

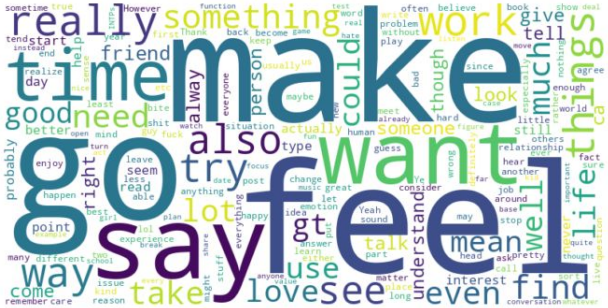
文字雲--社群平台

- 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。
- 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。
- 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。

Word Cloud for twitter

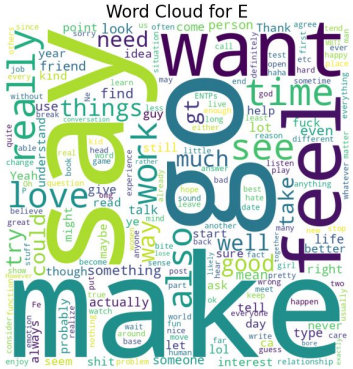
Word Cloud for PCF

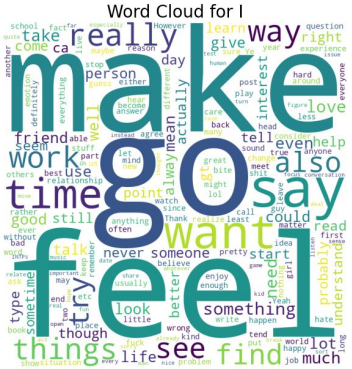
Word Cloud for reddit



文字雲--I/E、S/N

- 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。
- 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。

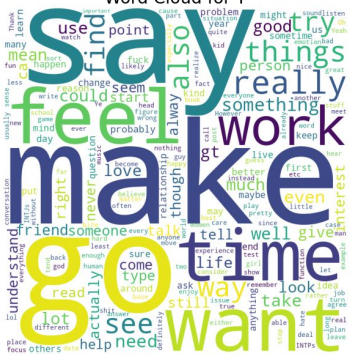




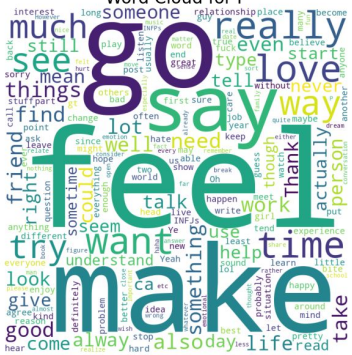
文字雲--T/F、J/P

- 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。

Word Cloud for T

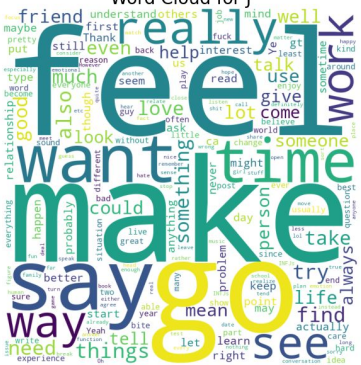


Word Cloud for F

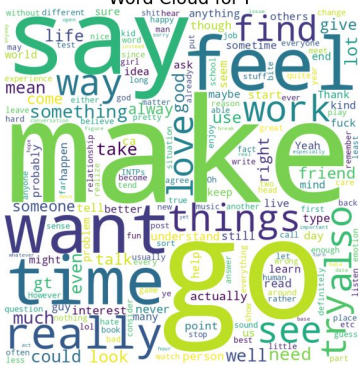


- J型人格的文字雲中，“want”、“make”和“way”相對較大，這可能意味著判斷型的人傾向於有計劃和組織性的溝通。
- P型人格的文字雲中，“want”、“say”和“things”較顯著，這可能代表感知型的人在探索和開放性的對話中較為活躍。

Word Cloud for J



Word Cloud for P

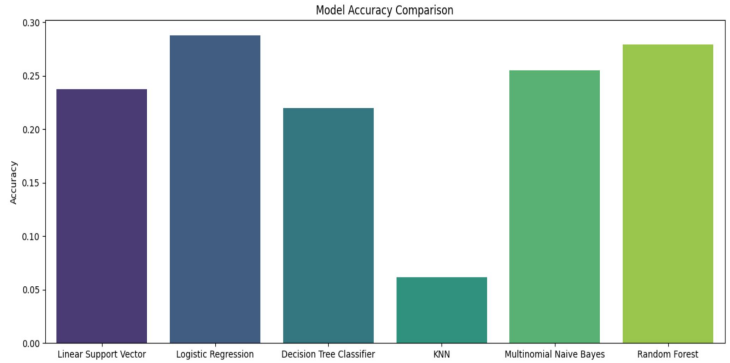


模型簡介

- Linear Support Vector Classifier
- Logistic Regression
- Decision Tree Classifier
- KNN
- Multinomial Naive Bayes
- Random Forest Classifier

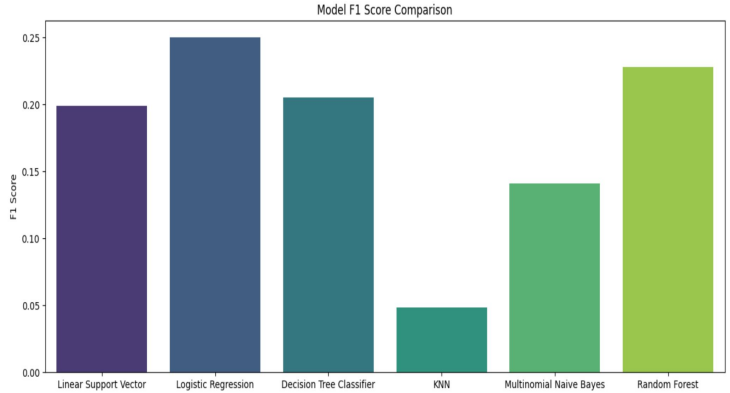
- **90% 訓練**
- **10% 測試**

16種MBTI 預測結果 - Accuracy & F1 Score



| Accuracy

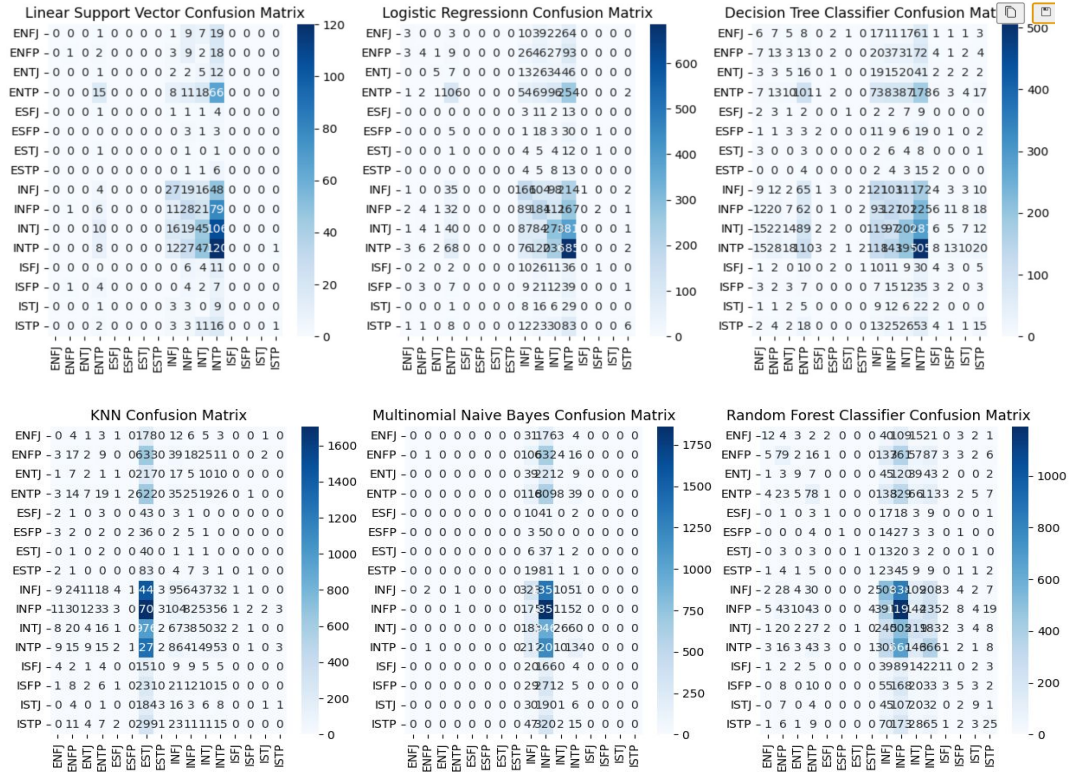
Linear Support Vector = 0.237
Logistic Regression = 0.2874
Decision Tree Classifier = 0.2198
KNN = 0.0618
Multinomial Naive Bayes = 0.2548
Random Forest Classifier = 0.279



| F1 Score

Linear Support Vector = 0.1992
Logistic Regression = 0.2501
Decision Tree Classifier = 0.2051
KNN = 0.04822
Multinomial Naive Bayes = 0.14136
Random Forest Classifier = 0.2283

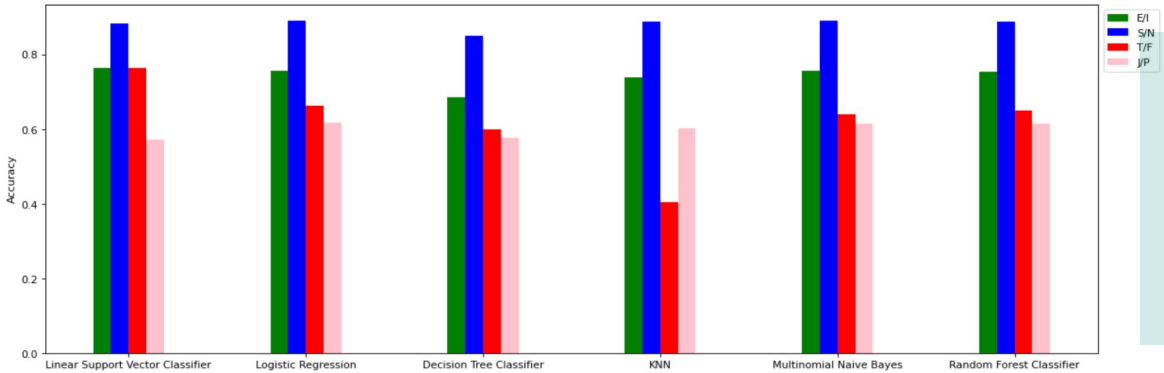
16種MBTI預測結果 - Confusion Matrix

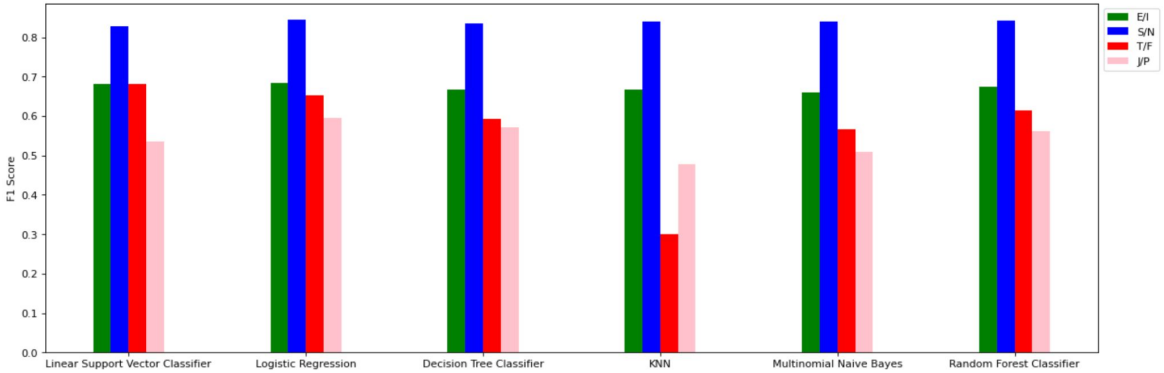


- X軸:模型預測結果
- Y軸:文章 MBTI 分類

- 樣本多集中於 I, N 類別

四大性格指標預測結果 - Accuracy & F1 score





- S / N 分類表現最佳
- E / I 分類表現次佳

模型檢討

16 型人格預測	6 個模型 Accuracy 皆未超過 0.3	
	Logistic Regression 表現最佳	
社群貼文 MBTI 預測效果不佳，其中以 KNN 模型表現最差	KNN 表現最差	
	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
	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
	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
	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
4 大性格預測	預測準確度普遍高於 0.5，部分預測結果高於 0.8	
四種性格分別預測結果稍優於16型人格預測	I, N 類別樣本數最多，模型更容易學習這兩類文本特徵,預測效能較佳	
	典型差異	4 大性格分別進行預測，在文本表達上可能有更明顯的典型性差異，使模型更容易區分

總結

1. 查看各模型的重要特徵，發現不具代表性 → 我們初步認為無法貼文內容分辨發文者之MBTI
2. 資料比例不平衡，導致模型可能有偏誤 → 未來可以針對各種預測平衡資料比例
3. 礙於運算資源有限，我們使用 tf-idf vector作為文章的特徵向量 → 未來可以使用BERT模型，去理解文章裡每個字之間的相關性，並準確反映每個字的特性