

TCFD報告書

Text Mining評鑑內容探討

政治大學統計系余清祥

2026年9月3日

Email: csyue@nccu.edu.tw

<http://csyue.nccu.edu.tw>



研究目的

- ❑ 分析氣候相關財務揭露（TCFD）報告書，以數位工具協助人工評鑑作業。
- 註記及查詢「治理」、「策略」、「風險管理」、「指標和目標」的相關文字。
- 協助專家判讀TCFD報告、判斷哪些文字為電腦生成(或抄襲)。
- 提供文字分析模型及方法，建立量化評鑑機制。



Bloomberg自2024年起使用AI寫報導



Artificial intelligence disciplines, such as natural language processing (NLP) and machine learning (ML), play a central role at Bloomberg.

We have always relied on text as a key underlying source of data for our clients. Over the past decade, we have made significant investments in statistical NLP techniques, extending our capabilities by building state-of-the-art technology for core document understanding, recommendation and customer-facing systems.

<https://www.bloomberg.com/company/values/tech-at-bloomberg/artificial-intelligence-ai/>



人工智慧的發展



□ 人工智慧是近年熱門議題，諸如ChatGPT、Claude、Gemini聊天機器人受到注目，AI工具的使用者擴及社會各界。

→ Bloomberg自2024年起部分報導使用AI；

→ 以AI生成影像、聲音愈發普遍，Deepfake、I2V vs. T2V都是典型範例。

註：整合人和AI將成為主流應用，
辨識兩者風格也是其中用途之一。



大數據時代的文字分析

□ 描述 → 辨識 → 解釋 → 生成 → 評估

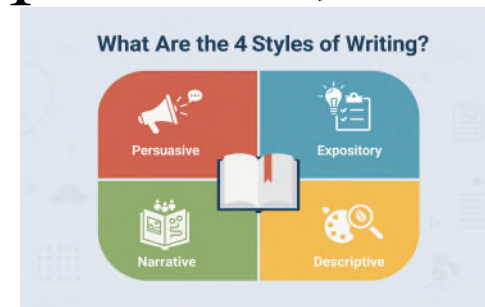
Characterization, Identification, Interpretation,
Generation, Evaluation

→ Writing Style Analytics(文字特徵)

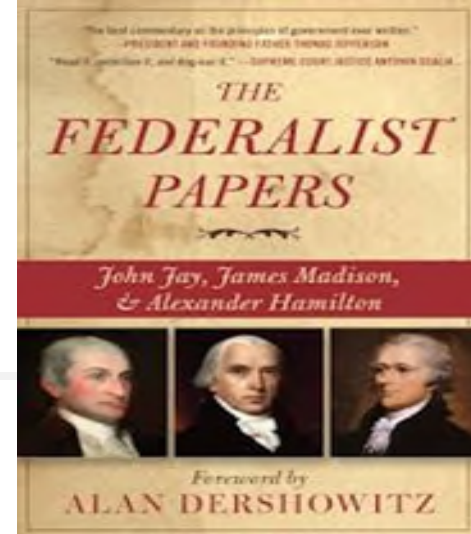
→ Authorship & AI Detection(作者辨識)

→ Information & Decision Analytics(資訊揭露)

→ LLMs & Intelligent Writing(AI人工智慧寫作)



<https://jenni.ai/essay/four-styles-of-writing-explained>



知名的英文分析範例

□ 莎士比亞(Shakespeare)的字彙總數

→ Efron and Thisted (1976) 以Poisson Process
估計莎士比亞字彙，推論1985年在莎翁故居
附近發現一首詩應是他的著作。

□ 《聯邦主義議文集》(The Federal Papers)

→ Mosteller and Wallace (1984) 以貝氏方法分析
「upon」之類虛字，判斷論文集中未署名的
12篇文章應是Madison、而非Hamilton。

常見中文分析議題《紅樓夢》

□ 《紅樓夢》相關研究開始較早

→ 陳炳藻(1980)《紅樓夢》字彙，UW-Madison。

→ 趙岡、陳鍾毅(1980)《紅樓夢研究新編》，
聯經出版社。(其他研究者：胡適、俞平伯)

→ 余清祥(1998)、杜協昌(2012)利用文本採礦探
討《紅樓夢》的作者爭議。

註：金庸《天龍八部》

(倪匡代寫，香港明報)



2023年TCFD報告評分模型

□ 首次以文字分析的角度，估計2023年TCFD報告書的評審分數。

→ 迴歸分析可提供不錯的結果。

	內部			外部		
	顯著變數 個數	Adj-R ²	變異係數	顯著變數 個數	Adj-R ²	變異係數
銀行	9(6)	0.83	0.25	9(5)	0.87	0.08
壽險	3(2)	0.59	0.24	7(4)	0.65	0.07
產險	4(3)	0.58	0.49	4(3)	0.67	0.09

辨識人類與LLM生成文字風格

- 以臺灣碩士生論文摘要為研究對象，比較三種LLM(ChatGPT、Gemini、Claude)生成文字。
- 使用六類文字變數：常見詞、生物多樣性、依存距離等，代入迴歸模型及機器學習模型。

Accuracy / F1	碩士生 vs. GPT	碩士生 vs. Gemini	碩士生 vs. Claude
Logistic Regression	0.9514 / 0.9505	0.9555 / 0.9548	0.9319 / 0.9307
Random Forest	0.9597 / 0.9596	0.9562 / 0.9560	0.9316 / 0.9304
XGBoost	0.9559 / 0.9555	0.9546 / 0.9545	0.9296 / 0.9290

AI工具選擇表



ChatGPT
(善於文字推理)

- 問題詢問
- 資訊蒐集
- 文字推理任務



Gemini
(善於多模態)

- 圖片生成
- 圖片修改
- YT影片摘要



Claude
(善於程式開發)

- 報告生成
- Vibe coding
- 建立Agent Skill

研究目標： TCFD評分模型、寫作風格



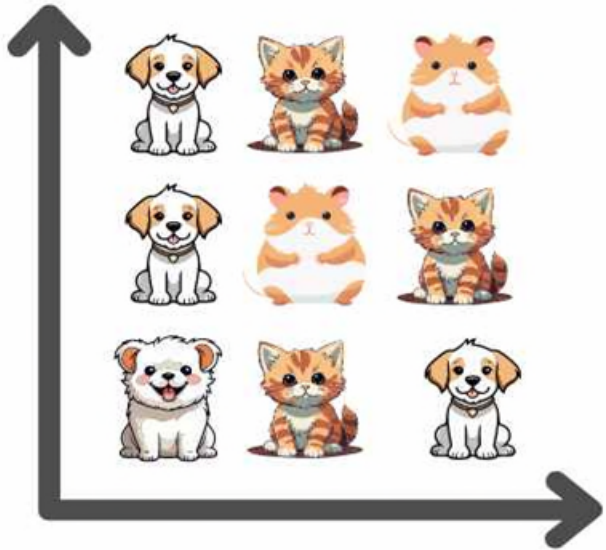
統計學習 vs. 深度學習

□ 統計學習、深度學習都以建立高準確率模型為目標，但兩者目標及思維不盡相同。

	Statistical Learning	Deep Learning
目標	變數間關係	準確性
切入點	數理模型	表徵學習 (Representation Learning)
詮釋	可解釋性	黑箱 (Explainable AI)
研究結果	Knowledge Discovery	最佳模型

什麼是表徵學習 (Representation Learning)

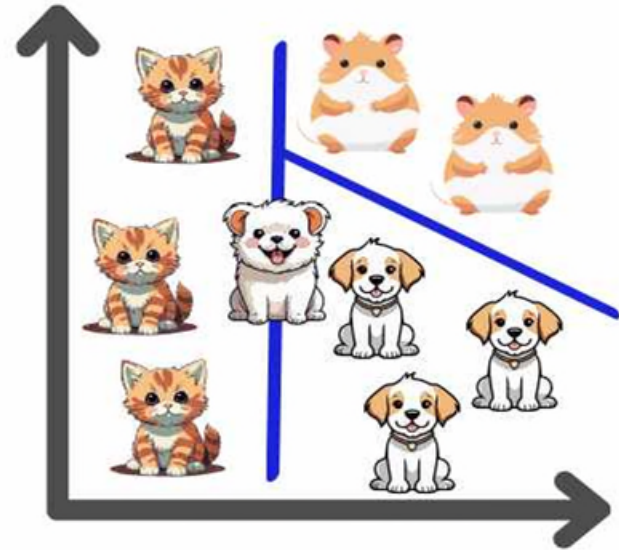
Default Representation



Representation Learning



Semantic Representation



<https://spotintelligence.com/2023/12/11/representation-learning/>



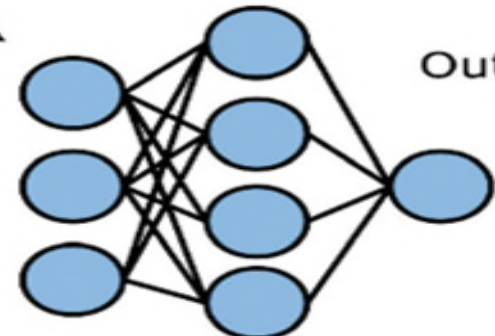
Feature extraction



OR

Raw pixel values

Machine Learning classifier



Output

Cat
or
Dog

研究目標



<https://www.kaoersi.com/xinwendongtai/25490.html>

□ 以2023～2025年TCFD報告及評分資料，檢視報告書的寫作風格、評分模型。

◆ 寫作風格

→ 文字及句型的特色。

→ 各產業的常見字詞？

◆ 評分模型

→ 迴歸模型是否不受時間、評審的影響？

→ 與深度學習模型(BERT)比較。

附表A-1、文字及公司變數

序號	本國銀行變數名稱	序號	壽險、產險變數名稱	序號	隨機抽樣500詞之平均Entropy	序號	隨機抽樣2000詞之平均Entropy
1	銀行狀態	1	有金控母公司	38	隨機抽樣500詞之平均Entropy	38	隨機抽樣2000詞之平均Entropy
2	有金控母公司	2	金控母公司是公股	39	隨機抽樣1000詞之平均Entropy	39	隨機抽樣3000詞之平均Entropy
3	銀行上市櫃狀態	3	資本額	40	隨機抽樣2000詞之平均Entropy	40	隨機抽樣4000詞之平均Entropy
4	公股	4	TCFD_會計師確信	41	隨機抽樣3000詞之平均Entropy	41	隨機抽樣5000詞之平均Entropy
5	資本額	5	TCFD_BSI查核	42	隨機抽樣4000詞之平均Entropy	42	隨機抽樣500詞之平均Simpson
6	TCFD_會計師確信	6	溫室氣體驗證	43	隨機抽樣5000詞之平均Entropy	43	隨機抽樣1000詞之平均Simpson
7	TCFD_BSI查核	7	有委任第三方驗證	44	隨機抽樣500詞之平均Simpson	44	隨機抽樣2000詞之平均Simpson
8	溫室氣體驗證	8	總字數	45	隨機抽樣1000詞之平均Simpson	45	隨機抽樣3000詞之平均Simpson
9	有委任第三方驗證	9	不同字數	46	隨機抽樣2000詞之平均Simpson	46	隨機抽樣4000詞之平均Simpson
10	總字數	10	隨機抽樣500字之平均TTR	47	隨機抽樣3000詞之平均Simpson	47	隨機抽樣5000詞之平均Simpson
11	不同字數	11	隨機抽樣1000字之平均TTR	48	隨機抽樣4000詞之平均Simpson	48	全部詞彙詞向量
12	隨機抽樣500字之平均TTR	12	隨機抽樣2000字之平均TTR	49	隨機抽樣5000詞之平均Simpson	49	全部詞彙詞向量_主成分分析
13	隨機抽樣1000字之平均TTR	13	隨機抽樣3000字之平均TTR	50	全部詞彙詞向量	50	前500大詞彙詞向量
14	隨機抽樣2000字之平均TTR	14	隨機抽樣4000字之平均TTR	51	全部詞彙詞向量_主成分分析	51	前100大詞彙詞向量
15	隨機抽樣3000字之平均TTR	15	隨機抽樣5000字之平均TTR	52	前500大詞彙詞向量	52	前10大詞彙詞向量
16	隨機抽樣4000字之平均TTR	16	隨機抽樣500字之平均Entropy	53	前100大詞彙詞向量	53	文字源於TCFD報告書
17	隨機抽樣5000字之平均TTR	17	隨機抽樣1000字之平均Entropy	54	前10大詞彙詞向量	54	文字源於ESG報告書
18	隨機抽樣500字之平均Entropy	18	隨機抽樣2000字之平均Entropy				
19	隨機抽樣1000字之平均Entropy	19	隨機抽樣3000字之平均Entropy				
20	隨機抽樣2000字之平均Entropy	20	隨機抽樣4000字之平均Entropy				
21	隨機抽樣3000字之平均Entropy	21	隨機抽樣5000字之平均Entropy				
22	隨機抽樣4000字之平均Entropy	22	隨機抽樣500字之平均Simpson				
23	隨機抽樣5000字之平均Entropy	23	隨機抽樣1000字之平均Simpson				
24	隨機抽樣500字之平均Simpson	24	隨機抽樣2000字之平均Simpson				
25	隨機抽樣1000字之平均Simpson	25	隨機抽樣3000字之平均Simpson				
26	隨機抽樣2000字之平均Simpson	26	隨機抽樣4000字之平均Simpson				
27	隨機抽樣3000字之平均Simpson	27	隨機抽樣5000字之平均Simpson				
28	隨機抽樣4000字之平均Simpson	28	總詞數				
29	隨機抽樣5000字之平均Simpson	29	不同詞數				
30	總詞數	30	隨機抽樣500詞之平均TTR				
31	不同詞數	31	隨機抽樣1000詞之平均TTR				
32	隨機抽樣500詞之平均TTR	32	隨機抽樣2000詞之平均TTR				
33	隨機抽樣1000詞之平均TTR	33	隨機抽樣3000詞之平均TTR				
34	隨機抽樣2000詞之平均TTR	34	隨機抽樣4000詞之平均TTR				
35	隨機抽樣3000詞之平均TTR	35	隨機抽樣5000詞之平均TTR				
36	隨機抽樣4000詞之平均TTR	36	隨機抽樣500詞之平均Entropy				
37	隨機抽樣5000詞之平均TTR	37	隨機抽樣1000詞之平均Entropy				

迴歸模型使用的變數

□ 主要使用字詞多樣性及文句結構，沒有加入字詞的意義、詞性、關聯性等資訊。

→ 字詞的TTR、Entropy、Simpson Index

□ 公司相關變數也加入模型。

→ 資本額、官股、溫室氣體驗證

註：本次報告的文字分析以整篇報告為單位，不呈現四要素的個別模型。

量化模型

透過量化模型描述觀察結果：

$$\text{觀察現象} = \text{模型} + \text{誤差}$$

或是

$$y = f(x) + \text{error} ; \text{觀察值} = \text{訊號} + \text{雜訊}$$

□ 數量化模型的關鍵：

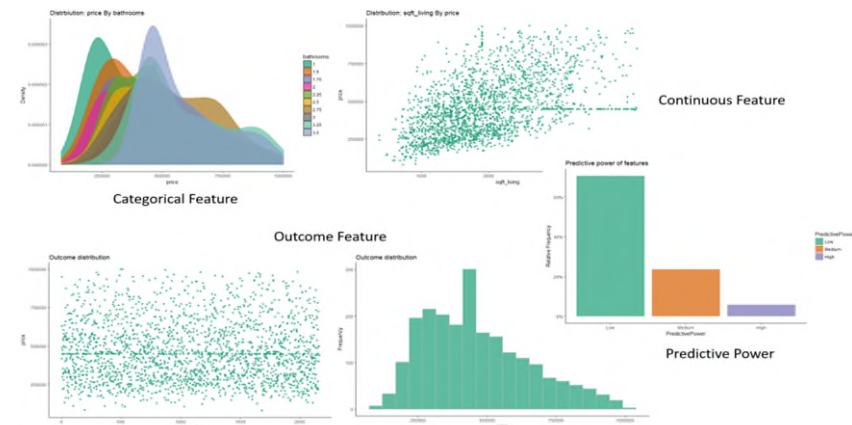
→ 量化目標值 y ：定義問題！

→ 選取關鍵變數： $x_1, x_2, \dots, x_p$

→ 建立量化模型：統計學習、機器學習。

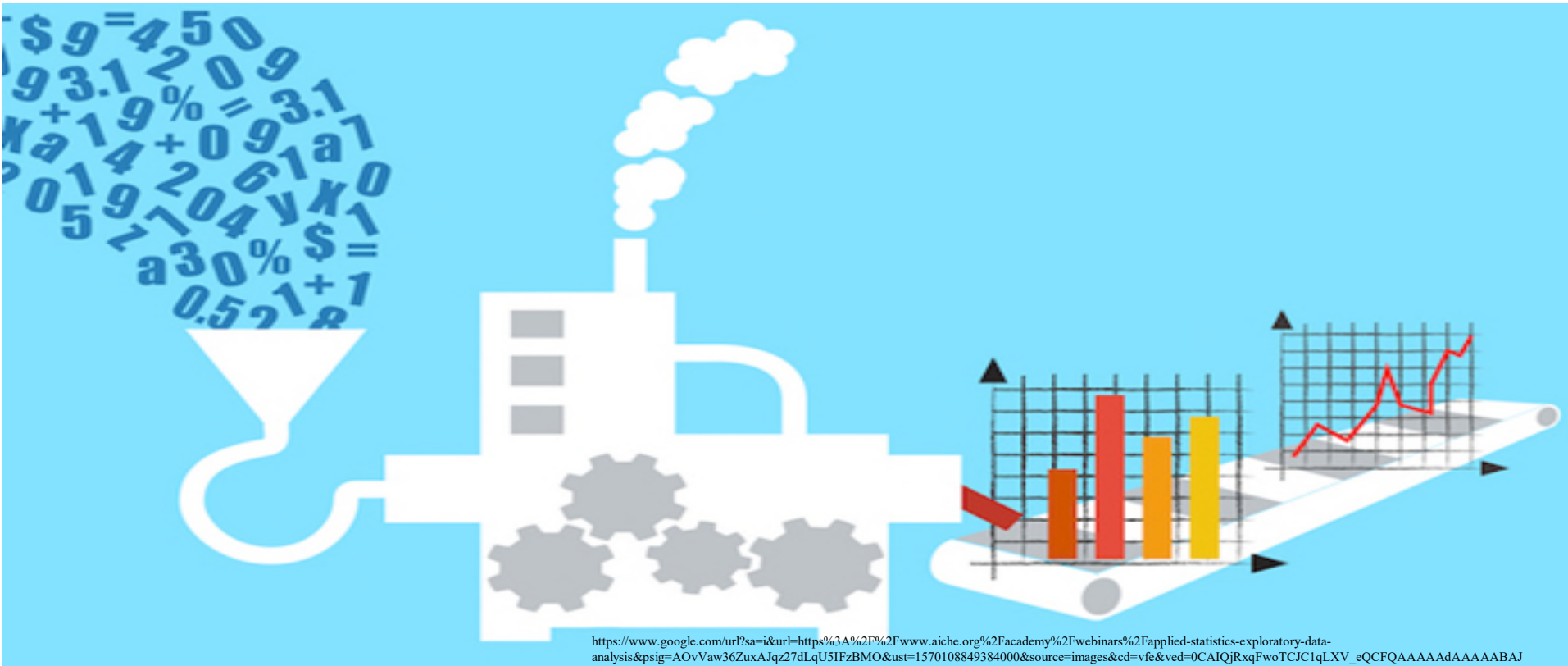
探索性資料分析(EDA)

- EDA首要目的在於資料偵錯、獲得資料的大略資訊、驗證已知結果。
- 圖形、表格在EDA中扮演重要的角色；並由分析結果中尋找合適的下一步分析方法。
- 使用任何的統計方法前，先確定該方法需要的假設條件是否滿足。



探索性資料分析(資料驅動)

- Exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics ... EDA is for seeing what the data can tell us beyond the formal modeling. ---Wikipedia





True/Fake



True/Fake



True/Fake



True/Fake

能夠分辨那個影像是電腦合成嗎？



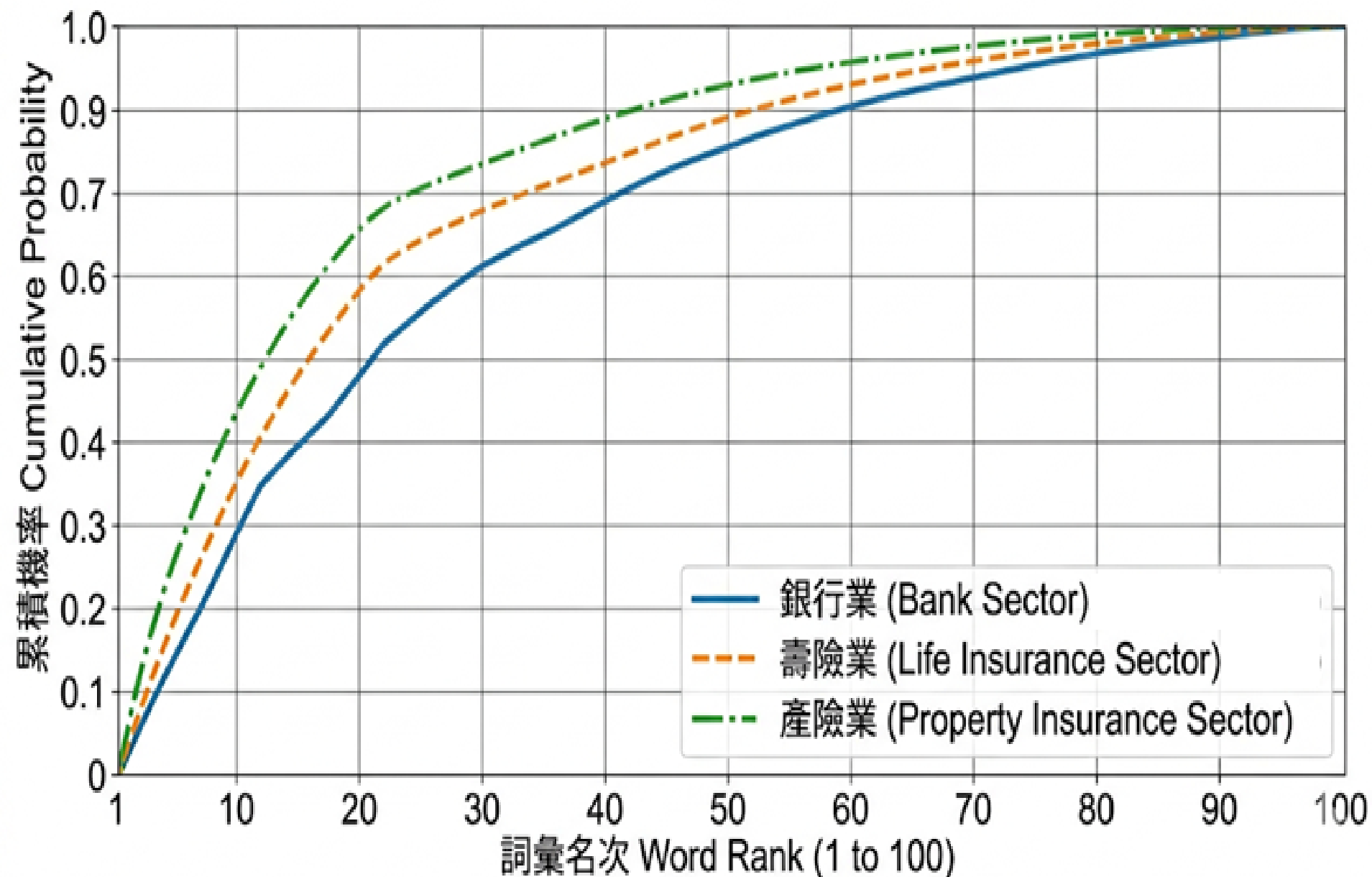
註：影像來源 Celeb-DF-v2，真實影像與深偽影像範例

TCFD 評分模型與文字風格

<https://www.logicraysacademy.com/blog/wp-content/uploads/2023/08/DA2.jpg>



各產業前百大詞彙累積機率



2025年 TCFD報告書銀行業常見詞

Rank 1 2 3 4 5 6 7 8 9 10

銀行

1~10 風險 氣候 變遷 影響 評估 授信 轉型 營運 實體 產業

11~20 銀行 金融 策略 業務 企業 機會 客戶 部位 議題 集團

壽險

1~10 風險 氣候 公司 投資 影響 變遷 營運 評估 轉型 產業

11~20 機會 策略 業務 金融 人壽 企業 實體 委員 低碳 減碳

產險

1~10 風險 氣候 公司 投資 保險 營運 影響 評估 變遷 轉型

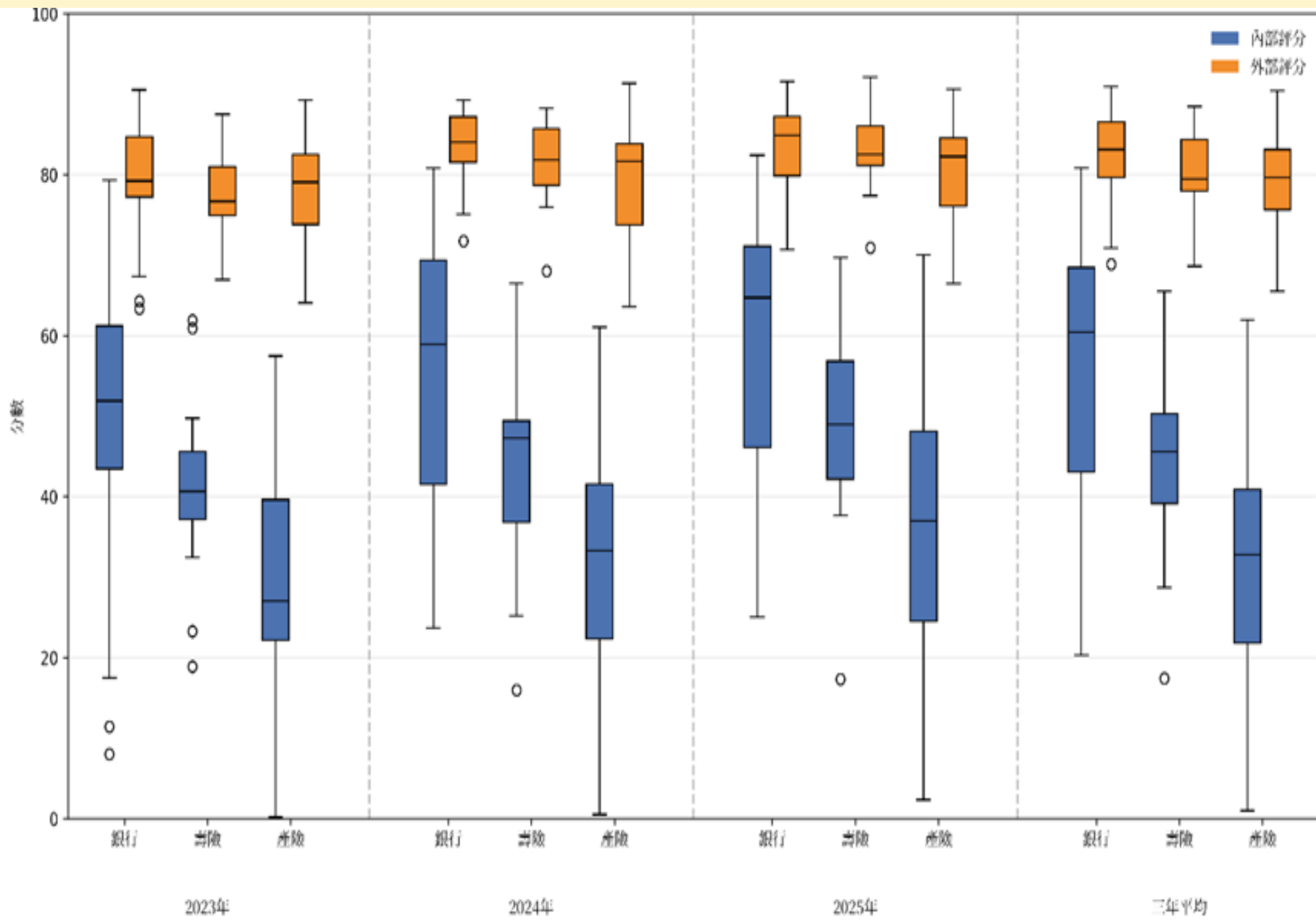
11~20 機會 業務 產險 實體 策略 商品 措施 企業 低碳 金融

產業間的用詞相似性

- 以三種指標測量產業間的用詞相似性：
- 壽險和產險相似性最高；
 - 銀行和產險相似性最低。

產業	Jaccard Similarity	Cosine Similarity	Yue Index
銀行 vs. 壽險	0.400	0.876	0.626
銀行 vs. 產險	0.386	0.866	0.567
壽險 vs. 產險	0.500	0.936	0.671

2023~2025 評分相當一致！



內部評分有較高的變異係數 $CV = \sigma/\mu$

年度	評審分數	銀行	壽險	產險
2023	內部	0.35	0.24	0.49
	外部	0.08	0.07	0.09
2024	內部	0.31	0.27	0.52
	外部	0.05	0.06	0.10
2025	內部	0.27	0.25	0.51
	外部	0.06	0.06	0.08

2023~2025年迴歸模型的估計結果

		內部			外部		
		變數個數	R ²	Adj.R ²	變數個數	R ²	Adj.R ²
2023	銀行	3	0.666	0.636	4	0.783	0.757
	壽險	4	0.411	0.264	5	0.691	0.588
	產險	5	0.817	0.747	4	0.800	0.743
2024	銀行	6	0.785	0.744	3	0.721	0.696
	壽險	9	0.832	0.694	6	0.832	0.760
	產險	10	0.966	0.923	11	0.985	0.960
2025	銀行	5	0.684	0.635	2	0.598	0.575
	壽險	4	0.700	0.620	3	0.613	0.545
	產險	6	0.918	0.877	3	0.870	0.844

5.8±2.3

.75±.16

.68±.19

4.5±2.7

.77±.12

.72±.14

2025年迴歸模型的顯著變數



銀行業



壽險業



產險業

內部模型

外部模型

內部模型

外部模型

內部模型

外部模型

- 銀行狀態
- TCFD_BSI 查核
- Entropy_Unigram500
- Simpson_Unigram500
- 不同詞數

- 資本額
- 不同詞數

- 資本額
- TTR_Bigram500
- Entropy_Unigram500
- Simpson_Unigram500

- 資本額
- 不同字數
- TTR_Bigram500

- TCFD_會計師確信
- 有任第三方驗證
- 總字數
- 不同字數
- Entropy_Unigram500
- Entropy_Bigram500

- 金控母公司是公股
- TCFD_會計師確信
- 有任第三方驗證

迴歸模型與BERT的比較

- 根據023~2025年所有資料($82 \times 3 = 246$ 筆)，透過交叉驗證評估迴歸模型、BERT。
 - 將資料隨機分為80% Training及20% Testing；
 - 以80%資料建構模型，再將模型套入20%資料計算Testing Errors；
 - 評估標準： R^2 、MAE(絕對)、RMSE(平方)；
 - 重複模擬300次，計算誤差平均及變異數。

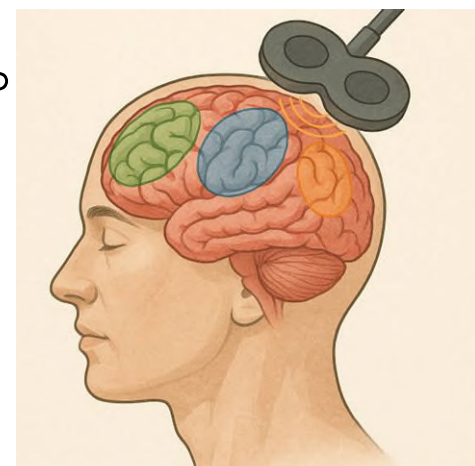
□ 迴歸模型未必不如大語言模型！

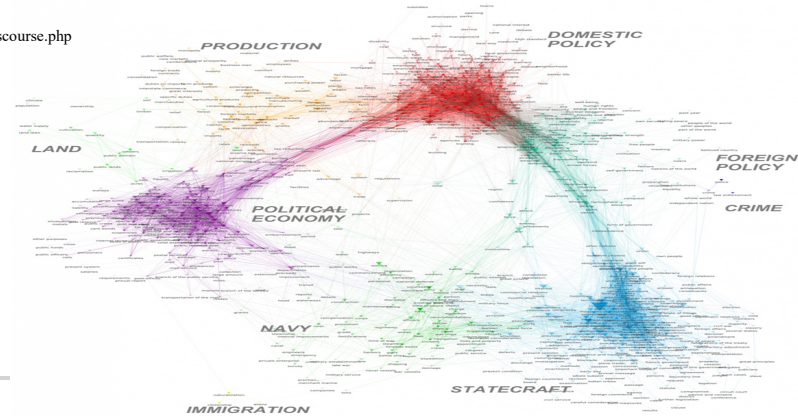
→ 兩者的準確率不相上下，迴歸模型可用於詮釋，
計算速度也較快。

模型		R^2	MAE	RMSE
內部	Regression	0.457±0.304	9.790±0.983	11.948±1.973
	BERT	0.517±0.128	8.701±1.010	11.378±1.280
外部	Regression	0.485±0.111	3.250±0.338	4.148±0.450
	BERT	0.371±0.152	3.555±0.387	4.552±0.523

分析發現

- 根據定義的解釋變數，迴歸等模型大致能夠描述評審的想法，效果與BERT接近。
 - 解釋變數可用於實質詮釋。
 - 2023～2025年分析結果相對穩定。
- 銀行、壽險、產險TCFD的文字風格
 - 三種產業常用詞有很高的重疊性。
 - 銀行業的用詞較為多元。





討論與建議

- 文字分析僅置入多樣性指標、不均度
 - 常見詞及共現性指標(關鍵詞叢)、目標詞特徵、依存距離、顯性主語句比例。
- 模型泛化能力？
 - 本研究可應用至其他類型報告？（變數）
- 四個核心要素的個別模型？
 - 四個子模型的效果不如大模型。

報告完畢，
敬請指教！

