以統計方法辨識 ChatGPT 生成之論文摘要

余清祥[1]　　　　張祐瑜[2]
Jack C. Yue　　　　You-Yu Chang

摘要

大型語言等模型為生活帶來便利，但也帶來潛在隱患，屢傳獲獎文章及創作依賴 ChatGPT 等 AI 工具，不僅引發公平性的討論，也顛覆大眾對於教育、創新等之定位。偽造論文會污染學術研究，讓社會大眾把誤謬當成事實傳播，導致錯誤政策及資源浪費，目前真偽偵測大多依賴機器學習模型，缺乏跨領域的判斷標準。本文以民國 107～109 年臺灣經濟學門碩士論文摘要為研究對象，以 ChatGPT 生成相同議題的摘要，透過探索性資料分析及統計方法比較兩種文本的寫作風格，作為判別論文真偽的依據，所需變數個數及計算時間較大型語言模型更有效率。研究整理兩種文本的字、詞、句長及多樣性等基本統計量，結合常見字詞、模糊性詞彙作為解釋變數，用於區隔文字是否為電腦生成，並比較羅吉斯迴歸與機器學習模型的分類效果。分析發現 ChatGPT 生成摘要傾向使用短句，碩士論文則長短句交錯；生成文本使用「提升」、「提供」、「建議」及「此外」等詞彙的比例較高，碩士生運用虛字「之」頻率較高。由探索性資料分析挑選之變數代入羅吉斯迴歸、隨機森林等機器學習模型，分辨文本是否為電腦生成的準確率相當高。另外，本文方法也與常見大型語言模型 BERT 比較，兩者的準確率大致相同，本文所需變數個數及計算時間較大型語言模型更有效率，並能建議哪些特徵可用於辨識寫作風格，未來可應用於其他非結構資料的分析。

關鍵詞：文字分析、探索性資料分析、ChatGPT、大型語言模型、寫作風格

[1] 余清祥，國立政治大學統計所教授，email：csyue@nccu.edu.tw
[2] 張祐瑜，國立政治大學統計學系碩士班研究生，email：112354021@nccu.edu.tw

A Statistical Approach to Identify Taiwan's Master and ChatGPT-generated Abstracts

Jack C. Yue[3] You-Yu Chang[4]

Abstract

Large language models (LLM) has brought convenience to our lives, but it also poses potential risks. Award-winning articles and creative works using artificial intelligence tools such as ChatGPT have not only sparked controversy over fairness but also reshaped public perceptions of innovation. This paper examines abstracts of master's dissertations in economics from Taiwan between 2018 and 2020. Using ChatGPT to generate abstracts on the same topic, we compared the writing style through tools from exploratory data analysis (EDA) and basic statistics, such as the length and diversity of words, phrases, and sentences of the two types of texts, combined common words and vocabulary as explanatory variables. We applied these variables into logistic regression and machine learning models to distinguish computer-generated text. The classification accuracy of the proposed approach was high, comparable to that of the frequently used LLM model BERT. The proposed method uses fewer variables, is computationally more efficient, and provides information on which features can be used to distinguish writing styles. Also, the results showed that ChatGPT-generated abstracts tend to use shorter sentences, while abstracts of master's theses use a mixture of long and short sentences. On the other hand, the generated texts tend to use more words like "enhance," "provide," "suggest," and "in addition," and master's abstracts use the functional word "之" more frequently.

Keywords: Text Analysis, Exploratory Data Analysis, ChatGPT, Large Language Models, Writing Style

---

[3] Professor, Department of Statistics, National Chngchi University, email: csyue@nccu.edu.tw
[4] Master, Department of Statistics, National Chngchi University, email: 112354021@nccu.edu.tw

一、緒論

隨著大數據時代的興起，文字、語音與圖像等非結構型資料（Unstructured Data）成為時下熱門研究議題。文字分析可追溯古希臘學者亞里斯多德（Aristotle）研究語言的修辭和邏輯結構，他的想法影響古代至今的修辭理論，包括著作《修辭學》是修辭學家視為很重要作品（Golden et al., 2007）。二十世紀中期（約 1950 年代）文字分析與計算機科學結合，圖靈測試（Turing test，亦稱為模仿遊戲）是其中知名範例，這個實驗由人工智能之父艾倫・圖靈（Alan Turing）提出，檢視電腦對於問題的文字回覆，作為判斷電腦是否有智慧及思考能力的依據。伴隨電腦運算能力的大幅提升，二十一世紀的文字分析邁向人工智慧階段，許多研究嘗試以機器模型解讀文字特色，像是《哈利波特》的作者 JK 羅琳曾以筆名羅勃・蓋布瑞斯撰寫犯罪小說《杜鵑的呼喚》（The Cuckoo's Calling），被英國《星期日泰晤士報》（The Sunday Times）記者委託文字分析專家辨認出（Archer and Jockers, 2016）。

用於文字分析的方法中，大型語言模型（Large Language Model，LLM）屬於近年廣受歡迎的深度學習模型（Deep Learning Model），藉由大量文章、影音、書籍中學習單詞和句子之間的關係，使其擁有回答問題、翻譯、生成文字及辨識寫作風格的功能。LLM 為具有大量參數（通常數十億至萬億級的參數）類神經網路組成的語言模型，在 2018 年左右出現，以自監督學習或半監督學習（Self- or Semi-Supervised Learning）對大量文字進行訓練。較為知名者包括 ChatGPT (Generative Pre-trained Transformer)和 BERT (Bidirectional Encoder Representations from Transformers)，這兩者都是 Transformer 架構的自然語言處理（Natural Language Processing, NLP）模型，其中 ChatGPT 用於生成文本，BERT 用於理解文本，是許多人日常生活不可或缺的 AI 工具。

然而 ChatGPT 之類的生成式人工智慧（Generative AI）軟體雖大幅提高工作效率，卻可能存在未知隱患，如數據安全、侵犯個人隱私與知識產權等（郭小東，2023）。近年屢傳得獎作品來自 AI，甚至有些期刊論文將 ChatGPT 列為共同作者，引起許多討論與爭論（Kuta, 2023；Stokel-Walker, 2023）。現在大多數人的共識是作者應全權為投稿文章負責，因此 ChatGPT 之類的 AI 工具不應被視為作者。為此《科學》（Science）期刊更新了編輯政策，明確規定 AI 生成之文字、圖片、影像皆不可用於投稿文章，甚至不得使用 AI 協助內容製作，而且愈來愈多

期刊要求作者標註文章是否由電腦生成，簽署文章並非 AI 生成的保障書[5]。

ChatGPT 等 AI 工具的盛行也影響了學校教育，筆者近幾年教授的統計課程中，使用 ChatGPT 協助撰寫作業及報告的比例快速增加，學生們不必理解方法的原理及精髓，直接要求 AI 提供操作範例及程式碼。以卡方獨立性檢定為例，輸入「*幫我用 R 軟體執行卡方檢定*」之類的文字，ChatGPT 會以一個例子展示計算過程及說明結果，學生們可以根據程式碼模仿操作。在這個例子裡我們可以把 ChatGPT 視為助教，協助學生熟悉課程內容的小幫手，但其中也有囫圇吞棗的隱憂，學習者有直接背誦電腦提供資訊之虞，缺乏質疑及交叉驗證的雙向溝通。筆者嘗試檢查過 ChatGPT 提供的參考文獻，發現作者和著作不時有出入，而引述內容也會有張冠李戴的現象，似乎 ChatGPT 嘗試自我創作、東拼西湊，因此現在還無法完全信賴 AI 工具的產出品質。然而，有些研究生整理相關文獻時，直接抄錄 AI 提供的清單，沒有逐篇檢查文獻內容的真實性，讀者如果不是相關領域的專家，可能無法分辨孰真孰假。

筆者所在的學校雖然採用論文比對系統，但僅止於偵測論文抄襲，尚未考量是否內容來自於生成式軟體，因此本文探討是否可藉由統計方法及文字分析的角度，辨識人、電腦產生文字的差異，希冀研究結果可作為分辨電腦生成文字風格的參考。本文以碩士論文的摘要為研究對象，由於 ChatGPT 是 OpenAI 於 2022 年提出，為避免研究生使用 AI 工具撰寫論文，我們選擇民國 107 年至民國 109 年（2018～2020 年）臺灣經濟學類碩士論文摘要。首先將真實的碩士論文摘要代入 ChatGPT，生成相同議題的摘要，再將論文分成訓練集及測試集，以交叉驗證測試選擇的變數及分類模型的穩定性。以下各節將分別回顧相關研究、資料特性，以及資料分析的想法，最後再呈現研究結果，希望可以提供一個具解釋性的方法判讀真假論文摘要，幫助讀者分辨文本是否為電腦生成。

二、資料與研究方法

數量化文字分析大致可追溯至 1940 年代，圖靈提出的 Turing 估計量（Turing's Estimate）原先在於估計一個生態系有多少不同物種（Species），之後被推廣為 Jackknife 估計量，並用於估計莎士比亞知道多少種不同字彙（Efron and Thisted,

[5] 參考 Science 期刊的投稿資訊網頁，Science Journals: Editorial Policies，https://www.science.org/content/page/science-journals-editorial-policies

1976; Thisted and Efron, 1987）。1980 年代自然語言處理（Natural Language Processing, NLP）受到符號主義（Symbolism）方法影響，仰賴專家編寫規則來解析語言，之後引進統計方法，並在自然語言處理獲得廣大的應用，例如：Church（1988）提出馬可夫鏈（Markov Chain）和隱馬可夫模型（Hidden Markov Model, HMM）的統計方法來標註詞性，打破以往基於規則的方式標註詞性。大數據的興起使得文字資料取得更為便捷，可藉由資料庫建立預訓練語言模型（Pre-trained Language Models, PLM）之類的工具，像是 Vaswani 等人（2017）提出的 Transformer 架構徹底改變 NLP，顯示文字分析的 AI 時代已經來臨。

大型語言模型在自然語言處理大多能取得高準確率，仰賴複雜模型結構及龐大（動輒數千億個）參數，如 GPT-3 參數數量已達 1750 億，因此需要搭配高性能電腦及較長運算時間。另外，模型準確性通常不是實務應用時的唯一考量，經常需要羅列有哪些關鍵變數，以作為詮釋結果及驗證模型之用，像是上述辨認出 JK 羅琳以筆名發表小說的重要變數之一是「she」的使用率，由於西方偵探小說大多站在男性的觀點，鮮少著墨於女性角色，但現在 AI 模型尚無提供這類型資訊的功能。因應詮釋及關鍵變數的需求，文字分析可採納 Tukey 提倡的探索性資料分析（Exploratory Data Analysis, EDA），藉由檢視各變數特性及變數間的關係，找出可分辨寫作習慣及作者風格的重要因素。事實上，完整的統計分析可分為 EDA 和 CDA（Confirmatory Data Analysis，驗證性資料分析），先以 EDA 找出關鍵資訊，再選擇適當的 CDA 模型確定目標變數與解釋變數間的關係，藉以確定屬於作者獨特的創作指紋。

以中文分析為例，《紅樓夢》作者歸屬至今仍是文學界一大謎團，對此 Hu 等人（2014）選取詞頻、句長與詞性分佈等特徵，再輔以主成分分析（Principal Component Analysis）驗證《紅樓夢》一書前 80 回與後 40 回風格差異，認為《紅樓夢》應是多位作者的集體創作。余清祥（1998）檢視白話文常見虛字「兒」、「在」、「了」、「著」的出現頻率，以及區隔中國南北方常見詞的出現比例，透過 t 檢定、卡方檢定、變動點分析（Chang-point Analysis），顯示《紅樓夢》前 80 回與後 40 回文字風格存在顯著差異。我們也認同 EDA 和 CDA 相輔相成的角色搭配，本文先由探索性資料分析的角度出發，探討碩士生撰寫與 ChatGPT 生成文本的風格特色，從中篩選出能夠反映出兩者差異的文字變數，作為判別論文摘要是否為電腦生成的依據。

本文蒐集 107～109 學年度臺灣經濟學類碩士論文摘要共計 1484 篇，其中

66 筆資料遺失論文題目或摘要內容，遺失資料占比約 4.4%，最終合計 1418 筆原始資料。沒有選擇 110 學年之後資料的原因是 ChatGPT 出現於 2022 年，110 學年入學者畢業時間通常為 2023 年及之後，避免研究生參考 AI 工具撰寫論文而影響資料品質，不考慮 110 學年以後的碩士論文摘要。另外，統計分析的誤差與樣本數有關，考量人力、時間、問題要求等因素，通常會視情況需要抽取部分資料，本文將從 107～109 學年度隨機抽取 500 篇論文摘要作為研究對象[6]，再分別將這些論文的標題、摘要輸入 ChatGPT，產生與碩士論文類似議題的摘要，並以本文提出方法驗證是否能夠區隔人、電腦的風格差異。預期由標題產生的文字會比較容易辨識，與摘要產生文字、碩士論文摘要差異較大，我們將詳細探索這三者的差異，作為挑選區隔人、電腦文字關鍵變數的參考。

英文等語言可用空格做為分詞依據，中文方塊字必須依賴適當的斷詞以解讀語意，不同斷詞方式會得出大相逕庭的結果。大家耳熟能詳的「下雨天留客天留我不留」就是典型範例，可解讀為「下雨天，留客天，留我不留？」(歡迎客人留宿)，也能詮釋為「下雨天留客，天留我不留！」(主人不樂意招待)，就看語氣轉換所在位置而定。現代中文書寫以白話文為主，描述時大多透過雙字或多字等詞彙，而非單一字彙為基本單位，因此斷詞通常是文字分析的重要步驟，這與分析英文時需先辨識字根（root）及詞性有很大的差異。Jieba 和 CKIP 是臺灣常用的中文斷詞軟體(余清祥、葉昱廷，2020)，雖然 Jieba 斷詞執行速度通常較快，但用於正體中文的精確性不如 CKIP，這或許與 Jieba 根據《人民日報》等簡體中文文本為資料集有關，因此本文使用 CKIP 斷詞。CKIP 是中央研究院中文詞知識庫小組（CKIP Lab）開發的斷詞套件，近年推出以 BiLSTM 架構訓練的 CKIP Tagger（Li 等人，2020）及 CKIP Transformers 等版本，除了具有基本的斷詞功能，還能進行詞性標註（Part-of-speech Tagging，或稱 POS Tagging）和句法分析(Parsing)[7]。經過 CKIP 軟體的斷詞後，再計算字詞相關的敘述性統計量，包括字詞總數、使用比例，作為比較不同文本的參考依據。

此外，本文也引進豐富度（Richness）與生物多樣性（Species Diversity）等指標，用以測量字詞等寫作風格，這些指標常見於資訊理論及語言學，可用於描述三種文本的特性及差異。相異字比例（Type-Token Ratio, TTR）原先用於字彙

[6] 本文嘗試抽取 100、200、500 篇碩士論文，但樣本數少時的變異數偏大，在此僅呈現隨機抽取 500 篇碩士論文的結果。
[7] 詞性、句法的分辨與語言學及文章領域有關，本文不考慮這方面的文字分析。

的豐富度，衡量文本中不同字彙的比例，可定義為 TTR = $\frac{字彙種類}{總字數}$，這個指標也適用於描述詞彙的豐富度。由於字彙種類通常不隨著總字數而上升，或是 TTR 數值在總字數較多時會偏低，因此計算時會規定一定數量的字數，這種方式或稱為標準化相異字比例（Standardized TTR, STTR），本文假設隨機抽取 100 字（抽出放回）計算 TTR。另一個多樣性指標為熵（Entropy）或稱為 Shannon Index，Shannon 於 1948 年根據熱力學第二定律提出，Entropy 的定義$-\sum_i p_i \log(p_i)$，其中p_i為第 i 種字彙（或詞彙）使用比例，與 STTR 的作法相同，本文也是隨機抽取 100 字計算熵值。

本文將上述方法挑選出之變數代入分類模型，總共有三種模型：羅吉斯迴歸（Logistic Regression）為本文主要探討的模型，具備可解釋性及計算量較少的優點；隨機森林（Random Forest；Breiman, 2001）為集成式決策樹模型，透過多樣性與多數決機制提升泛化能力；XGBoost（Chen and Guestrin, 2016）則為梯度提升樹（Gradient Boosting Tree）的強化版本，具備優異的學習效率與預測性能。本文比較這三種模型、BERT 的分析結果，其中 BERT 執行步驟與本文透過 EDA 挑選不同，雖然通常也會經過斷詞（Tokenization）決定置入模型的基本分析單位—詞元（Token），但並非上述提到的 CKIP 或是 Jieba。

三、探索性資料分析

探索性資料分析由統計學家 John W. Tukey 於 1970 年代提出，他主張統計分析不應只著眼於 CDA 等模型的，而是先透過資料視覺化（Data Visualization）與基本統計量，摘要出資料中潛藏的結構、模式與異常特性，藉此提出可能的研究議題及驗證假設。常見的 EDA 工具包括統計圖表，如折線圖、箱型圖等，以及摘要統計量，平均數、標準差及離群值，以此獲得大略的變數分布與資料的基本特性。除了字詞之外，本文也將標點符號的使用列入考量，中文標點符號雖然有 2000 年以上的歷史，但直至 1920 年才由中華民國教育部正式公告，對於白話文的推廣非常重要。不過，中華人民共和國在 1951 年也公布了標點符號用法的相關規定，有幾種標點符號的使用方式與臺灣不同，像是兩岸雖然都有單引號、雙引號，但臺灣是「」及『』等直角引號，中共則採用英文格式的‘’及“”。由於大語言模型的訓練資料有不少來自海峽彼岸，由此產生的文字多少會保留部分標點

符號，亦即標點符號也是重要的參考變數。

表 1 為兩種 ChatGPT 生成摘要（代入標題、摘要）及經濟學門碩士生摘要的文字基本統計量。以標題生成的文字特性最為不同，雖然每篇摘要的篇幅較長，但多樣性卻最低，包括字彙、詞彙、標點符號的種類最少，以摘要生成的文字與碩士生摘要比較接近，但詞彙和標點種類仍有一段差距。另外，本文仿造先前研究，以五種標點符號「，。；！？」作為斷句的判斷依據（何立行等人，2014），無論平均每句使用的字數、詞數都以碩士生摘要比較長，但與 ChatGPT 生成摘要的差異沒有字詞等種類來得大。最常見的字、詞、標點所佔的比例也可仿造表 1 的方式呈現（詳見表 2），以標題生成的文字同樣有較低的多樣性，相同的種類數卻有較高的比例，相對而言，碩士生摘要的字詞及標點多樣性最高，但與摘要生成的文字差異比較小。

表 1、摘要文字使用的基本統計量

		標題生成	摘要生成	碩士生
	篇數	500	500	500
字	總數	281,188	181,193	201,884
	種類	1,699	1,940	2,089
詞	總數	158,482	103,522	116,584
	種類	5,766	6,978	8,103
標點	總數	20,655	16,437	17,988
	種類	26	34	46
句	總數	17,116	11,310	11,604
	字數	16.43	16.02	17.40
	詞數	9.24	9.07	9.96

表 2、摘要的常見字詞及標點符號的使用比例

		標題生成	摘要生成	碩士生
	常見	500	500	500
字頻	100	52.27%	47.36%	47.10%
比例	500	92.46%	89.12%	88.47%
詞頻	100	29.81%	23.16%	21.39%
比例	500	53.45%	44.71%	41.32%
標點	5	89.63%	80.07%	77.40%
比例	11	97.83%	93.78%	91.46%

圖形大多能呈現更為多元的資料面貌，像是直方圖（Histogram）、箱型圖（Boxplot）可提供較為詳細的資訊，透過視覺輔助選取適當的解釋變數。先以常見詞彙的累積機率為例，表 2 僅呈現最常見 100 個詞、500 個詞，我們可仿造累積分布函數（Cumulative Distribution Function，CDF）的計算，得出最常見 500 個詞的出現機率（圖 1）。碩士生摘要的 CDF 累積速度較慢，可詮釋為文字多樣性較高，與表 1、表 2 的結果相當一致。另外，也可用圖形呈現每句詞數的分布（圖 2），其中摘要生成、碩士摘要文字的使用趨勢較為接近，以標題生成之摘要有較高比例在 1～3 個詞，較低比例在 4～8 個詞[8]。換言之，每句詞數可用於判斷摘要為碩士生撰寫、或是電腦生成，本文將採記每句平均字數及平均詞數、字數標準差及詞數標準差。

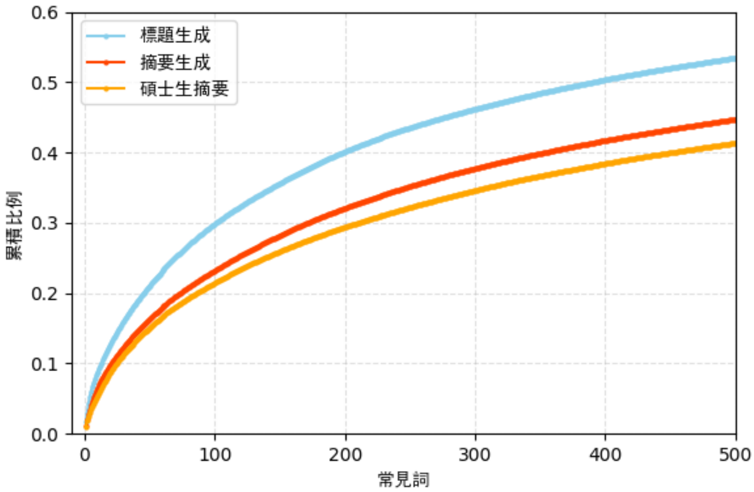

圖 1、常見詞彙的累積比例

8 三種文本的常見字彙累積機率、每句字數的圖形也與圖 1、圖 2 類似，在此省略。

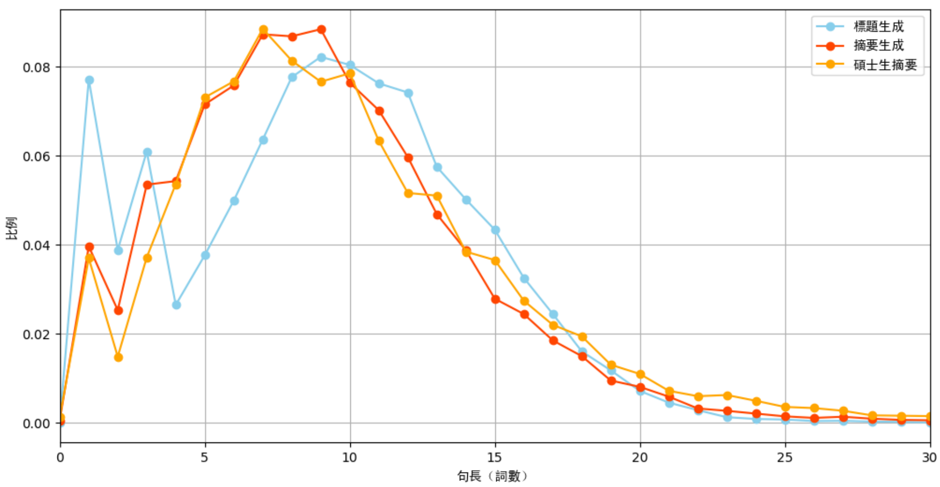

圖 2、句長（詞數）折線圖

上一節提到 TTR 和 Entropy 是反映字詞多樣性（豐富度、均勻性）的重要指標，在此以字為單位，計算隨機抽取一萬字、兩萬字，比較碩士生、電腦生成摘要的 TTR 及 Entropy 數值（表 3）。與字詞的比較結果類似，無論是隨機抽出一萬字或兩萬字，碩士生摘要有最大的 TTR，顯著高於以標題生成、以摘要生成的文本，代表電腦生成文本的豐富度較低。相同的現象也出現在 Entropy，碩士生摘要有最大的數值，顯示碩士生的用字更為均勻，不會集中在某些特定字詞。也可仿造字彙的 TTR 和 Entropy 計算，用於測量詞彙的使用特性，由於結果大致與表 3 類似，在此省略篇幅不再贅述。

表 3、三種文本的字彙 TTR 與 Entropy（平均數、95%信賴區間）

		標題生成	**摘要生成**	**碩士生摘要**
TTR	**一萬**	0.0922 (0.0919, 0.0924)	0.1061 (0.1058, 0.1063)	0.1099 (0.1096, 0.1102)
	兩萬	0.0548 (0.0547, 0.0550)	0.0638 (0.0637, 0.0640)	0.0663 (0.0662, 0.0665)
Entropy	**一萬**	8.6204 (8.6164, 8.6244)	8.9095 (8.9022, 8.9096)	8.9301 (8.9262, 8.9339)
	兩萬	8.6668 (8.6643, 8.6693)	8.9652 (8.9628, 8.9676)	8.9874 (8.9846, 8.9903)

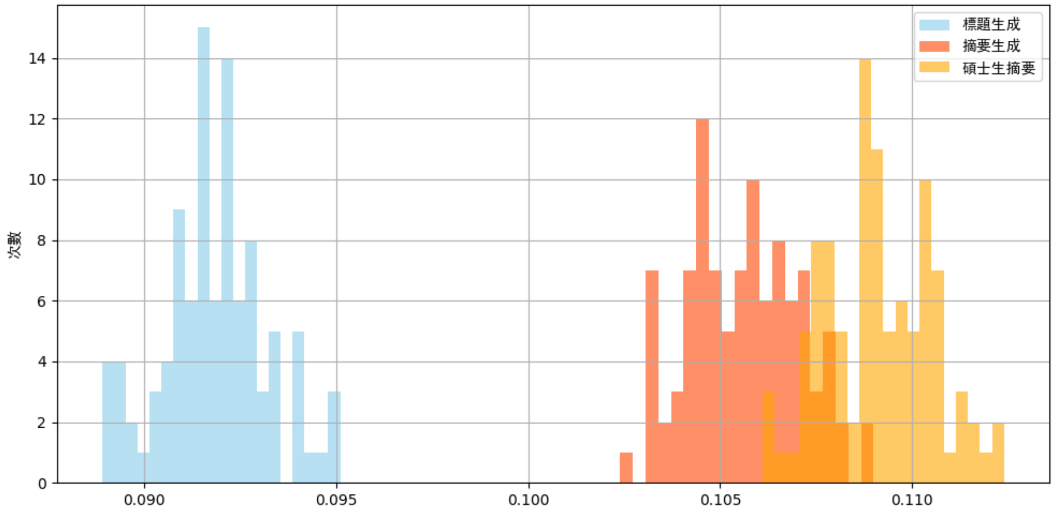

圖 3、500 篇摘要之 TTR 分布直方圖

根據上述的模擬比較，電腦生成文字的多樣性明顯較低，亦即 TTR 及 Entropy 之類的指標可用於區隔碩士生、電腦生成的摘要。不過，每篇摘要的平均字數不到 500 字，有些甚至不到 300 字，每篇摘要標準化相異字（詞）比例、Entropy 等之計算，以簡單隨機抽取 100 個字及詞，重複這個步驟 30 次，得出 TTR 及 Entropy 之平均數、標準差將作為區隔碩士生、電腦生成摘要的解釋變數。圖 3 為隨機抽取 100 字 TTR 之 30 次模擬平均數，三種摘要文本各有 500 篇摘要，以標題生成摘要之 TTR 分布最不相同，平均數明顯小於碩士生、以摘要生成之 TTR 分布，碩士生、以標題生成摘要的分布差異最為明顯，預期辨識這兩者的模型準確率應該很高。

此外，本文也考慮各文本常見詞彙，作為分辨不同文本的解釋變數。作者通常有屬於自己的寫作風格（或習慣），即便題材不同也會保留某些用字習慣，上述提到 JK 羅琳在創作偵探小說時，女性第三人稱「she」的使用量是這類型小說的兩倍，這是非常明顯的寫作密碼。先前測試 ChatGPT 的生成文字時，發現也與真人作者類似，字詞使用會有某些特定趨勢，可以視為 AI 寫作風格或用詞規則（余清祥等人，2025）。本文仿造余清祥、葉昱廷（2020）的作法，先統計三種文本最常出現的詞彙，接著篩選出現大於 50 次、且詞頻是其他文本 2 倍以上的詞彙。圖 4(a)為滿足這些條件、在標題生成、碩士生使用比例不同之詞彙，圖 4(b)則為摘要生成、碩士生出現頻率不同的詞彙，相異詞彙數值與上述結果一致，碩士生及摘要生成文本的風格比較接近，因此總數較少。ChatGPT 生成文字傾向於使用「顯示」、「進一步」及「提升」等詞，碩士生摘要則會使用「我們」口語

化用語，以及「無論、沒有、不論、不會」等表達語氣的文字。

圖 4 (a)、標題與碩士生摘要詞彙差

圖 4(b)、摘要生成與碩士生詞彙差

以 EDA 挑選之文字特徵視為解釋變數，後續將代入常見的羅吉斯迴歸、隨機森林、XGBoost 等模型，檢視分類辨識的結果，與常見大語言模型的軟體比較。限於篇幅，由於 EDA 進行程序通常耗時，而且經常需要反覆比對確保分析結果，本節並未逐一討論所有解釋變數的選取過程，以及呈現變數確實可作為區隔不同文本的依據。附錄一為經過 EDA 挑選的所有解釋變數，以及每個變數的定義與測量方式（包括如何從每篇摘要中轉換），總共考慮了 37 個變數，讀者可以根據附錄一的說明操作。

四、模型分類結果比較

本節將討論模型分類的結果，以交叉驗證（Cross-validation）的方式將 500 筆摘要分成 80%訓練集（training set）、20%測試集（testing set），而且相同標題的摘要會同時出現在訓練集（及測試集），並重複模擬 500 次計算平均準確率。分類模型的探討將分為兩個步驟說明，先比較標題生成、碩士生摘要的分析結果，接著再呈現摘要生成、碩士生摘要，除了準確率的數值之外，同時也大略說明解釋變數的貢獻程度。其中，準確率也包括精確率（Precision）、召回率（Recall）、F1 分數（F1-score），而準確率所有預測值是正確的比例，精確率是預測陽性且為真陽性的比例，召回率則是真陽性被找到的比例，F1 則是精確率與召回率的調和平均（Harmonic Average）。

先比較標題生成、碩士生摘要的分析，主要檢視本文挑選變數與分類模型結合，是否足以作為分辨人、電腦生成文字的依據，同時也用於評估各類型變數之解釋力。由於這兩種文本的差異比較明顯，因此三種模型的分類準確率都很高，

在此僅以兩種變數組合（表 4）說明分析結果，其中組合二變數個數為根據所有資料得出之結果，交叉分析中以 80%的訓練集得出之變數個數稍微少一些。僅使用字詞多樣性的準確率相對較差，隨機森林、XGBoost 的結果很接近，而羅吉斯迴歸的數值最低，在召回率、F1 指標甚至比另兩種模型低了 10%。代入常見詞等字詞使用頻率不同的變數組合二，在三種模型都有非常高的準確性，模型之間沒什麼差異，顯示以標題生成的文字和碩士生摘要的用字差異大，僅使用 31 個文字變數即能區隔[9]。我們也使用大語言模型常見的中文 BERT 模型（BERT-base-chinese）區隔標題生成、碩士生摘要，其準確率也是非常接近 100%，由於模型考慮的變數（亦即 Token）很多，將 max length 設定為 512，優化器使用 Adam 收斂條件2×10^{-5}以減少過擬合（Overfit）風險。不過，BERT 代入的變數高達兩萬餘個，包括文字、標點等，執行時間相對較長，每次模擬約為 20 分鐘，使用羅吉斯迴歸、隨機森林等模擬 500 次時間僅需 5 分鐘。

表 4、標題生成、碩士生摘要的模型分類結果

	模型	Accuracy		Precision		Recall		F1	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
組合一	LogisticRegression	0.8297	0.02	0.9344	0.02	0.7097	0.04	0.8058	0.03
	RandomForest	0.9032	0.02	0.9090	0.03	0.8971	0.03	0.9025	0.02
	XGBoost	0.9131	0.02	0.9241	0.03	0.9010	0.03	0.9120	0.02
組合二	LogisticRegression	0.9775	0.01	0.9840	0.01	0.9709	0.01	0.9773	0.01
	RandomForest	0.9837	0.01	0.9824	0.01	0.9852	0.01	0.9837	0.01
	XGBoost	0.9803	0.01	0.9775	0.01	0.9836	0.01	0.9804	0.01

註：組合一為「熵標準差、TTR 標準差、平均句長、句長標準差、短句比例、括號比例」6 個變數，組合二為「虛字、常見詞、模糊詞彙」31 個變數。

表 5、模型特徵—摘要生成、碩士生摘要

組合一	熵及 TTR 標準差、句長、句長標準差、短句比例、括號比例（6）
組合三	虛字、常見詞、模糊詞彙（29）

[9] 本文也嘗試維度縮減，透過 Boxplot 等 EDA 圖形輔助判斷，大約 18 個解釋變數的 F1 指標就能達到 0.98 以上。

組合四	統計量、虛字、常見詞、模糊詞彙（35）
組合五	21,128 個中文字及標點

接下來比較摘要生成、碩士生摘要兩個文本，考量的變數組合參考表 5，其中組合四和組合五變數個數根據所有資料得出之結果，交叉分析中以 80%的訓練集得出之變數個數稍微少一些。組合一的變數與表 4 相同，可作為對照組說明摘要生成、碩士生摘要相對不容易區隔；組合三及組合四仿造組合二，將各文本比較常見的詞彙視為解釋變數，其中組合四還加入了組合一的六個統計量，組合五為上述提到的 BERT 軟體，在此記錄為 BERT-512。摘要生成、碩士生摘要確實比較不容易區隔（表 6），將變數組合一代入三種分類模型的準確率明顯低於表 4，F1 指標大約只有 70%。即便同時考慮字詞多樣性、常見詞，各種分類準確指標僅達 85%左右，羅吉斯迴歸的數值依然是三者最低。計算量較大的大型語言模型 BERT-512，在區分摘要生成、碩士生摘要無法獲得很高的準確率，F1 分數與組合四的羅吉斯迴歸相當，但精確率與召回率確有不小出入，差距超過 15%，反觀本文提議的變數組合、三種分類模型，並沒有出現這個現象，猜測 BERT-512 有過擬合的可能。

表 6、摘要生成、碩士生摘要的模型分類結果

	模型	**Accuracy**		**Precision**		**Recall**		**F1**	
		Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**	**Mean**	**SD**
組合一	**LogisticRegression**	0.6709	0.03	0.7063	0.04	0.5887	0.05	0.6405	0.03
	RandomForest	0.7028	0.03	0.7096	0.03	0.6894	0.05	0.6984	0.03
	XGBoost	0.6997	0.03	0.7083	0.03	0.6817	0.05	0.6937	0.03
組合三	**LogisticRegression**	0.8335	0.02	0.8339	0.03	0.8320	0.04	0.8332	0.02
	RandomForest	0.8296	0.03	0.8382	0.03	0.8335	0.04	0.8301	0.02
	XGBoost	0.8323	0.02	0.8303	0.03	0.8372	0.04	0.8329	0.03
組合四	**LogisticRegression**	0.8536	0.02	0.8609	0.03	0.8449	0.03	0.8522	0.02
	RandomForest	0.8498	0.02	0.8432	0.03	0.8611	0.03	0.8514	0.02
	XGBoost	0.8405	0.02	0.8392	0.03	0.8441	0.03	0.8410	0.02

模型		Accuracy		Precision		Recall		F1	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
組合五	BERT-512	0.8841	0.07	0.9594	0.06	0.8110	0.14	0.8684	0.09

表 6 的結果顯示 BERT-512 有過擬合或不穩定的問題，在此詳細說明其中原因。首先，比較表 6 中組合四、BERT-512 的各準確率指標之平均數及標準差，兩者的 Accuracy 及 F1 平均數差別不大，但兩種指標的標準差卻有明顯差異，BERT-512 的標準差是組合四的 3 倍以上。另外，組合四所有模型的 4 種指標間之平均數及標準差相當接近，但 BERT-512 的 Precision 很高、Recall 卻明顯較低，而且這兩種指標的標準差也明顯高於組合四。如果組合四、BERT-512 的準確率指標數值有類似的分配，因為指標平均數接近，預期標準差也相近，但 BERT-512 準確率指標的標準差異常地高，猜測準確率數值可能是左偏分配。如果是左偏分配，代表 BERT-512 模型的分類結果不穩定，多數情況會有不低的準確率，但偶而會出現離譜的分類結果。

這個猜測可藉由檢視 BERT-512 的 500 次模擬結果確認（圖 5），雖然四種準確率的中位數都大於 0.8，但有不小機會出現 0.6 或甚至更低的數值，尤其是召回率及 F1 有接近或小於 0.2 的可能。透過小提琴圖（Violin Plot，圖 6）確定我們的推論，變數組合四代入羅吉斯迴歸的模擬準確率接近常態分配，而 BERT-512 的準確率卻是左偏分配，其中又以召回率數值的變異程度最大。如果以 t 檢定比較這兩類模型的四種準確率指標[10]，兩種模型在四種準確率都有顯著差異，其中組合四代入羅吉斯迴歸的 F1 準確率略低於 BERT-512 的 F1 數值，但使用較少的解釋變數及運算時間（表 6）。

[10] 由於 BERT 準確率為左偏分配，t 檢定未必合適，在此僅作為示範。

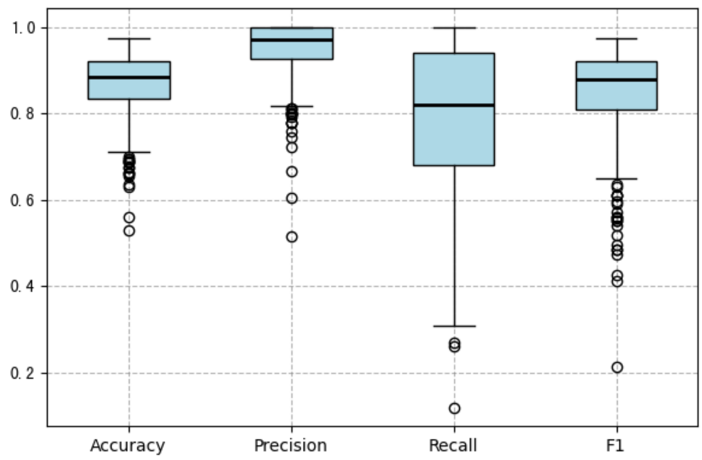

圖 5、BERT-512 的準確率指標箱型圖

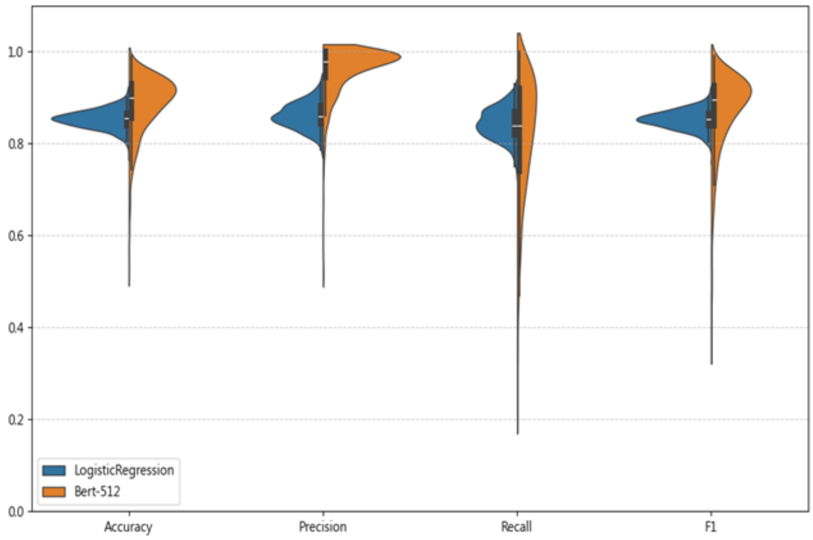

圖 6、組合四羅吉斯迴歸與 BERT-512 準確率小提琴圖

表 7、組合四羅吉斯迴歸與 BERT 成對樣本 t 檢定

Accuracy		**Precision**		**Recall**		**F1**	
t-value	**p-value**	**t-value**	**p-value**	**t-value**	**p-value**	**t-value**	**p-value**
-9.734	< 0.001	-33.743	< 0.001	5.178	< 0.001	-3.820	< 0.001

五、討論與建議

本文以臺灣經濟學門碩士論文摘要為研究依據，比較碩士生摘要與 ChatGPT 生成摘要的差異，運用探索性資料分析與統計方法，辨識真人撰寫與 AI 生成之寫作特徵。研究顯示生成的文本在語言風格、句子結構及用詞選擇上存在差異，

碩士生摘要與標題生成摘要的差異最為顯著，與摘要生成的辨識率約為 85%。研究者也可選擇 AI 工具區隔碩士生、電腦生成摘要，像是大型語言模型的 BERT 模型，得出之分類準確率與本文方法很接近，但本文使用的變數個數遠少於 AI 模型，所需運算時間也更少。而且，本文透過 EDA 篩選出具有辨識度的變數，可算是一種維度縮減的方法，篩選之變數可提供研究者挖掘探究的參考，加強實質詮釋的可能空間。

舉例而言，ChatGPT 生成摘要傾向使用較短的句子，而且句子間的長短差異較小；另外，標題生成摘要傾向於使用模糊詞藻，如「提升」、「提供」、「建議」、「此外」等，以及「穩定性」、「安全性」、「適用性」、「依賴性」等中性詞藻，這些詞語頻繁出現容易導致文章看似合理，實則空泛的描述。碩士生撰寫的摘要則經常交錯使用長短句，短句的比例明顯低於生成的摘要，字詞和標點符號的多樣性也較生成摘要來的高，而且較常使用文言文虛字「之」，以及「我們」這類口語化的字眼。不過，由於語言特色、專業領域等之不同，本文結果未必能套用至其他文本，實務應用時應鉅細靡遺地進行資料前處理、EDA 等步驟，以期找出 AI 與人類常用詞彙的差異。

我們選擇 107～109 學年經濟學類碩士論文為研究對象，原意在於減少 2022 年 ChatGPT 正式發佈後的影響，然而仍然無法完全避免 google 翻譯或其他 AI 軟體的影響。換言之，由於各種輔助寫作軟體的盛行，碩士生摘要雖然是真人創作，但也接近於某種類型的共同創作。以這種角度思考，ChatGPT 的摘要生成文字或許也是一種共同創作，AI 扮演的角色多為潤飾、修改，難怪本文方法、BERT-512 的分類準確率只有 85%。相反地，以標題生成的摘文字沒有參考碩士生摘要，更容易看出與真人撰寫的風格差異，不到 10 個變數即有 90%的準確率，30 個變數則可達到 98%。隨著 AI 工具的技術發展愈發成熟，使用者比例與日遽增，我們認為辨識人寫、電腦生成文字的困難度將更困難，不易認定何謂論文抄襲，衝擊現行對於學術論文的要求標準。

本文主要特色之一是以 Tukey 的 EDA 想法，選取與研究目標有關的重要資訊，有效減少使用變數的個數，並且大幅降低運算時間。我們認為如果能夠找到關鍵資訊及變數，未必要依賴強大的分類模型，本文研究結果支持上述的猜測，統計分析常見的羅吉斯迴歸準確率也具有競爭力，與機器學習模型的隨機森林、XGBoost 差異很小，對於不熟悉電腦科技的使用者非常方便。另外，我們重複本文的研究步驟，重新隨機抽取另外 500 篇碩士論文摘要，或是非碩博士論文的短

篇摘要，都獲得非常類似的結果，顯示 Tukey 推廣的 EDA 概念同樣適用於文字資料。自從 2010 年 IBM 公司推出大數據這個概念以來，不少人認為統計已經失去了舞台（或甚至瀕臨淘汰威脅），AI 以及電腦科技已經足以滿足數量分析的需求，本文研究主旨之一即在探討大數據時代的統計角色，事實證明統計分析仍有電腦無法取代的特色及優勢。

不過，筆者並非電腦專業，使用 ChatGPT 產生摘要時盡量保持該軟體的初始狀態，讓摘要之間的文字不會因經驗累積而有關聯，但這種設定未必符合實務需求。未來可探討不同提示語（prompt）對生成結果之影響，分析 ChatGPT 不同版本的生成模型在語言風格模仿上的異同，包括可將原始文本的篇幅分布給模型作為訓練集，讓生成的文本分布與真人撰寫更加相似。此外，本文發現 BERT-512 之類的大語言模型雖具備自動特徵學習能力，但模型穩定性不如傳統方法，或許可先用 NLP 模型分析每篇文本摘要，只保留關鍵部分，再代入分類器提升其在低資源環境中的表現。當然，除了經濟學門碩士論文外，後續研究也可納入更多領域的文本資料，檢驗本文方法是否可以廣泛應用。

參考文獻

一、中文文獻

- 何立行、余清祥、鄭文惠（2014）「從文言到白話：《新青年》雜誌語言變化統計研究」，《東亞觀念史集刊》，7，頁 427-454。
- 余清祥（1998）。「統計在紅樓夢的應用」，《國立政治大學學報》，76，頁 303-367。
- 余清祥、葉昱廷（2020）。「以文字探勘技術分析臺灣四大報文字風格」，《數位典藏與數位人文》，6，頁 69-96。
- 余清祥、許永明、徐韶汶（2025）。「TCFD 報告與文字分析」，to appear in《經濟論文》。
- 郭小東（2023）。「生成式人工智能的風險及其包容性法律治理」，《北京理工大學學報》（社會科學版），25(6)，頁 93-105。

二、英文文獻

- Archer, J. and Jockers, M.L. (2016). The Bestseller Code: Anatomy of the Blockbuster Novel(葉妍伶翻譯), St. Martin's Press.
- Breiman, L. (2001). Random Forests, *Machine Learning*, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324
- Chen, T. and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System, *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. ACM. https://doi.org/10.1145/2939672.2939785
- Efron, B. and Thisted, R. (1976). Estimating the Number of Unseen Species: How many Words did Shakespeare Know?, *Biometrika*, 63, 435–447.
- Golden, J.L., Berquist, G.F., Coleman, W.E., Golden, R., and Sproule, J.M. (2007). *The Rhetoric of Western Thought: From the Mediterranean World to the Global Setting*, 9th ed. Dubuque, IA (USA): p.67.
- Kuta, S. (2023). Art made A.I. Won a State Fair Last Year. Now the Rules are Changing, *Smithsonian Magazine*.

- Li, P.H., Fu, T.J., and Ma, W.Y. (2020). Why Attention? Analyze BiLSTM Deficiency and Its Remedies in the Case of NER, in *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(5), 8236−8244.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space, *Proceedings of Workshop at ICLR*. https://doi.org/10.48550/arXiv.1301.3781
- Shannon, C.E. (1948). "A Mathematical Theory of Communication", *Bell System Technical Journal*, 27(3), 379–423.
- Stokel-Walker, C. (2023). ChatGPT listed as Author on Research Paper: Many Scientists Disapprove, *Nature,* 613.
- Thisted, R. and Efron, B. (1987). Did Shakespeare Write a Newly-discovered Poem?, *Biometrika*, 74, 445−455.

附錄一、解釋變數列表

1. 熵與 TTR 離群值：採用簡單隨機抽樣抽取 100 個字詞，重複模擬 30 次，以 1.5 倍 IQR 判定離群值，並計算 30 次模擬中共出現幾個離群值。
2. 熵與 TTR 標準差：同樣採上述簡單隨機抽樣的方法計算得出熵 TTR 的標準差。
3. 平均句長：以逗號（，）、句號（。）、分號（；）、驚嘆號（！）與問號（？）等五種全形標點符號進行斷句，計算出標題生成與碩士生摘要的平均句長。
4. 句長標準差：以句長的五種全形標點符號斷句，計算文本的句長標準差。
5. 短句比例：定義短句為字數小於 6 字，計算每篇摘要出現短句的比例。
6. 括號使用比例：碩士生摘要傾向使用半形括號。
7. 虛詞：「與」、「之」、「應」、「所」及「了」等五字出現頻率。
8. 常見詞：計算兩文本詞頻比例相差兩倍以上、且兩文本合計出現 100 次以上的詞彙（共計 20 個常見詞）。
9. 模糊性詞彙：生成摘要更容易出現安全性、針對性及透明性等模糊用詞，參照常見詞篩選方法，得到「進一步」、「競爭力」、「穩定性」、「透明度」、「報酬率」以及「房地產」等 6 個詞彙。