中原大學

資訊工程學系

專題研究報告

題目

AI 生成文本判別

指導教授：莊秀敏

學生：11127229 廖翊崴

目錄

圖目錄

表目錄

壹、緒論

一、摘要

本研究透過深度學習的方法，採用 BERT 預訓練語言模型進行微調，建立一套可辨識 AI 生成與人類撰寫文本的辨識系統。在資料處理上加入了詞性標註（POS tagging）、依存句法分析（Dependency Parsing）以及 TF-IDF 關鍵詞萃取，藉此強化模型對語言結構與語意特徵的理解能力。將訓練完成的模型部署到後端系統，並搭配設計簡潔、操作直觀的前端介面，使用者可直接上傳文本進行檢測，系統能夠輸出整體判斷結果，並標示出可能為 AI 生成的句子，另外提供了一個調節器，依使用者需求做更精確的標示，透過計算每個詞元（Token）對此句被判別為 AI 生成的貢獻度作為可修改的地方建議，給使用者一個具體且可視化的分析參考。

二、研究動機

由於 AI 生成文本的快速發展與普及，許多文章充斥著 GPT 生成的痕跡，雖然現在生成文本與人類創作文本之間的界線日益模糊，但仍可透過某些語言特徵、句構和風格差異來發現一些蛛絲馬跡，本研究希望透過深度學習的方式，訓練出一個有能力分辨的 AI。

三、研究貢獻

提出一個可有效區分 AI 生成與人類撰寫文本的深度學習模型，該模型將應用於一般文章檢測場景，像是網路文章、新聞報導、學生心得作業等，為了因應現實中 AI 生成內容與人類撰寫內容混雜於同一篇文章的情況，模型設計聚焦於句子和詞元（Token）層級的精準標示，提升偵測粒度與實用性，同時研究也將開發一套操作簡便、介面清晰的前端系統，讓非技術使用者亦能輕鬆操作與理解分析結果，提升本研究在內容審查領域的應用價值。

貳、文獻探討

隨著大型語言模型（Large Language Models, LLMs）如 GPT 系列的快速發展，AI 生成文本（AI-generated text）的品質已逐漸逼近人類撰寫水平，促使辨識生成文本成為一項關鍵課題。Ippolito 等人[1]指出，雖然某些 decoding 策略（如 top-k sampling）會生成更具欺騙性的文本，容易誤導人類，但反而更容易被自動分類器辨識，顯示人與模型在辨識線索上的差異；而當生成策略如 nucleus sampling 平衡了多樣性與流暢性，則更難被兩者辨識，凸顯辨識任務的挑戰性。在另一方面 Jawahar 等人[2]對生成文本偵測方法進行系統性綜述，強調預訓練語言模型（如 BERT, RoBERTa）雖然在二分類判別任務上展現優勢，但因受限於訓練資料導致其泛化能力差，並指出提升人機協作的辨識能力需依賴更多語意與事實層面的分析輔助。

綜合上述研究可知，目前辨識 AI 生成文本的方法多聚焦於模型風格與統計特徵的差異，但在語意推理、語法結構辨識與實際應用場景上仍具發展空間，因此本研究在模型架構設計上，除了延續使用 BERT 預訓練模型微調外，更進一步引入詞性標註、依存句法分析與 TF-IDF 特徵，期望能提升模型對語言結構與語意線索的理解力，並透過句子層級的偵測與可視化回饋，達成更精準且具實用性的 AI 生成內容辨識系統。

參、研究方法

一、資料處理

本研究的非 AI 生成文本資料主要來自 2010 年以前的網路文章與報導，由於從 2017 年開始 AI 文字生成工具（如 2017 年推出的 QuillBot）已逐漸成熟並被廣泛應用於語言改寫與輔助寫作，因此我們將 2010 年作為可信非 AI 生成資料的時間界線。

為了蒐集非 AI 文本，我們在 Google 搜尋引擎中使用「關鍵詞＋ before:2010」的格式（例如：「氣候暖化 before:2010」、「大腸癌 before:2010」）進行查找，找到符合時間條件的文章後，點入連結檢視內文，若確認為目標內容，便將其下載或複製並進行處理。文章整理過程包括移除所有空格與換行，並以句號作為主要分割依據，為了避免誤切分且為追求統一，我們排除了句號後接引號（如「……。」）或括號（如（……。））的情況，使用的 Python 正則式如下：sentences = re.split(r'(?<=。)(?![」|）)', text)，整理後的文章會依句子切分，每個句子標記為 0（表示非 AI 生成）存入 CSV 檔案中，其格式如表一，並在第一個句子所在列的第三行標上文本出處，如圖一。

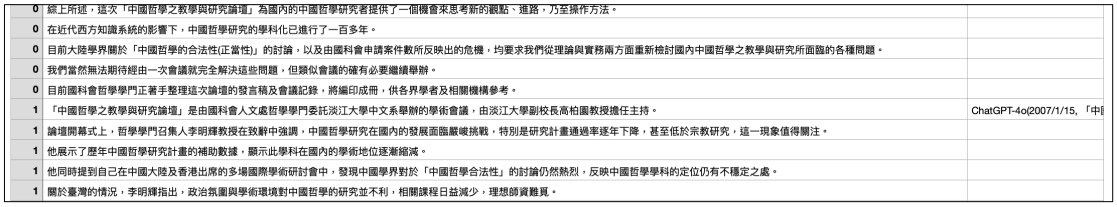
標記	文本	文本出處
0 or 1	一個句子	年份、來源、哪個工具生成的
……	……	……

表一

1	此外，永續發展必須建立在環境保護、經濟發展和社會正義三大基礎之上。	
1	專家們也提醒，在建立新指標時，必須先確立目標，避免因加入非經濟因素而產生干擾。	
1	國際指標不一定完全適用於台灣，因此需要根據台灣的特殊情況來制定合適的指標。	
1	同時，如何讓一般民眾理解這些指標也是一個重要課題，以避免錯誤解讀。	
0	「中國哲學之教學與研究論壇」是由國科會人文處哲學學門委託淡江大學中文系承辦的會議，主持人是副校長高柏園教授。	2007/1/15,「中國哲學之教學
0	哲學學門召集人李明輝教授於開幕致辭中說明論壇之宗旨。	
0	他首先指出:「近年來國科會的研究計畫申請案，中國哲學申請案的通過率有逐年下降的趨勢，去年甚至落後於宗教研究的申請案，實為值得警惕的現象。	
0	」他藉由圖表比較歷年獲得補助的申請案及其分配情形，顯示國內的中國哲學研究正在萎縮中。	
0	其次，他略述他於去年十二月在廣州中山大學及汕頭大學出席「西方哲學東漸與中國社會現代化國際學術研討會」，在深圳大學出席「『中國哲學』建構的當代反思與未來前瞻國際學術研討會」，以及在香港中文大學出席「中國哲學研究方	
0	他指出:「最近大陸學界關於「中國哲學的合法性(正當性)」的討論顯示中國哲學作為一門學科的地位尚未穩定。	

圖一

接著將這些所找到的非 AI 文章輸入至各類 GPT（Generative Pre-trained Transformer）模型中生成相似內容，以產出 AI 生成的對應文本。由於不同 GPT 模型在單次生成的最大字數上有所限制，若原始文章過長，我們會將其切分為適當長度再分批輸入。使用的 Prompt 格式為：「整理以下並生成相似文章，不要有標題和列點，x 字左右：」，其中的 x 為輸入文本的字數，目的是產出與原文篇幅相近的生成內容，以保持非 AI 與 AI 文本的總句數分布相對均衡。最終生成的 AI 文本亦經相同的整理與句子切分流程，並將每句標記為 1（表示 AI 生成），來源同樣記錄在第一句的第三欄，如圖二所示。



0	綜上所述，這次「中國哲學之教學與研究論壇」為國內的中國哲學研究者提供了一個機會來思考新的觀點、進路，乃至操作方法。	
0	在近代西方知識系統的影響下，中國哲學研究的學科化已進行了一百多年。	
0	目前大陸學界關於「中國哲學的合法性(正當性)」的討論，以及由國科會申請案件數所反映出的危機，均要求我們從理論與實務兩方面重新檢討國內中國哲學之教學與研究所面臨的各種問題。	
0	我們當然無法期待經由一次會議就完全解決這些問題，但類似會議的確有必要繼續舉辦。	
0	目前國科會哲學學門正著手整理這次論壇的發言稿及會議記錄，將編印成冊，供各界學者及相關機構參考。	
1	「中國哲學之教學與研究論壇」是由國科會人文處哲學學門委託淡江大學中文系舉辦的學術會議，由淡江大學副校長高柏園教授擔任主持。	ChatGPT-4o(2007/1/15, 「中
1	論壇開幕式上，哲學學門召集人李明輝教授在致辭中強調，中國哲學研究在國內的發展面臨嚴峻挑戰，特別是研究計畫通過率逐年下降，甚至低於宗教研究，這一現象值得關注。	
1	他展示了歷年中國哲學研究計畫的補助數據，顯示此學科在國內的學術地位逐漸縮減。	
1	他同時提到自己在中國大陸及香港出席的多場國際學術研討會中，發現中國學界對於「中國哲學合法性」的討論仍然熱烈，反映中國哲學學科的定位仍有不穩定之處。	
1	關於臺灣的情況，李明輝指出，政治氛圍與學術環境對中國哲學的研究並不利，相關課程日益減少，理想師資難覓。	

圖二

最終本研究的總資料量為以下表二，單位為筆，以句號做切分，一個句子為一筆，另外在訓練集和測試集上我們以 8:2 做分割，在類別上加入了分層抽樣，以確保訓練集與測試集中，各類別的比例和整體資料集中是一致的，分割參數：test_size=0.2, random_state=1, stratify=fb_label_df['labels']，詳細資料量如表三。

總資料量	4054 筆
0（非 AI 文本）	1973 筆
1（AI 文本）	2081 筆

表二

	0（非 AI 文本）	1（AI 文本）	總和
訓練集	1578 筆	1665 筆	3243 筆
測試集	395 筆	416 筆	811 筆

表三

最後在處理好的資料上加入了詞性標註（POS tagging）、依存句法分析（Dependency Parsing）以及 TF-IDF 關鍵詞萃取。

我們採用了 CKIP 的 CkipPosTagger()工具對文本進行詞性標註（POS tagging），CkipPosTagger()為中央研究院中文處理小組（CKIP）所開發之工具，CkipPosTagger 是 CKIP Lab 釋出的 Python 介面套件 CKIPTransformers 中的模組之一，採用 BERT-based 深度學習模型，專為繁體中文語料設計，能夠辨識詞語間的詞性資訊，用以強化模型對詞性的特徵學習。另外在依存句法分析（Dependency Parsing）的部分用了 Stanford NLP 團隊於 2020 年發表的 Stanza 套件，Stanza[3]是一個基於神經網路架構的多語言自然語言處理工具包，涵蓋詞性標註、命名實體識別及依存句法分析等多項任務，該工具有高準確度與良好擴展性，其特別針對中文語料展現了優異的表現，在中文依存句法分析方面，Stanza 利用深度雙向 LSTM 結合多層注意力機制，有效捕捉中文語法特性，顯著提升了句法分析的精準度，藉此用以強化文本句構。儘管 BERT 語言模型已被證實能在其內部隱含地學習語言結構資訊[4]，然而有研究指出，在某些語言理解任務中，明確提供詞性與語法特徵仍可對模型效能帶來正面影響[5]。

為了進一步強化模型對文本的理解，我們導入了 TF-IDF（Term Frequency–Inverse Document Frequency）[6]關鍵詞萃取技術，此方法可凸顯具代表性的關鍵詞，進而提供模型額外的線索。在實際觀察中，我們發現人類撰寫的文本詞彙使用較為多變，詞彙分佈相對均勻，反之，AI 生成的文本常重複使用特定字詞，藉此可作為參考判斷依據。

因此，本研究將三者作為附加的語言特徵，期望能提升模型對 AI 生成文本與人類撰寫文本間語言特徵差異的辨識能力。

二、模型設計

本研究使用中研院語言學研究所 CKIP Lab 釋出的繁體中文 BERT 模型[ckiplab/bert-base-chinese][7]作為文本分類模型的語言理解基礎，該模型基於 BERT[8]架構，使用大量繁體中文文本進行預訓練，能有效擷取中文語意中的上下文關係，由於市面上多數中文 BERT 模型以簡體語料為主，此模型提供繁體中文場景下更準確的語言表示能力，更能適配實際應用。

三、介面實作

前端介面如下圖所示。

圖三

後端的部分除了載入模型執行 AI 生成文本的預測任務外，當文本被判別為 AI 生成時，會進一步計算每個詞元（Token）對此分類結果的貢獻度。本研究採用集成梯度法（Integrated Gradients）作為詞級貢獻度的計算基礎，該方法由 Sundararajan 等人提出[9]，是一種具備理論公設基礎的特徵歸因技術，能為深度神經網路的輸入特徵提供可解釋性分數。具體而言，本研究將 Integrated

Gradients 應用於 BERT 模型的嵌入層（embeddings），透過計算嵌入向量與 baseline（全零向量）之間的積分路徑梯度，估算每個詞元 token 對最終分類結果的貢獻度。在實作層面上，我們使用 Captum 套件實現集成梯度計算，指定模型輸出為 softmax 後第二類別（即 AI 生成文本）的機率，並搭配 attention mask 作為額外輸入以保留 BERT 的注意力邏輯，最後每個詞元的重要性分數由該詞元在所有嵌入維度上的歸因結果進行總和計算獲得。

肆、結果與討論

以下爲訓練資料的原始文本與經過詞性標註（POS tagging）、依存句法分析（Dependency Parsing）及 TF-IDF 關鍵詞處理後之文本的訓練結果比較，此結果與 Glavaš 等人[5]的研究論述一致，加入額外的詞性與語法特徵仍可對模型效能帶來正面影響。透過觀察表四能發現每項指標皆有進步，並且除了泛化能力，其他指標的數值表現均不錯。

	mcc	acc	F1	F1-P (1)	F1-N (0)	eval_loss
原本	0.61623	0.80025	0.79585	0.82581	0.76590	1.4733
處理	0.64194	0.81504	0.81189	0.83624	0.78754	1.4037
增減	+ 0.026	+ 0.015	+ 0.016	+ 0.010	+ 0.022	- 0.070

表四

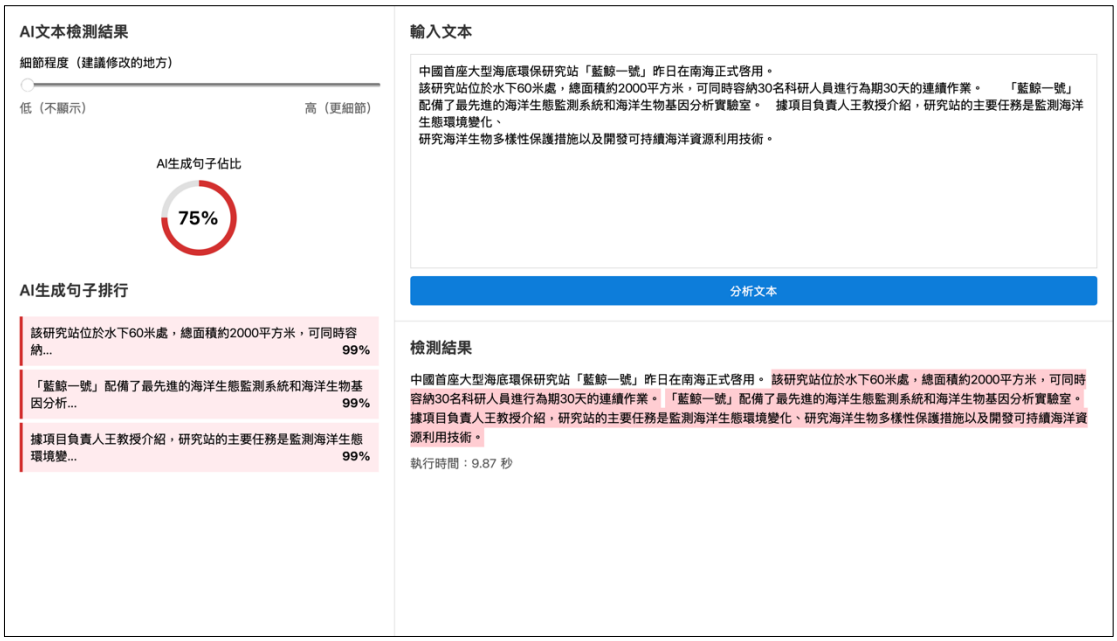
為了確認是誤差還是有優化，下圖呈現原始資料和經處理的資料各進行 8 次訓練所得準確率（Accuracy）的雙尾檢定（Two-Tailed Test）結果，結果明顯是有優化的。

```python
import numpy as np
from scipy import stats

before = np.array([0.779, 0.805, 0.790, 0.793, 0.798, 0.789, 0.793, 0.778])  # 優化前
after = np.array([0.803, 0.806, 0.809, 0.807, 0.815, 0.795, 0.799, 0.813])   # 優化後

t_stat, p_value = stats.ttest_rel(after, before)

print(f"t-statistic: {t_stat:.4f}")
print(f"p-value: {p_value:.4f}")

alpha = 0.05
if p_value < alpha:
    print("有優化")
else:
    print("無優化")
```

```
t-statistic: 3.8926
p-value: 0.0060
有優化
```

圖四

以下為介面實際運作情況，測試文本為 ChatGPT-4o 隨機生成的新聞內容。圖五顯示在未調整細節程度的情況下系統輸出的結果，被判定為 AI 生成的句子會被整句標示，同時也展示了當使用者輸入含空格與換行的文本時，系統整理與處理的效果。另外圖六為調整細節程度後的輸出結果，細節顯示建議調整的地方，而在圖七中放入了修改後的文本，修改建議調整的地方後偵測結果為 0%。



圖五


AI文本檢測結果
細節程度（建議修改的地方）
低（不顯示）
高（更細節）
AI生成句子佔比
75%
AI生成句子排行
該研究站位於水下60米處，總面積約2000平方米，可同時容納...
99%
「藍鯨一號」配備了最先進的海洋生態監測系統和海洋生物基因分析...
99%
據項目負責人王教授介紹，研究站的主要任務是監測海洋生態環境變...
99%
輸入文本
中國首座大型海底環保研究站「藍鯨一號」昨日在南海正式啓用。該研究站位於水下60米處，總面積約2000平方米，可同時容納30名科研人員進行為期30天的連續作業。「藍鯨一號」配備了最先進的海洋生態監測系統和海洋生物基因分析實驗室。據項目負責人王教授介紹，研究站的主要任務是監測海洋生態環境變化、研究海洋生物多樣性保護措施以及開發可持續海洋資源利用技術。
分析文本
檢測結果
中國首座大型海底環保研究站「藍鯨一號」昨日在南海正式啓用。 該研究站位於水下60米處，總面積約2000平方米，可同時容納30名科研人員進行為期30天的連續作業。 「藍鯨一號」配備了最先進的海洋生態監測系統和海洋生物基因分析實驗室。據項目負責人王教授介紹，研究站的主要任務是監測海洋生態環境變化、研究海洋生物多樣性保護措施以及開發可持續海洋資源利用技術。
執行時間：12.81 秒

圖六


AI文本檢測結果
細節程度（建議修改的地方）
低（不顯示）
高（更細節）
AI生成句子佔比
0%
AI生成句子排行
輸入文本
中國首座大型海底環保研究站「藍鯨一號」昨日在南海正式啓用，此研究站在水下60米處，總面積約2000平方米，可同時容納30名科研人員進行30天的連續作業。其配備了先進的海洋生態監測系統和海洋生物基因分析實驗室，根據項目負責人王教授介紹，此研究站的主要任務是監測海洋生態的環境變化、研究海洋生物多樣性的保護措施和開發海洋資源永續利用技術。
分析文本
檢測結果
中國首座大型海底環保研究站「藍鯨一號」昨日在南海正式啓用，此研究站在水下60米處，總面積約2000平方米，可同時容納30名科研人員進行30天的連續作業。 其配備了先進的海洋生態監測系統和海洋生物基因分析實驗室，根據項目負責人王教授介紹，此研究站的主要任務是監測海洋生態的環境變化、研究海洋生物多樣性的保護措施和開發海洋資源永續利用技術。
執行時間：14.25 秒

圖七

伍、參考文獻

[1] D. Ippolito, D. Duckworth, C. Callison-Burch, and D. Eck, “Automatic detection of generated text is easiest when humans are fooled,” in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 1808–1822.

[2] G. Jawahar, M. Abdul-Mageed, and L. V. S. Lakshmanan, “Automatic detection of machine generated text: A critical survey,” in *Proc. 28th Int. Conf. Computational Linguistics (COLING)*, Barcelona, Spain (Online), Dec. 8–13, 2020, pp. 2296–2309.

[3] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza : A Python Natural Language Processing Toolkit for Many Human Languages,” in *Proc. 58th Annual Meeting of the Association for Computational Linguistics (ACL)* , Jul. 5–Jul.. 10, 2020, pp. 101–108.

[4] I. Tenney, D. Das, and E. Pavlick, “BERT rediscovers the classical NLP pipeline,” in *Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, Florence, Italy, Jul. 28–Aug. 2, 2019, pp. 4593–4601.

[5] G. Glavaš and I. Vulić, “Is supervised syntactic parsing beneficial for language understanding tasks? An empirical investigation,” in *Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, Online, Apr. 2021, pp. 3090–3104.

[6] K. Sparck Jones, “A statistical interpretation of term specificity and its application in retrieval,” *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[7] CKIP Lab, “ckiplab/bert-base-chinese,” Hugging Face, 2021. [Online]. Available: https://huggingface.co/ckiplab/bert-base-chinese

[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” *Proc. NAACL-HLT*, pp. 4171–4186, 2019.

[9] M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic attribution for deep networks,” in *Proc. 34th Int. Conf. Machine Learning (ICML)*, Sydney, Australia, Aug. 6–11, 2017, pp. 3319–3328.