2025 年【科學探究競賽-這樣教我就懂】

普高組 成果報告表單

題目名稱：假評論終結者-LSTM 與爬蟲合成的最終兵器

一、摘要

本研究結合網路爬蟲與 LSTM 模型，開發出一款 Google 插件，供使用者透過文字分析評論內容。藉由該程式，用戶能輕易過濾假評價，統整大量評論意見，創造更透明有效率的購物環境。

二、探究題目與動機

在現今數位時代，網路評價已成為消費者決策的重要參考依據。然而，現有的評分機制大多採用一到五顆星的簡單量化方式，這種方式往往無法準確反映使用者的真實體驗，甚至可能出現評分與內容不符的情況。例如，一則四星評論可能實際上充滿負面意見，而某些高分評價則可能來自機器人或非真實使用者，導致我們消費者難以獲得準確資訊，商家也無法獲取有價值的回饋。

為了解決這個問題，我們運用人工智慧技術，結合文字情感分析，開發一套更精確的評論評估系統。我們設計的系統將透過深入分析評論內容的情緒與語意，提供比傳統評分機制更具代表性與參考價值的評估結果。這不僅有助於消費者做出更明智的購買決策，也能讓商家獲得更具洞察力的產品反饋，並且提升市場透明度與產品品質。

三、探究目的與假設

（一）探究目的

本實作的研究目標是開發一款 Google 擴充功能，協助使用者迅速分析評論，以提升對商品評價的判斷效率與準確性。此外，透過此實作，我們期望學習文本情感分析的基礎概念，並深入探索如何運用人工智慧技術來建構情感分析專案。同時，我們也將學習 Google 擴充功能的開發技術與完整流程，包括設計、功能實現及測試，藉此累積實作經驗，為未來相關專案奠定穩固基礎。

（二）探究假設

本實作假設程式能夠正常運作並成功抓取網頁內容，進行精準的評論分析，為使用者進行資料的整理以及篩選。

四、探究方法與驗證步驟

（一）探究方法

1、文獻分析法：本研究透過網路資料蒐集與文獻閱讀，讓我們了解當前情緒辨識技術的發展現況、所使用的模型架構，以及爬蟲技術的最新進展。我們分析不同情緒辨識模型的特點與應用，並評估爬蟲技術在資料擷取與處理上的發展程度，期望提供更全面的理解與參考。

2、實作研究法：研究以 LSTM 模型為核心，結合網路爬蟲技術，開發一款具備直觀

且簡潔介面的 Google 插件，讓使用者能夠輕鬆運用該插件進行直觀的評論分析。

（二）驗證步驟

1、收集訓練資料

我們使用 Python 的 requests 函式庫撰寫爬蟲程式，從博客來網站擷取數據。爬蟲的核心功能包括解析網站內容，篩選出評論區塊，並透過標籤識別進行分類。最終，整理後的數據會以 CSV（逗號分隔值）格式儲存，方便後續分析與處理。

```
55      # 從網站擷取評論數據
56      headers = {"User-Agent": "Mozilla/5.0 (Macintosh; Intel Mac OS X 10_14_4) AppleWebKit/537.36 (KHTML, like
    Gecko) Chrome/74.0.3729.169 Safari/537.36"}
57      product_id = str(url.split('/products/')[1].split('?')[0])
58      comments_url = f"https://www.books.com.tw/booksComment/getCommemt/{product_id}"
59      res = requests.get(comments_url, headers=headers)
60      soup = BeautifulSoup(res.text, 'html5lib')
61      comments = soup.find_all("span", calss="comment-content")
62
63      # 以CSV檔保存評論
64      csv_output = io.StringIO()
65      csv_writer = csv.writer(csv_output)
66      for comment in comments:
67          csv_writer.writerow([comment.getText().strip()])
68
69      csv_output.seek(0)
70      df = pd.read_csv(io.StringIO(csv_output.getvalue()), header=None)
71      comment_list = df[0].tolist()
```

圖一、爬蟲程式碼

2、清洗資料

我們選用 Jieba 作為中文分詞工具。由於中文文本不像英文單詞那樣具有明確的界限，因此分詞成為模型建構過程中的關鍵步驟。Jieba 能夠將長篇中文評論拆解為具備語義意義的詞彙，並提供可擴充的中文停用詞庫。在人工標註數據時，會將標點符號與常見但無情感意義的詞語納入停用詞列表，以提升模型的辨識準確度。如下圖二所示，顯示導入停用詞語料庫的截圖。

```
#  停用詞載入
def load_stopwords(file_path):
    with open(file_path, 'r', encoding='utf-8') as file:
        stopWords = set(line.strip() for line in file)
    return stopWords
```

圖二、導入停用詞程式碼

為了確保評論內容在資料處理過程中的一致性，並避免簡繁體混用，我們使用 OpenCC（Open Chinese Convert）工具對所有 CSV 檔案進行文字轉換，確保內容統一為繁體中文。

此外，我們將評論的標註標準定義為 0 代表正面評論，1 代表負面評論，並透過人工方式為爬取的數據標記情感分類，完成資料準備階段。

3、訓練模型

我們選擇 LSTM 模型作為核心架構，其主要優勢在於能夠保留長時間的依賴關係，因此特別適用於文字分析、語音辨識等涉及長篇內容的任務。尤其是在處理較長或情感較為複雜的評論時，例如書評，其中可能同時包含正面與負面的評價詞彙，LSTM 能夠更精確地理解並分析這類內容。

在模型建構方面，我們設定輸出維度（output_dim）為 128，並配置 256 個神經元。當模型建置完成後，即進行訓練，並將訓練內容的最大長度限制為 300 個字，以避免內容過長影響模型效能。

```
def predict_sentiment(texts):

processed_texts = [list(jieba.cut(text, cut_all=True)) for text in texts]
processed_texts = [list(filter(lambda a: a not in stopWords, content)) for content in processed_texts]


sequences = token.texts_to_sequences(processed_texts)
sequences = sequence.pad_sequences(sequences, maxlen=300)


predictions = model.predict(sequences)

return predictions
```

圖三、LSTM 參數設定值

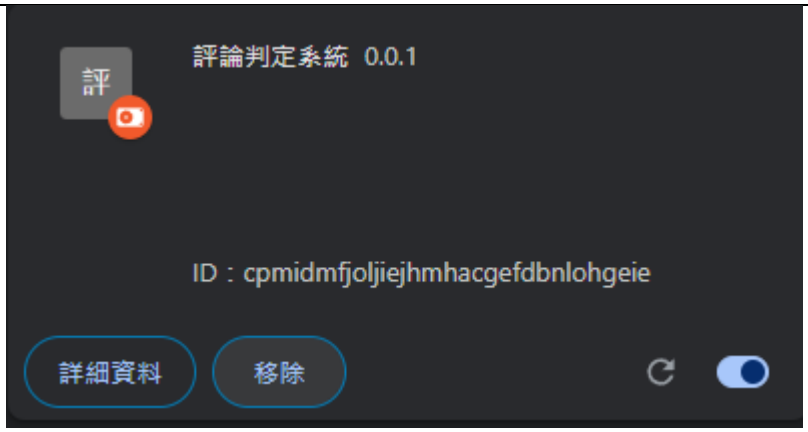
模型的輸出結果包含 Comment 和 Sentiment Score 兩個部分。其中，Comment 為爬蟲程式擷取的網站評論內容，而 Sentiment Score 則是模型進行情感分析後的評分結果。當評論表達正面情緒時，分數越趨近 0；當情緒較為負面時，分數則接近 1；若評論未明顯呈現正面或負面的情感詞彙，則分數會落在中間區間。

4、製作成 google 插件

為了提升使用者的便利性，我們最終選擇以 Google Chrome 擴充功能 來呈現模型。開發 Chrome 擴充功能時，主要使用 HTML、CSS 和 JavaScript，並需包含一個 manifest.json 檔案。該檔案的作用是讓 Chrome 瀏覽器識別並執行此擴充功能，使其能夠與瀏覽器互動，進而實現所需的功能。

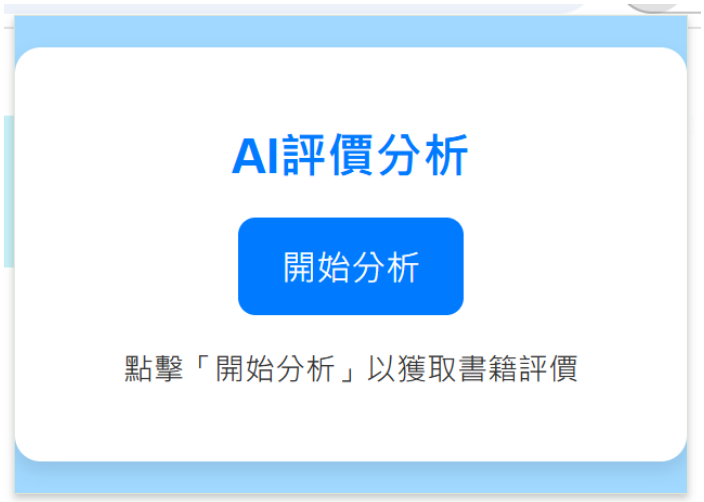
在專案的檔案結構中，popup/index.html 被定義為擴充功能的彈出視窗，因此我們需建立一個名為 popup 的資料夾，並在其中建立 index.html 來設計視窗內容，搭配 style.css 進行樣式設定，以確保視覺呈現的一致性與使用體驗的最佳化。

完成所有必要的檔案後，開啟 Google Chrome 並進入 擴充功能管理頁面，接著啟用 開發人員模式。啟用後，點擊 「載入未封裝項目」，選擇包含擴充功能的資料夾並載入。完 成這些步驟後，擴充功能即可順利運行並開始使用。



圖四、插件啟用畫面

這個程式的主要功能可以分為三個部分：首先，提供一個直觀且易於操作的 UI 介面，使用者能夠輕鬆地與程式進行互動；其次，程式能夠擷取當前網頁的網址，並依據網頁內容進行處理；最後，程式會接收並顯示從模型回傳的數據，讓使用者能夠清楚地了解模型的分析結果或建議。



圖五、UI 介面

五、結論與生活應用

（一）結論

本次實作訓練的模型可分辨情緒。透過與文字辨識模型的結合，該技術可延伸到其他平台如 Google Map、Uber eats 等，幫助使用者得到真實評價，令商家能夠統整用戶的反饋。在問卷調查方面，也可以運用情緒分析處理填表者撰寫的內容，讓評分結果更精準，而非單單依賴 1~5 星的按鈕。

並且在本次研究中，我們所使用的模型在精確度方面尚未達到理想水準，顯示參數設定與結構仍有進一步優化的空間。首先，模型的層數與每層節點數可能需要調整，以更符合數據特性並提升學習能力。此外，增加高品質的訓練數據，有助於改善模型的表現，尤其是在現有數據不足以涵蓋目標任務的多樣性時。另一方面，資料預處理階段採取更精細的停用詞處理策略將能有效去除無關資訊並突顯關鍵特徵。未來我們可進一步優化這些策略，以提升模型的準確性與穩定性，從而建立更高效、更精確的模型。

（二）生活應用

本專案透過 AI 技術，提供更精準的商品評價分析，有助於消費者在網路購物時快速篩選出有價值的意見，做出更明智的決策。其生活應用包含：在選購商品時，快速了解商品評價趨勢，節省瀏覽大量評論的時間；協助消費者判斷網路評論的真偽，避免受到不實資訊誤導；提升購物決策效率，並創造更透明的購物環境。

（三）SWOT 分析

（四）生活應用 **Strengths 優勢**	**Weakness 劣勢**
1.提供博客來使用者快速找好書的方式。 2.介面簡潔，使用者容易上手。	1.安裝流程對部分用戶來說仍有困難。 2.此專案只能在博客來擷取評論並分析，無法適用於任意網站。
Opportunities 機會	**Threats 威脅**
1.若有更合適的模型出現，情緒辨識 AI 的效率與準確率皆可能進一步上升。 2.該模型可延伸擴充至其他網站，如社群媒體、外送平台、購物網站等。	1.網站的反爬蟲程式不斷改進，爬蟲功能可能因此無法使用或需要降低爬取速度。

參考資料

1、 楊宗恩 (2017)。以 LSTM 深度神經網路語言模型建構英文課程重點摘要。國立屏東科技大學資訊管理學系碩士班：碩士論文。

2、 Wai-kin Wong, Wang-chun Woo. (2015, September 19) Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting.

3、 楊杰淮 (2022)。網絡爬蟲與反爬蟲相關研究。明新科技大學電機工程系碩士班:學術論文

4、 Trupti V. Udapure, Ravindra D. Kale, Rajesh C. Dharmik(2014) Study of Web Crawler and its Different Types. IOSR Journal of Computer Engineering (IOSR-JCE) ,Volume 16, Issue 1, Ver. VI (Feb. 2014), PP 01-05

5、 Graves, A. (2012, January 1). *Long Short-Term Memory*. Springer Nature Link.

6、 houdt, G. V. (2020, May 13). *A Review on the Long Short-Term Memory Model*. Springer Nature Link.