NLP期末報告

利用NLP分析”新冠肺炎文章/新聞”

作者： 張子中 陳旻凱 林茂宸
指導老師： 廖元勳

摘要

近年來**Covid-19**肆虐全球，死亡人數日漸攀升，各地發展情況不同出現資訊不對稱以及醫療資源分配不均，結合疫情實際大數據推測出該地區大眾恐慌程度。運用ptt-web-crawler爬出JSON檔，再用Json to txt，將文章分成正面和負面，使用bert-base-chinese預訓練模型，再以資料集進行微調(fine-tuning)訓練下游任務，並使用損失函數、混淆矩陣、正確率和 F1指標評估，結果訓練資料少的情況下，BERT依然擁有十分良好的表現。

中華民國 一一一年一月

圖目錄

表目錄

目錄

壹、 前言

1-1 研究動機

近年來**Covid-19**肆虐全球，死亡人數日漸攀升，各地發展情況不同出現資訊不對稱以及醫療資源分配不均，導致出現過度恐慌、囤積物資或是沒有讓疫情重災區最優先獲得足夠的資源。

1-2 研究目的

透過**NLP**分析文章及社群留言了解實際疫情與民眾情緒的相關性。

1-3 解決問題

希望解決疫情實際情況與大眾的資訊不對等的問題，降低恐慌以利政府調配可使用資源，達到最有效的利用。

1-4 技術方法

以網路爬蟲瀏覽數上所有文章，再以**NLP**分析社群網站文章及留言尋找**COVID-19**相關關鍵字過濾網頁，並且以情感分析判斷發文者當下情緒，結合疫情實際大數據推測出該地區大眾恐慌程度。

貳、 文獻探討

2-1 自然語言處理

自然語言處理(**N**atural **L**anguage **P**rocessing,**NLP**)，主要是透過人工智慧與語言學利用電腦將語言變成變成有意義的符號或是關係等等，在對數據進行統計分析。自然語言處理有些常見的機制與方法以下分別介紹TF-IDF、RNN、LSTM與Transformer。

1.TF-IDF(**t**erm **f**requency–**i**nverse **d**ocument **f**requency)
常用於資料檢索與文字挖掘的技術，tf-idf是一種統計方法，用來評估在這一詞在檔案集或語料庫的重要程度，字詞的重要性受到出現次數影響，字詞出現頻率愈高越重要，由tf-idf加權形式常被當作搜尋引擎使用。

2.RNN(循環神經網路Recurrent Neural Network)適合應用於序列資訊的處理，而每個樣本和之前存在的樣本會有關聯，也就是說RNN能夠額外考慮上下文的關係，提升預測準確率。

3.LSTM(長短時記憶網路Long Short-Term Memory)論文首次發表於1997年。適合用於處理和預測時間序列中延遲和間隔非常長的事件。LSTM是一種含有LSTM區塊或其他的類神經網路，因為它的特性，在文獻中它可被描述成智慧型網路單元。
4.Transformer
Transformer於2017年由Google發表，取代許多RNN模型，因為其採用注意力的機制(attention)，不需要像RNN一樣基於時間順序處理，也就是說注意力機制可以從任何位置上下左右提取資訊，並達成並行化處理，大幅減少了訓練時間與準確率。

2-2 文本挖掘

文本挖掘（Text mining）主要工作為文字分析，一般是將文字經過處理後產生較為重要的文字資訊，並透過自然語言處理和分析，將這些訊息通過分類和預測來產生最終評價和解釋。文字挖掘的方法有很多種，方法如下。

關鍵詞提取:

對較長的文本進行分析，找出對文本較為關鍵的關鍵詞。

文本摘要:

對大型文檔或多個文本做出簡要概述。

聚類:

聚類是一種從未標註的文本中取得隱藏的數據結構的一種技術。

文本分類:

利用監督是學習的方式，對未知數據的分類進行預測的一種方法。

觀點抽取:

對文本進行分析，找出核心的觀點，並判斷為正/負面的言論，主要是對特定主題進行分析，如美食、飯店、汽車等等的評論。

情感分析:

對文本做感情的判斷，主要分為正面、負面及中立三個分類，常用於話題監控、輿情分析等。

2-3 情感分析

1. 文本情感分析

文本情感分析的目的是為了找出作者在某個話題上所抱持的情感態度，也就是他當時做出言論的情緒狀態，而情感分析最基本的步驟，就是將文本中某段已知的文字分類成積極的、消極的或者是中立的，更高級的還可以以更複雜的情緒狀態做分類，例如：「高興」、「生氣」、「悲傷」等等。

2. 中文情感分析的困境

要處理文本前必須將句子進行分割，通常是透過標點符號作為分割符號。再來要進行斷詞的工作。斷詞是為了瞭解中文文章的意義，由於中文詞是開放的集合，在處理不同領域的文件時，常常造成分詞系統產生錯誤的切分，為了解決這個問題就必須補充該領域的詞彙，以增加辨識情感的準確率。

2-4 BERT模型

BERT(Bidirectional Encoder Representations from Transformers)，中文意思是變換器的雙向編碼器表示技術，**BERT**起源於預訓練的上下文表示學習，**BERT**是一種深度雙向的、非監督學習的語言表示，且僅使用純文字語料庫進行預訓練的模型。而上下文無關模型（如**word2vec**或**GloVe**）為詞彙表中的每個單詞生成一個詞向量表示，因此容易出現單詞的歧義問題。**BERT**其注意力機制可以考慮到單詞出現時的上下文。例如，詞「離開」的**word2vec**詞向量在「有事先離開了」和「爺爺離開人世了」是相同的，但**BERT**根據上下文的不同提供不同的詞向量，詞向量與句子表達的句意有關，解決歧義的問題，BERT深度且雙向的其使用上下左右文來表示字或詞，BERT在這方面的表現都高於其他模型。

BERT採用一個簡單的方式進行學習，遮蔽輸入中百分之十五的單詞，通過深度雙向Transformer編碼器執行整個序列，然後預測被遮蔽的單詞。
例如:
輸入: 有個人走進 [MASK1]，他要買一 [MASK2] 牛奶。
標記: [MASK1] = 商店, [MASK2] = 加侖

而為了學習句子與句子之間的關係，給定BERT兩個句子A與B，可以判斷B是否為A的下一個句子或是語料庫中隨機的一個句子。
例如:

句子A:有個人走進商店
句子B:他要買一加侖牛奶
標記:是下一個句子

句子A:有個人走進商店
句子B:企鵝不會飛
標記:不是下一個句子

然後，Google在大型語料庫維基百科和 BookCorpus上訓練了一個大型模型（12層到24層的Transformer），花費了十分長的時間（100萬升級步驟）。使用BERT有兩個步驟，預訓練與微調，預訓練的成本相當昂貴需要在4到16個Cloud TPU上執行4天才跑得完，但每種模型的訓練都是一次性的，而微調成本不高，只需要從預訓練好的模型開始，大約可以在30分鐘內完成，而且BERT目前總共可以支援超過109種語言，剩下的大多是文法結構較難或無法處理。

2-5 統計分析

統計分析(Statistical Analysis)，顧名思義就是將訊息統整併分析，主要透過收集數據、整理數據及分析數據並從中尋找出有用的資料。

統計方法利用字詞的前後文為依據，若此字詞有出現相似的前後文，則有相似的詞意，以「我晚餐吃牛排，我宵夜吃漢堡」為例，晚餐與宵夜有相同的前後文「我--吃」，因此晚餐與宵夜之間，比起晚餐與牛排之間有更相近的關係。

參、 研究方法

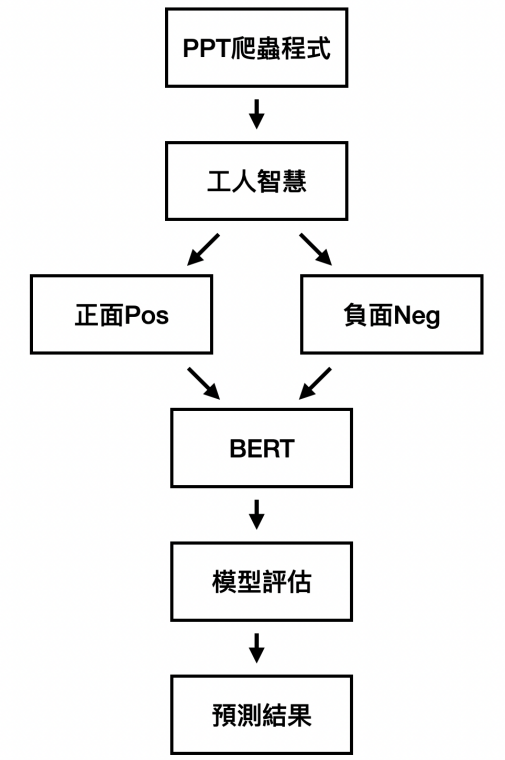

圖3-1 實驗架構

3-1 資料集與工具

本資料集的母體來源是針對PTT網站**nCoV19**版文章尋找**COVID-19**相關文章為目標，透過 **Python** 網路爬蟲程式抓取，並利用工人智慧將資料一筆一筆加以分類及標籤，探討詮釋文章意義，以挖掘隱含在這些文章中的重要資料與訊息。

一、 ptt-web-crawler (PTT 網路版爬蟲)

我們利用Github上開源的爬蟲程式ptt-web-crawler，抓取PTT上關於新冠肺炎的文章，參考圖3-2，並將抓取下來的的json檔轉換成txt檔，參考圖3-3，並將內容作提取，分類成正負面兩面的資料集，給程式做訓練。

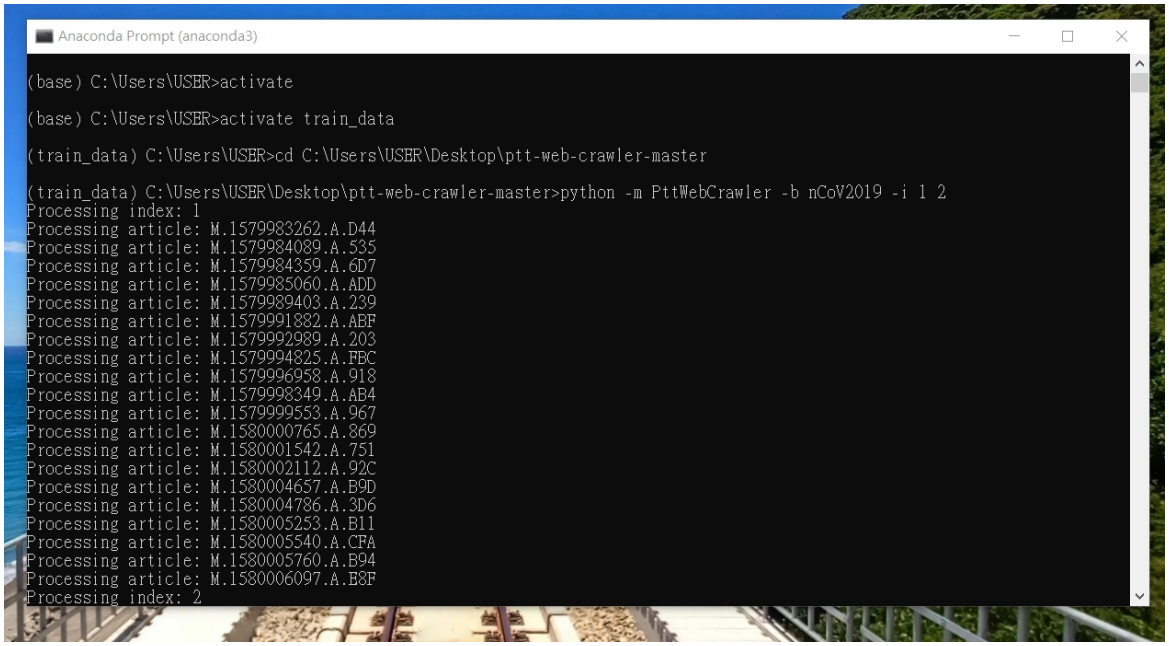

圖3-2. PTT爬蟲

爬蟲參考網址:https://github.com/jwlin/ptt-web-crawler

Json to txt

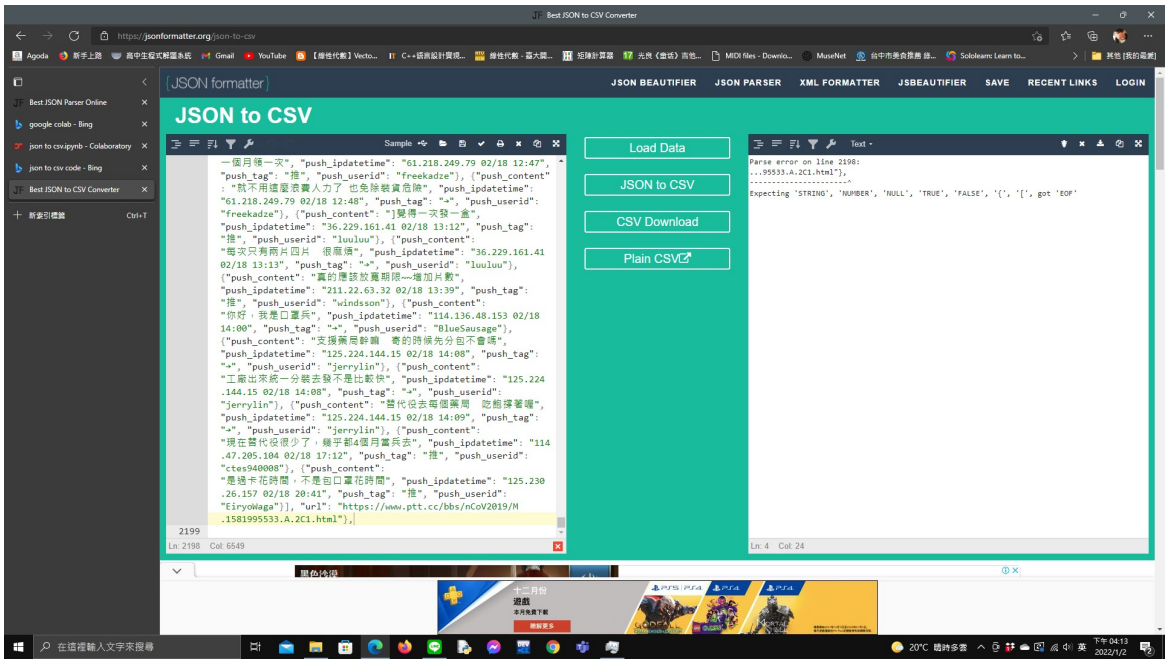

圖3-3. 轉檔

參考網址:https://jsonformatter.org/json-parser

二、 Huggingface Transformers

Huggingface團隊提供許多API與預訓練模型來執行不同形式的任務，這些模型主要可以應用在文本、圖像與音訊三大領域，Transformers還可以結合多種模式來執行任務，例如從掃描文檔中提取訊息、視覺問答與光學字符識別等等多種任務。本次實驗的採用的Huggingface提供的“bert-base-chinese”預訓練模型進行微調，該模型以繁簡體中文訓練而成，擁有大約2.1萬個tokens，共144組(12層 * 12heads)注意力參數。

3-2 系統設計

本實驗經由網路爬蟲(參考圖3-4)取得資料集，如上節3-1所示，取得資料後進行文本分類，將文章分類為正面(pos:1)、負面(neg:0)，本實驗使用bert-base-chinese預訓練模型，再以上述提到的資料集進行微調(fine-tuning)訓練下游任務，請參考圖 3-5.。

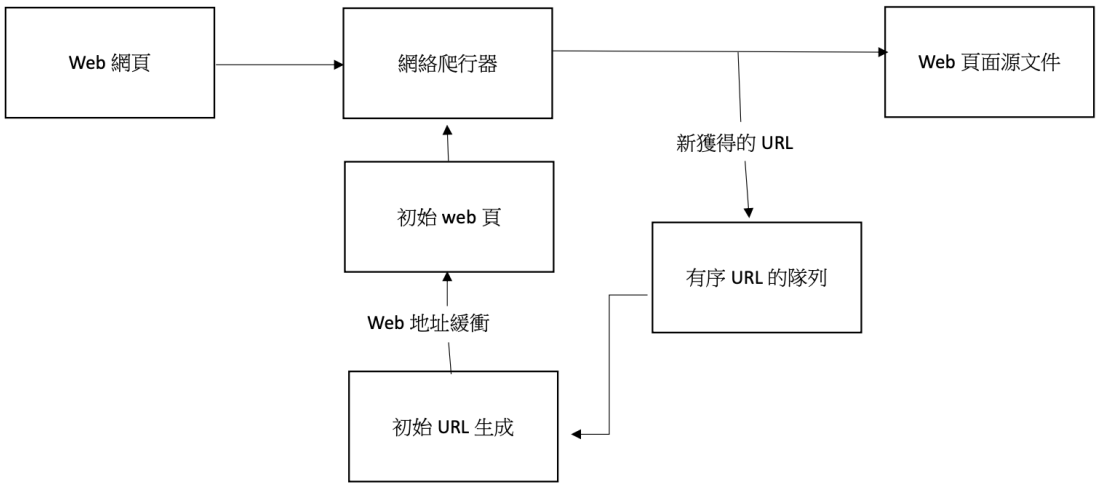

圖3–4. 網路爬蟲

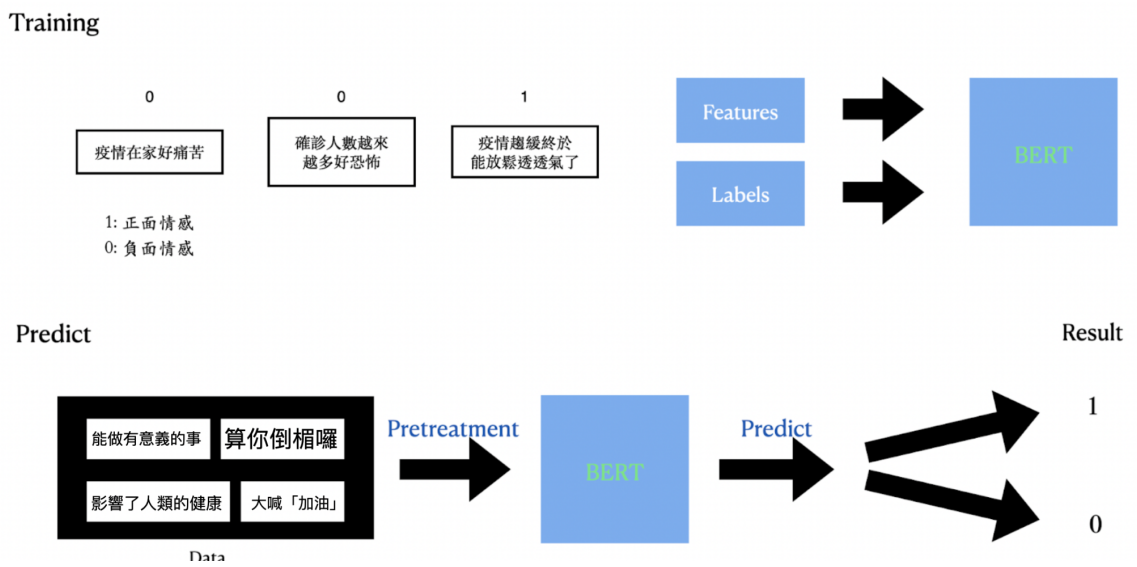

圖3-5. BERT訓練

3-3 模型評估

根據所擁有的數據集以及預期成果，選擇最適合的模型進行訓練，並評估此模型是否合適。

一、 損失函數

BERT模型使用交叉熵(Cross-Entropy)來當作損失函數，交叉熵為分類問題上主要使用的評估方式，交叉熵越小表示預測結果與實際情況越符合，公式如圖3-6，y為預測結果分佈，t為真實數據分佈。

$$E(t,y) = -\sum_i t_i log y_i$$

圖3-6. 交叉熵

二、 混淆矩陣

混淆矩陣是以最簡單的二分法進行分類而得出的一張表格

TP:實際為True，預測為Positive，預測與實際結果相同。

TN:實際為True，預測為Negative，預測與實際結果不同。

FP:實際為False，預測為Positive，預測與實際結果相同。

FN:實際為False，預測為Negative，預測與實際結果不同。

	模型預測為真	模型預測為否
真實情況為真	TP	FN
真實情況為非	FP	TN

表3-1. 混淆矩陣

三、 正確率(Accuracy)

正確率為模型評估最基本的指標，表示模型正確預測的分數，但如果分類不平衡的數據集，僅正確率不能說明所有的情況。公式參考圖3-7。

$$Accuracy = \frac{TP + TN}{all}$$

圖3-7. 正確率

四、 F1指標

在辨識和偵測相關的演算法中，為評估模型完整性，常會提到精確率(precision)和召回率(recall)，公式參考圖3-8及圖3-9，而F1指標為兩者的調和平均，F1指標能較全面地評斷模型的表現。公式參考圖3-10。

$$Precision = \frac{TP}{TP + FP}$$

圖3-8. 精確度

$$Recall = \frac{TP}{TP + FN}$$

圖3-9. 召回率

$$F_1 = \left(\frac{\mathrm{recall}^{-1} + \mathrm{precision}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}.$$

圖3-10. F1

肆、 實驗結果與討論

4-1 實驗環境

硬體及軟體	規格與版本
中央處理器(CPU)	Intel Core i5-4440 3.10GHz
記憶體(RAM)	16G DDR3
硬碟(Disk)	1TB HDD
圖形處理器(GPU)	GTX 1050 2GB
作業系統(OS)	Window 10
Torch	1.10.1+cu102
Tensonflow	2.6.0

表4-1 開發環境

4-2 實驗結果

由本次實驗解結果可以看出即是在訓練資料少的情況下，BERT依然擁有十分高的準確率，可見BERT在微調前BERT本身就對語言擁有十分高的理解，實在令人佩服。參考4-2。

	precision	recall	f1-score	support
0	1.00	0.86	0.92	7
1	0.91	1.00	0.95	10
accuracy			0.94	17
macro avg	0.95	0.93	0.94	17
weighted avg	0.95	0.94	0.94	17

表 4-2. 評估指標

- macro average 表示計算每個標籤的 f1 的函數，並返回平均值而不考慮數據集中每個標籤的比例。
- weighted average 表示計算每個標籤的 f1 的函數，並返回考慮數據集中每個標籤的比例的平均值。

伍、 結論

5-1 研究限制

一、 技術問題

Cuda錯誤，在訓練階段收到錯誤"CUDA error: invalid device ordinal CUDA kernel errors might be asynchronously reported at some other API call,so the stacktrace below might be incorrect. For debugging consider passing CUDA_LAUNCH_BLOCKING=1."，該錯誤表示使用CUDA執行程式碼時出現了問題，但是使用GPU訓練不會報錯，建議使用CPU執行，最後指定CPU又正常運行，推測是CUDA版本與系統沒有調適好導致錯誤，無法使用GPU的情況下訓練大量資料必定消耗更多時間成本。

二、 資料集

在沒有已成形資料集的情況下，製作資料集有以下困境:

1.假新聞

網路上因為各種因素如：網軍、專業寫手以及各種政治因素導致假新聞氾濫，假新聞以帶有目的性散佈不實謠言來混淆民眾，而寫假新聞的人通常是受到利益導致，而這樣的情況對我們資料集非常不利，不僅違反我們當初設定（文章為發文者當下情緒代表），更有可能因為數量龐大而導致實驗失去正確性。

2.準確率

本次實驗文章分類全靠人工，費神又費力，因此資料樣本數稀少，而且人工分類並非百分之百正確，人因為成長環境、教育程度等各種因素，對於同一項事物可能擁有不同的見解，尤其在台灣各類文章可能涉及到當前時事及意識形態，因此資料應經過更加嚴謹的篩選及評斷。

三、 BERT

BERT模型是本次實驗最大的優點同時也是缺點，BERT模型當初在訓練時用4-16個Cloud TPU訓練了4天前前後後共花了1.2萬美元，訓練BERT的成本是非常昂貴的，所幸他們開源提供了預訓練模型，但如果要自己改造或者是製作一個BERT，我想那不是一個人使用者負擔的起的。

5-2 未來改進與方向

因為假新聞的影響，應考慮先將資料透過假新聞過濾程式如cofact來去除掉不適合的資料，BERT方面還提供了許多的預訓練模型可以嘗試，像是縮減版的Distill-bert更輕量化但準確率也非常高，以及更進階的中文預訓練模型BERT-wwm使用全詞遮罩準確率又提升更多，還有其他基於不同語料庫開發的許多開源模型可以做嘗試，希望能在找到更好的結果。本次實驗僅完成了模型的測試，未來還可以進行網站建置及時的新冠肺炎輿情系統，提供使用者對於疫情與民眾情緒關係的視覺化分析。

參考文獻

蘇文群 (2021)。真的假的？！BERT 你怎麼說？

黃若蓁 (2020)。運用BERT模型對中文消費者評價之基於屬性的情緒分析。

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

混淆矩陣(confusion matrix)介紹:混淆矩陣(confusion matrix)介紹. 混淆矩陣：常見用來評估分類模型的指標 | by CHEN TSU PEI | NLP-trend-and-review | Medium

準確率、精準率、召回率、F1，我們真瞭解這些評價指標的意義嗎？: https://www.gushiciku.cn/pl/pylt/zh-tw

lstm: https://zh.wikipedia.org/wiki/%E9%95%B7%E7%9F%AD%E6%9C%9F%E8%A8%98%E6%86%B6

tf-idf: https://zh.wikipedia.org/wiki/Tf-idf