演算法與資料分析期末報告

基於 LSTM & BERT 機器學習
之網路輿情分析

國立金門大學
資訊工程系

作者：簡志融

指導教授：張珀銀 教授

目次

一、摘要

在這幾年間，機器學習、深度學習相關的應用獲得了許多關注，其中LSTM（Long Short-Term Memory）是一種循環神經網絡（RNN）的變體，旨在解決RNN在長序列上的梯度消失或梯度爆炸問題。LSTM是於1997年提出的，是一種有效、好用的模型，被廣泛應用於自然語言處理（NLP）和時間序列預測等領域。

BERT（Bidirectional Encoder Representations from Transformers）是一種自然語言處理技術，其基於Transformer網絡架構和預訓練方法，已經被廣泛應用於各種自然語言處理任務中。網路輿情是一個需要處理自然語言的任務，如:識別評論的主題、意見和情感等，而BERT可以幫助我們更加有效率的處理這些任務，因為它使用大量的無標籤數據進行預訓練，從而學習到有關於自然語言的通用表示。在網路輿情的分析中，BERT可以用來建立分類模型，該模型可以將評論分為正面、負面或中立等類別，這樣可以幫助用戶、店家與企業更好的理解其服務或產品在消費者中的聲譽及看法，並針對特定類別的評論進行快速回應，從而改進服務或產品的設計和銷售策略。

本計畫預期貢獻包含：(1)用Selenium動態抓取各種網路輿情；(2)運用LSTM進行訓練並生成語言模型(3)運用BERT進行訓練並生成語言模型；(4)運用語言模型對評論進行分類及分析，透過結果及各種指標對語言模型進行評估。

二、緒論

1.研究背景

近年來，隨著網路蓬勃發展的影響，帶來許多便利的改變，人們可以在網上購物、娛樂、學習、分享、交換意見等.....，大大節省了時間和精力；同時，網路也成為了人們瞭解世界重要且主要的渠道，人們可以通過網路了解全球各地的新聞、時事和文化，此外，網路還為人們提供了更多的溝通和交流機會，人們可以通過社交網站、即時通訊軟體等，與世界各地的人進行交流，建立龐大的社交網絡，這使得人們更容易交流不同文化和觀點，豐富了人們的生活經驗，但也衍生了許多問題，因為網路的易於接觸，一些不懷好意的人藉著極佳的方便性，到處散布假消息，刻意的操弄輿論，重傷他人，因此在這世道下，有效的分辨言論真假是非是相當重要的議題。

2.研究動機

自然語言評論分析是一個快速發展的領域，研究動機主要源於以下幾個方面：(1)電子商務的蓬勃發展，越來越多的產品和服務經由網際網路被推向市場，消費者可以通過網路上的平台對這些服務和產品進行評論；(2)社交媒體的普及，越來越多的人在社交媒體平台上發表自己的看法；(3)自然語言處理技術的進步，自然語言評論分析的準確度和效率得到了大幅提升，用來分析的價值也隨之提高，使得它成為一個被廣泛應用的工具。自然語言評論分析可以幫助用戶、店家與企業更好地理解大眾對產品和服務的看法，更好地分析現代消費者對產品和服務的真正需求及痛點，從而因應趨勢進行改進。

3.研究問題

因此，本計畫所要探討【研究問題】包含：

1. 如何找到最合適的爬蟲來獲取不同的網路輿情？
2. 如何把獲取到的資料進行預處理並製作成資料集？
3. 如何將從模型獲取的資料進行客觀分析？
4. 如何將模型訓練得更加精確？
5. 對於政府或店家能夠有效地獲得民眾的反饋？

4.研究目的

1. 多方參考資料文獻尋找方法。
2. 應用爬蟲套件，將不同的資料存入資料集。
3. 提出一種基於機器學習的自動化比對技術。
4. 透過真人反饋對比模型分析分數調整。
5. 論述說明本計畫研究成果對於訊息過濾貢獻及未來發展。

三、文獻探討與回顧

網路輿情評論是大部分民眾對特定實體的第一印象，幾乎所有特定實體都會有屬於自己的評論，透過手機在網路上就能夠看到其他人對特定實體的大致評價，同時自己也能夠流下自己的意見，而主要的評論大可分為以下三種:

一、

評論與表達內容大致相符合

二、

評論與表達內容相差甚遠

三、

評論與表達內容毫不相干

第一種表達與評論內容一致，這類型的評論是占最主要的，但有介於每個人標準不一，即使服務水準再高，或許仍有人感到不滿；相反地，即使是一些路邊攤，有人則滿足於此，給與高度評價，而且就算是同樣的地點，每個人當下的心情及感受都會不一樣，有可能給出低於或高於真正水準的評價，以致評論蒐集的分數事實上並不客觀。

第二及第三種表達和評論內容大相逕庭的，因為網路過於發達，常常有人當現代范仲淹，即使沒去過岳陽樓，還能隔空寫出岳陽樓記，這件事對古人來說或許很強，但對現代人來說，卻是絕佳的攻擊工具。常常有新聞報導，某某店家對於弱勢族群顧客惡言相向，或者是某某店家對於顧客的不耐煩及不適當行為都被媒體報導，以至於各種網路正義評論家就紛紛湧入店家頁面，一窩蜂湧入店家頁面不負責任地流下情緒性字眼，導致店家身心受創，害人家生意做不下去；亦或是一些以評論嘲諷店家的人，故意以一些似是而非的言論嘲諷店家，導致評論機制失衡。

由於各種複雜因素導致的不足，進而藉機開發撰寫一支程式，運用不同的Python爬蟲套件，因應不同資料需求，在網路上多方汲取評價，並整理資料，再由程式處理遺漏值，抽掉空白值，將數據處理至最佳後，拆分訓練集合測試集，運用BERT預訓練模型，後以測試集測試模型的完成度，並找出問題，反覆修正，直到訓練出理想中的模型，再來就可將模型的應用擴大，不只用來分析評論，也能衍生至分析任何文本，如：新聞中立與否、假訊息、文章邏輯矛盾等，廣泛應用於生活當中。

Kaggle

Kaggle是一個數據建模和數據分析競賽平台。企業和研究者可在其上發布數據，統計學者和數據挖掘專家可在其上進行競賽以產生最好的模型。眾多策略可以用於解決幾乎所有預測建模的問題，Kaggle的目標則是試圖通過眾包的形式來解決這一難題，進而使數據科學成為一場運動。而在本報告當中，模型最初的訓練資料便是來自Kaggle的一自然語言處理競賽。

LSTM

長短期記憶模型（Long short-term memory）為一特殊的RNN模型（遞歸神經網絡），目的是要解決「長序列」訓練過程中的梯度消失和梯度爆炸問題，相比RNN，LSTM能夠在更長的序列中有較好的表現，因此本報告中使用LSTM做為第一種模型。

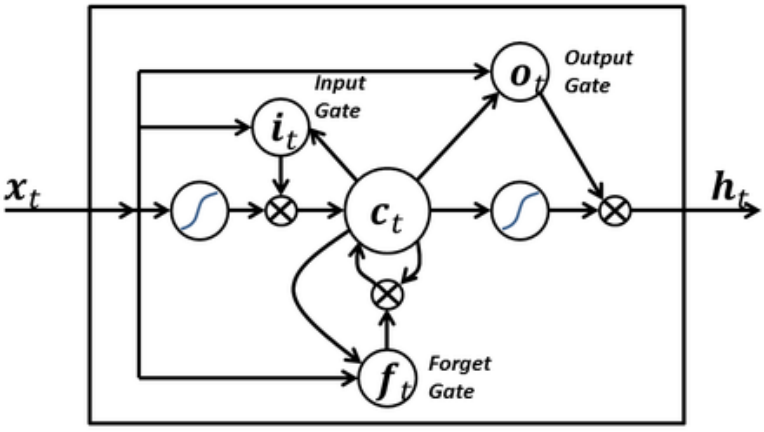

圖3.1、LSTM基礎架構圖

BERT

BERT（Bidirectional Encoder Representations from Transformers）由Google於2018年提出，是一種基於Transformer架構的語言模型。它是一種預訓練模型，在使用大量無標籤文本上進行預訓練，學習到了豐富的語言表示。BERT的特點在於它能夠雙向理解上下文信息，並且在各種自然語言處理任務中取得了顯著的成果，因此本報告中使用LSTM做為第一種模型。

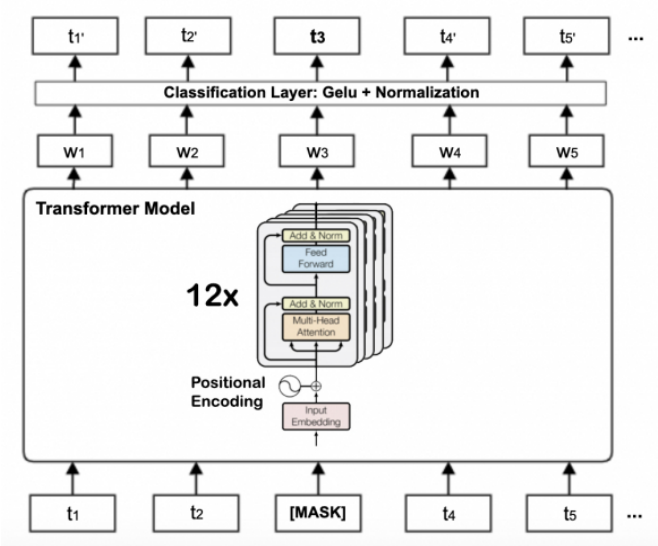

圖3.2、BERT架構圖

四、研究方法與步驟

1. 理想系統分析情境圖

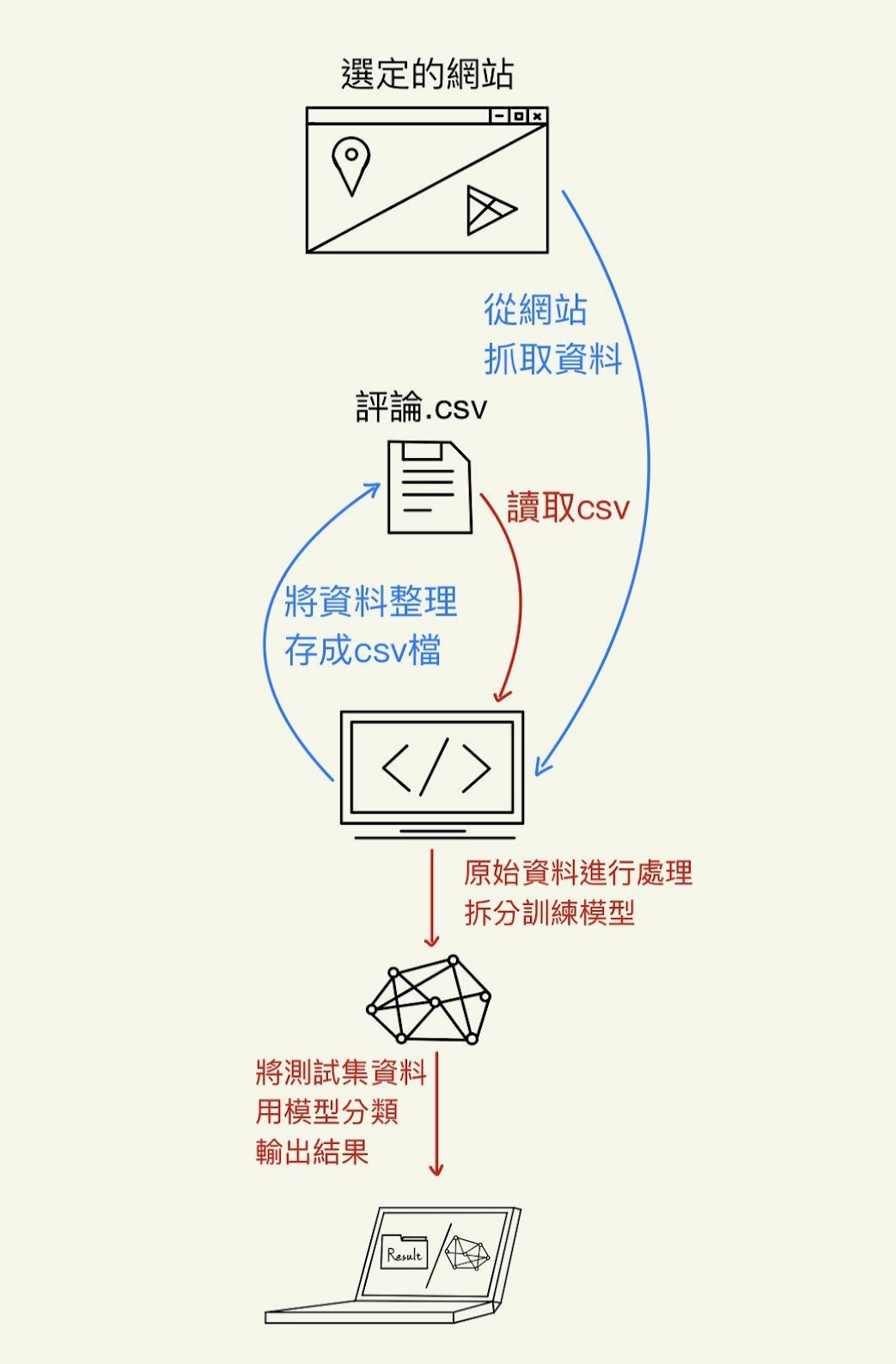

圖4.1、系統情境圖

2. 目標：

- 抓取網路輿情
- 用抓取的資料訓練模型
- 準確客觀分析出評論的好壞
- 判斷是否為優良店家
- 讓語言模型進行評分並評估

3. 步驟:

a. 使用網路上的資料集對不同的模型進行訓練
b. 依結果進行評估、調整
c. 使用爬蟲的資料對模型進行測試

4. 網路爬蟲收集資料:

使用Python的selenium模組撰寫自動化爬蟲程式抓取網站評論。在程式中，只需簡單設定關鍵字、範圍及一些爬蟲相關參數即可自動化的爬取資料。

```
keyword = '充電器'
page = 1
```

圖4.2、關鍵字及範圍設定

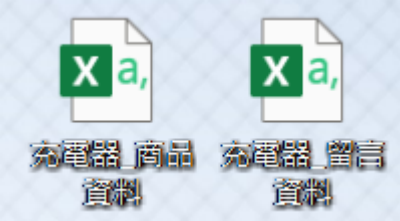

圖4.3、程式輸出的檔案

商品ID	賣家ID	商品名稱	商品連結	價格	品牌	存貨數	商品文案	上架時間	折數	可否搭配	可否大量	選項
3118427050	8776550	【Gooday	https://sho	254		186	✨推出	1.57E+09	6	FALSE	FALSE	[{'name':
21869801069	8908342	宏晉 直	https://sho	499		8045	宏晉	1.67E+09	6.2	FALSE	FALSE	[{'name':
2124483919	5910055	【現貨含	https://sho	229		152	💯在地	1.56E+09	4.5	FALSE	FALSE	[{'name':

歷史銷售	可否分期	是否官方	是否可預	喜愛數	商家地點	SKU	評價數量	五星	四星	三星	二星	一星
6816	TRUE	FALSE	FALSE	874	嘉義縣水	['雙槽充	3266	3157	98	6	3	2
1927	TRUE	TRUE	FALSE	1373	臺中市霧	['太平洋	743	708	18	8	5	4
17582	TRUE	FALSE	FALSE	1967	新北市三	['四槽充	5910	5646	203	37	5	19

圖4.4-4.5、商品資料檔案內容

商品ID	賣家ID	商品名稱	價格	使用者ID	是否匿名	留言時間	是否隱藏	訂單編號	給星	留言內容
2.38E+10	8291388	【綠聯】1(	1999	13761978	TRUE	1.67E+09	FALSE	1.235E+14	5	品質:良好^nCP值:ok^n^n蠻需
2.38E+10	8291388	【綠聯】1(	1999	2463024	TRUE	1.671E+09	FALSE	1.238E+14	5	品質:質感很好，手感類似Ap
2.38E+10	8291388	【綠聯】1(	1999	22277867	TRUE	1.671E+09	FALSE	1.246E+14	5	品質:待確認^nCP值:便宜^n^n
2.38E+10	8291388	【綠聯】1(	1999	2.87E+08	TRUE	1.677E+09	FALSE	1.304E+14	5	品質:好^n^n加購的線剛剛好
2.38E+10	8291388	【綠聯】1(	1999	1916292	FALSE	1.681E+09	FALSE	1.334E+14	5	超讚的出貨速度^n超讚的商
2.38E+10	8291388	【綠聯】1(	1999	5626158	TRUE	1.671E+09	FALSE	1.241E+14	5	品質:👍^nCP值:👍^n^n比想
2.38E+10	8291388	【綠聯】1(	1999	18266799	TRUE	1.678E+09	FALSE	1.295E+14	2	品質:差^nCP值:理想與現實差

圖4.6、留言資料檔案內容

5. LSTM:

使用Kaggle假新聞資料集進行訓練，進行以下步驟:

- 文本分詞（Text Segmentation）、建立字典
- 序列的 Zero Padding
- 將 Label 做 One-hot Encoding
- 拆分資料集、訓練(Data permitting)
- Prediction

id	tid1	tid2	title1_zh	title2_zh
0	0	1	2017养老保险又新增两项，农村老人人人可申领，你领到了吗	警方辟谣"鸟巢大会每人领5万"仍有老人坚持进京
3	2	3	"你不来深圳，早晚你儿子也要来"，不出10年深圳人均GDP将超香港	深圳GDP首超香港？深圳统计局辟谣：只是差距在缩小
1	2	4	"你不来深圳，早晚你儿子也要来"，不出10年深圳人均GDP将超香港	GDP首超香港？深圳澄清：还差一点点……

圖4.7、資料集內容

```
['2017', '养老保险', '又', '新增', '两项', '农村', '老人', '人人', '可', '申领', '你', '领到', '了', '吗']
```

圖4.8、使用Jieba將文字有意義的切割

```
array([   0,    0,    0,    0,    0,    0,  217, 1268,   32, 1178, 5967,
         25,  489, 2877,  116, 5559,    4, 1850,    2,   13], dtype=int32)
```

圖4.9、Zero Padding

```
array([[1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [1., 0., 0.],
       [0., 1., 0.]], dtype=float32)
```

圖4.10、One-hot Encoding

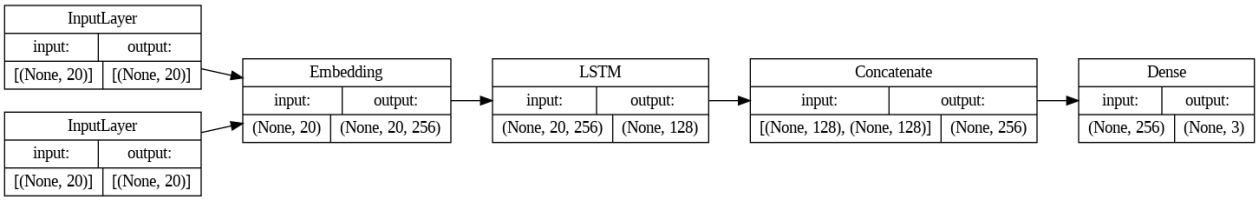

圖4.11、孿生神經網路模型

6. BERT:

使用Kaggle假新聞資料集進行訓練，進行以下步驟:

- 將原文本轉換成BERT相容的輸入格式
- 在BERT之上加入新layer成下游任務模型
- 訓練該下游任務模型
- 對新樣本做推論

[原始文本]
句子 1：苏有朋要结婚了，但网友觉得他还是和林心如比较合适
句子 2：好闺蜜结婚给不婚族的姜岚扔花球，倒霉的姜岚掉水里笑哭苏有朋！
分類 ：unrelated

[Dataset 回傳的 tensors]
tokens_tensor : tensor([101, 5722, 3300, 3301, 6206, 5310, 2042, 749, 8024, 852, 5381, 1351,
6230, 2533, 800, 6820, 3221, 1469, 3360, 2552, 1963, 3683, 6772, 1394,
6844, 102, 1962, 7318, 6057, 5310, 2042, 5314, 679, 2042, 3184, 4638,
4912, 2269, 2803, 5709, 4413, 8024, 948, 7450, 4638, 4912, 2269, 2957,
3717, 7027, 5010, 1526, 5722, 3300, 3301, 8013, 102])

segments_tensor : tensor([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1])

label_tensor : 2

[還原 tokens_tensors]
[CLS]苏有朋要结婚了，但网友觉得他还是和林心如比较合适[SEP]好闺蜜结婚给不婚族的姜岚扔花球
，倒霉的姜岚掉水里笑哭苏有朋！[SEP]

圖4.11、換成相容的輸入格式

```
[epoch 1] loss: 32.120, acc: 0.803
[epoch 2] loss: 19.275, acc: 0.845
[epoch 3] loss: 14.135, acc: 0.903
[epoch 4] loss: 10.738, acc: 0.868
[epoch 5] loss: 8.326, acc: 0.905
[epoch 6] loss: 8.947, acc: 0.930
CPU times: user 1min 41s, sys: 46 s, total: 2min 27s
Wall time: 2min 27s
```

圖4.12、訓練下游任務模型

	text_a	text_b	label	predicted
603	海口飞机撒药治白蛾	3月谣言盘点：飞机撒药治白蛾、驾考新规，你中“谣”了吗？	disagreed	disagreed
803	烟王褚时健去世	辟谣：一代烟王褚时健安好！	disagreed	disagreed
952	李宇春跟老外结婚	李宇春被传嫁给78岁老外？春爸被逼亲自辟谣：假的！	disagreed	disagreed
1752	海口飞机撒药治白蛾	紧急辟谣 飞机又来撒药治白蛾了？别再传了，是假的！	disagreed	disagreed
2646	12306数据泄漏	铁路12306 辟谣，称网站未发生用户信息泄漏！	disagreed	disagreed

圖4.13、對新樣本做推論

五、模型成果

經多次調整及實驗後，將結果上傳Kaggle後，LSTM模型得到0.7118分，而BERT模型得到0.8544分，由上述訓練之後歸納出以下幾點:

- 多種類分類問題對LSTM較不適用
- BERT在自然語言處理較卓越

在訓練LSTM時，透過Data permitting發現有過適問題，驗證集遺失值上升，查閱資料後，進行 Batch size 和 Epoch 的調整，發現在Epoch不變，調整BS只能起到些微作用；BS不變，調整Epoch對遺失值幫助雖不大，但準確率卻有所提升。

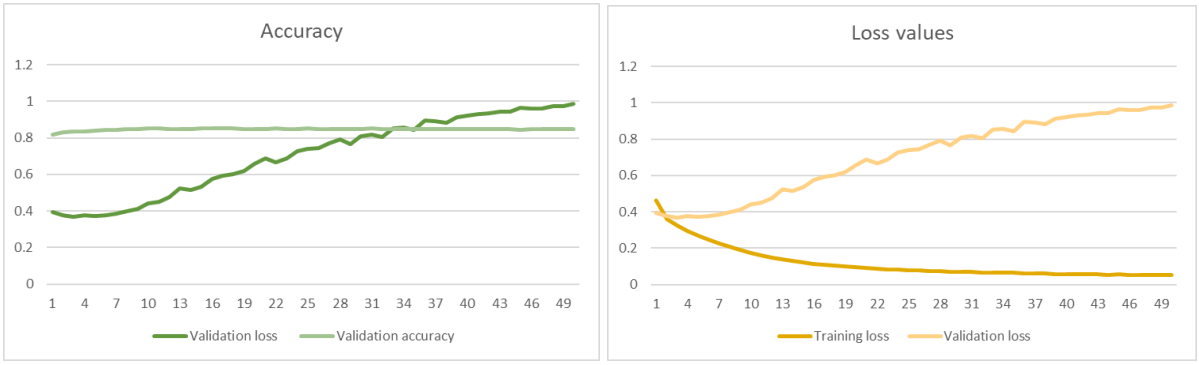

圖5.1-5.2、LSTM模型初始資料(BS:500；Epoch:50)

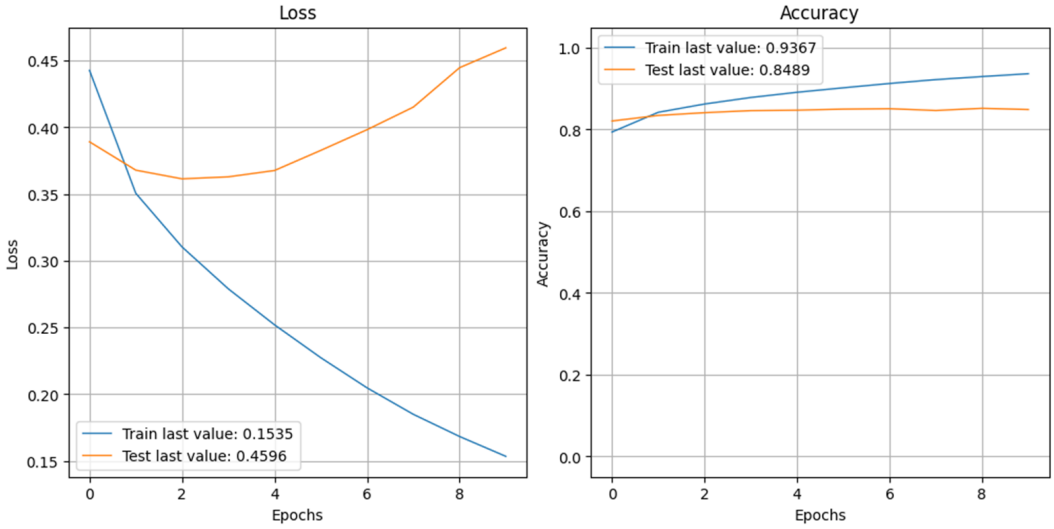

圖5.3、初次調整的結果，Loss一度下降(BS:250；Epoch:10)

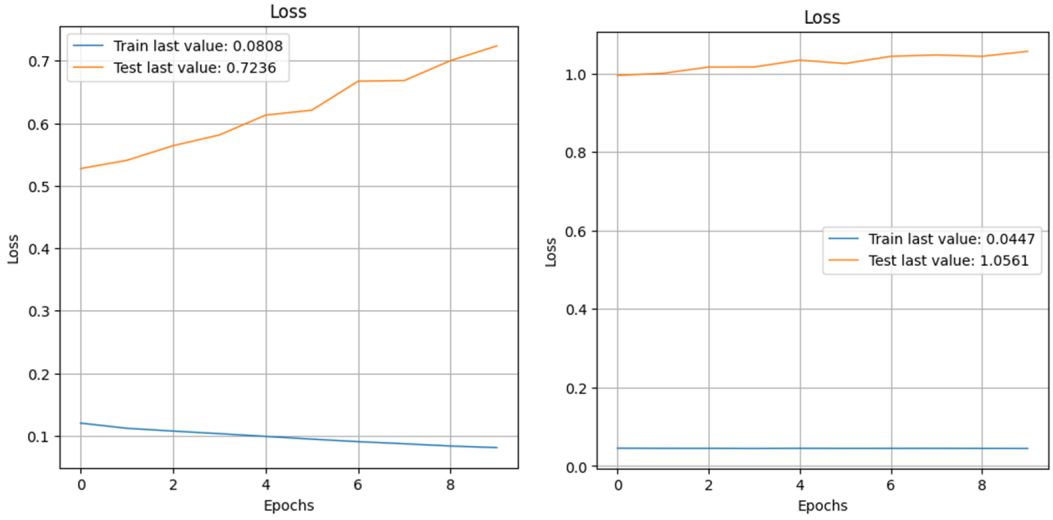

圖5.4-5.5、BS值上升可讓遺失值平緩但不收斂(左:500；右3000)

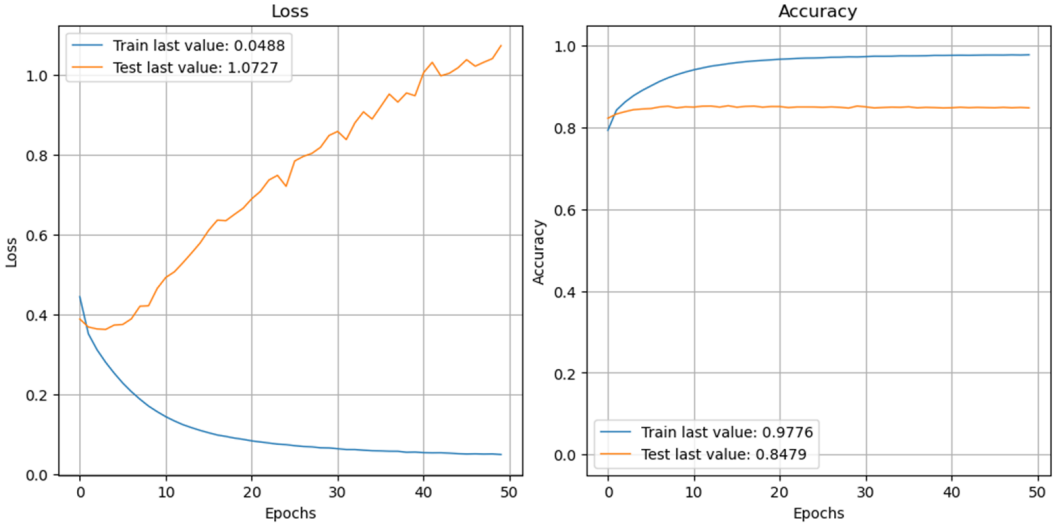

圖5.6、Epoch值上升可小幅提升準確率(BS:250；Epoch:100)

有鑑於BERT對自然語言的處理能力，將商品留言分類資料集交給BERT訓練調整並預測，爾後得出以下結果:

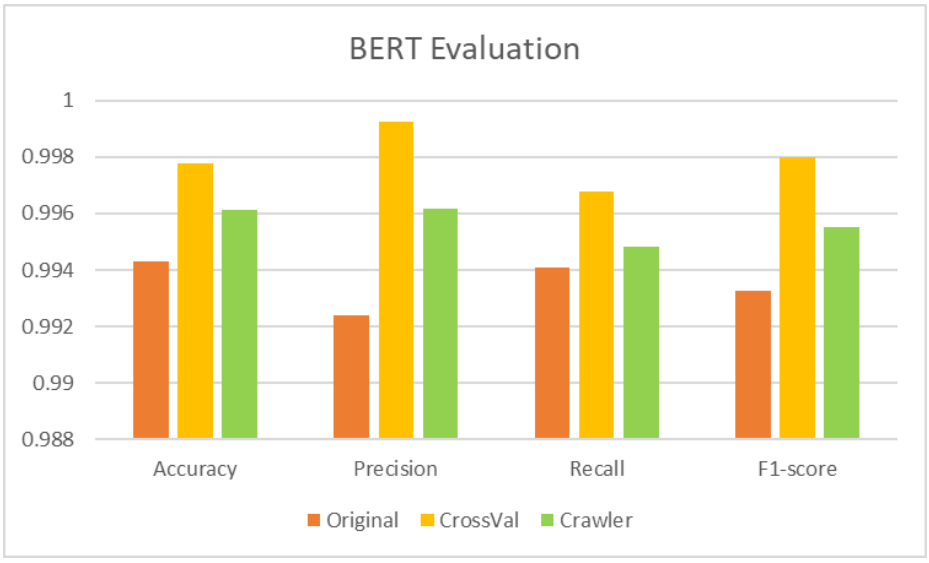

圖5.7、BERT評估表

相較三種分類問題，二分類問題對BERT來說相當容易，不僅在原資料集的測試集得到高分，更在其他沒見過的資料集取得0.995的F1-score，令人相當驚豔。

引用Google科學家的話:「有關BERT在上述自然語言理解任務中為何可以達到先進水平，目前還未找到明確的原因。目前BERT的可解釋性研究主要集中在研究精心選擇的輸入序列對BERT的輸出的影響關係，通過探測分類器分析內部向量表示，以及注意力權重表示的關係。」

六、結論

AI科技發展快速，其無非是世界一道不可阻攔的洪流，人們應保持開放樂見的心態面對，「縱浪大化中，不喜亦不懼。應盡便須盡　無復獨多慮。」在一波波的浪潮中，尋等機會，一舉站上AI的浪頭上，盡享AI帶來的便利及紅利。

資料集	模型	評估指標	其他
假新聞、評論	LSTM、BERT	Accuracy、Loss、Precision、Recall、F1-score	爬蟲

表6.1、統整表

七、參考文獻

BERT wiki

BERT？如何BERT？BERT的基礎介紹

BERT 自然語意演算法如何提升關鍵字理解能力

動態爬蟲：動態加載問題及在不同視窗間跳轉、滑動

Get text from span tag in BeautifulSoup

Finding web elements

進擊的 BERT：NLP 界的巨人之力與遷移學習

以神經網絡進行時間序列預測 — LSTM

Evaluation Metrics：分類模型