建立深度學習模型以預測各國新聞媒體的立場-以反送中為例

簡毓漩
巨資四 B
05170204

黃閔慈
巨資四 B
05170209

Abstract – 本篇文獻在研究各新聞媒體在反送中事件所扮演的角色，由於不同新聞的立場會影響到閱聽者對於每個事件的觀感和想法，故我們想藉由比較香港、臺灣、英國代表性媒體針對此議題的新聞立場，了解不同國家的媒體對於反送中議題的重要詞演變，並透過標記立場，建立深度學習模型藉此預測各媒體對於反送中議題的立場。

Keywords - Web Crawler, Word Segmentation, Text Mining, Sentiment Analysis, Deep Learning

I. INTRODUCTION

由於一個殺人案件衍生出來的反送中事件，持續了將近一年，還未見平復，仍然吵得沸沸揚揚，故勺丁业

Figure 1. 香港 01 新聞網

Figure 2. 台灣 UDN 新聞網

Figure 3. 英國 BBC 中文 新聞網

II. METHODS

我們針對三國媒體對於反送中議題的立場分析，會進行以下步驟的處理，首先是資料搜集，利用 Python 爬蟲套件至三家媒體的官網爬取新聞，並且人工標記每篇新聞對於反送中議題的立場，再進行斷詞後，先利用 TF-IDF 計算詞頻，觀察各媒體各月重要詞變化。

除了透過 TF-IDF 觀察各媒體變化外，我們提出以 PyTorch 套件所訓練的模型來預測各媒體立場，並和 Naïve Beyes 分類器做比較，並且視覺化呈現三間媒體個月的立場變化。

A. 資料搜集

1) Python Crawler

我們選擇了跟香港有歷史關係的英國，以其國內新聞媒體 BBC 為目標；有地緣關係的臺灣，以其國內新聞媒體 UDN 為例；香港自己本身，以香港 01 為對象，藉由上述三個國家跟香港有一定的關聯的前提下，使我們的資料更為完整。在每個新聞媒體中，我們先以「反送中」、「逃犯條例」來搜尋，並在搜尋後，抓取前 100 篇文章的 Hashtag ，來找還有哪些相關的關鍵字，再以各媒體關鍵字的前十名搜尋，抓取更多的新聞。此次新聞抓取的區間為 3/31~11/31 的所有反送中相關新聞，其中包含新聞標題、發佈時間、新聞內容、hashtag、URL，最後我們總共抓取到了 BBC 共 376 筆，UDN 共 3432 筆，香港 01 共 1570 筆。而為了之後訓練模型的需要，在抓取完資料後，我們開始以人工方式進行立場的標記，選擇每天新聞數量的一半，來判斷其立場為支持、反對亦或是中立。

TABLE I. UDN 之 HASHTAG 前十

標籤名	次數
香港	4303
反送中	3659
禁蒙面法	759
逃犯條例	344
立法會	276
北京	185
香港特首	165
警察	122
一國兩制	119
港警	104

TABLE II. 香港 01 之 HASHTAG 前十

標籤名	次數
逃犯條例	1046
反修例示威	291
逃犯條例相關聆訊	253
禁蒙面法	239
警務處	234
元朗黑夜	228
612 逃犯條例示威	153
警方記者會	145
警民關係	135
一國兩制	113

TABLE III. BBC 之 HASHTAG 前十

標籤名	次數
香港	193
逃犯條例	163
中國	159
政治	97
抗議	45
法律	41
示範	34
中美關係	29
美國	28
台灣	22

2) Word Segmentation

我們選擇利用Jieba套件進行斷詞，因為其方便使用且可以簡單地透過自定義字典來新增原本無法精確斷開的字詞。因為怕斷詞會不夠精確，故加入了531個詞彙的自定義字典後，再做斷詞的動作，表四為自定義辭典範例。

TABLE IV. 自定義字典範例

自定義字典範例		
逃犯條例	中國	政治
香港	抗議	示範
示威	中美關係	台灣
警務	美國	習近平
唐納德・特朗普	新聞自由	學生生活
社交媒體	犯罪	金融財經
航空旅行	20 國集團	旅遊
審查	朝鮮	一國兩制
環球時報	反修例示威	港獨
今日香港	港警	元朗黑夜
831 遊行	警方記者會	暴動
警民關係	中英聯合聲明	香港人權與民主法案

B. TF-IDF

初期我們希望利用 word2vec 來將新聞內容轉成向量，並利用 K-means 分群來觀察各媒體討論議題和用字區別，但我們發現 K-means 結果不盡理想，因此決定改變作法。故這個階段，我們利用 TF-IDF 套件做計算，以每個月為單位，來看各家媒體的新聞報導內容再每個月有甚麼不同的變化。

C. Model

我們提出以 PyTorch 套件寫成的名為 NEWS_NET2 的深度學習模型，藉此預測每一篇新聞的立場，並以隨機抽樣的方式和 Naïve Bayesian Classifier 比較模型效果，以下為我們的模型流程結構。

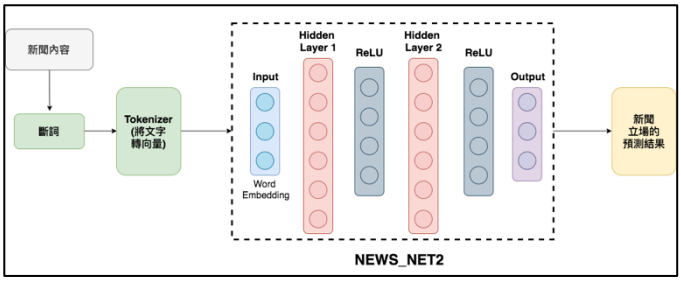

Figure 4.模型結構圖

1) Tokenizer

將斷好詞的新聞內容經由 keras 的 Tokenizer 套件轉成向量，運作形式會是將新聞內容的所有詞彙會總成一張詞彙表，詞彙表包含每一個詞彙和每個詞彙的索引值，而每一則新聞內容會依照該新聞內容所出現的詞彙轉換成出現詞彙的索引值，藉此將文字內容轉成向量。然而這樣轉向量的話，很有可能向量長度會不同，因為新聞內容出現的

詞彙數不等，為確保每一篇文章的向量長度相同，會在向量中剩餘的位置補 0。在將每篇新聞內容轉為向量後，才可放入 NEWS_NET2 模型中訓練。

2) NEWS_NET2 Model

NEWS_NET2 模型是以 PyTorch 套件所寫成，可以由上圖看出此模型包含輸入層、兩層隱藏層、輸出層。之所以會建立兩層隱藏層，也是經過實驗後，發現兩層以上隱藏層所得到的模型效果沒有更好，因此建立含有兩層隱藏層的 NEWS_NET2 模型。

此模型在輸入層時，會經過 Word Embedding 的轉換，將文字向量進行大小上的轉換，在進入第一層 Hidden Layer 進行線性計算，然而由於預測值為類別資料，因此每經過一層 Hidden Layer 會再經過 Relu 的激勵函數，藉此解決處理本次的非線性問題，而之所以選擇 Relu 是因為此模型仍屬於少量隱藏層且 Relu 有計算速度快和收斂速度快的優點，最後輸出 Tensor 形式的向量，再轉換後就能得到模型預測的新聞立場。

而此次因為資料集含有三家媒體，考量到每家媒體的用字遣詞不同，因此會在 NEWS_NET2 的模型架構下，設定不同參數，以使模型在不同訓練資料下得到更好的效果。

III. Results

A. DataSets

我們收集了分別來自香港 01、UDN、BBC 中文新聞網從 2019 年 3 月 31 到 2019 年 11 月 30 日的香港逃犯條例相關報導，每一則報導都含有標題、新聞內容、HashTag，而每一則報導都由組員分別獨立標記新聞立場，立場標記有反對、中立、支持，以下圖五顯示三家媒體在各立場的個數分佈，可以看到 BBC 中文的新聞被標記為中立的新聞最多，而反對立場的新聞最少，而 UDN 的新聞中被標記為支持立場的新聞最多，最少的新聞立場也是反對，而香港 01 的新聞被標記為中立的新聞為最多，反對和支持立場則相距不遠。

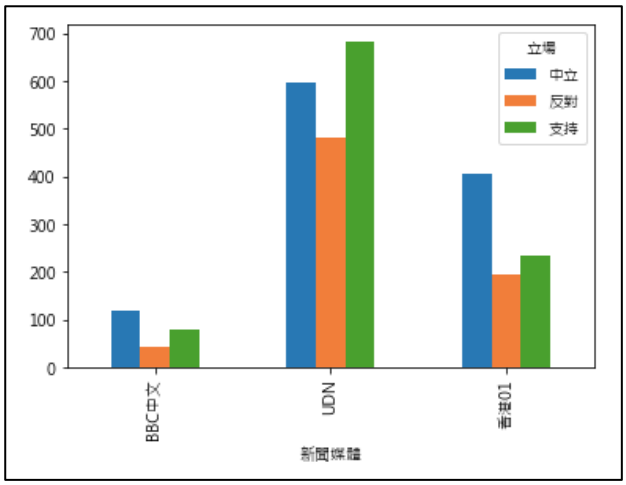

Figure 5. 香港 01、UDN、BBC 中文 新聞立場分佈圖

表五為資料集數據，我們一共從香港 01 爬回 1570 篇反送中相關新聞，而 UDN 則爬回 3432 篇反送中相關新聞，BBC 中文網爬回 376 篇相關新聞，其中有標記立場的資料分別是 833 筆、1759 筆、237 筆，而各媒體立場分佈如圖五所示。

表六為三間媒體的統計數據，可以看到 BBC 中文的新聞內容平均句數最多，可能在訓練模型時，會影響到模型效果。

TABLE V. DATASET

No. of Document	香港 01	UDN	BBC 中文
Train	666	1408	189
Validation	167	351	48
Test	737	1673	139
Total	**1570**	**3432**	**376**

TABLE VI. DATASET STATISTIC

Average Value	香港 01	UDN	BBC 中文
Average Sentence	46.68	32.27	90.65
Average word / per sentence	6.87	7.02	7.58

B. TF-IDF

根據前一個章節所說的方法，我們的結果如下列三表所呈現，而我們可以分別從其中去探討一些發現。

先以台灣的 UDN 新聞網為例，在反送中事件爆發的前期(四-六月)，由於台灣也是事件導火線的當事國之一(陳同佳於台灣犯罪後回港)，故大部分的報導偏向說明事件發生的原因以及香港政府對於此事的做法；而隨著事件慢慢地愈演愈烈，我們可以發現，報導內容開始頻繁出現中共方面的消息以及美國政府的做法，不難想像這是因為歷史及地理位置因素所造成的結果。而從前期到後期的新聞內容變化，結合現在台灣的本土意識，可以推定對於反送中這件事，整體而言，台灣的立場是由中立導向支持的態度。

再以香港本土的香港 01 新聞網為例，可以發現在一開始的內容，就跟中共有頻繁的連結，而到了後期開始出現對於中共一國兩制政策的內容，並強烈地表達香港人要爭取自由與人權的想法。故配合香港警察在這段期間對於香港記者的所作所為，我們可以說縱使新聞應該保持中立，旨在揭露各方面的事實，但在親身的經歷後，必然會有一定程度的影響新聞內容，推測香港的立場從一開始大多數就是支持的態度。

最後以英國的 BBC 新聞網來做結尾，從下表 VI 可以看到，英國在事件一開始關注的是經濟發展的部分，直到

近兩個月來才開始轉為對於香港人自治以及人權方面問題的層面。這並不難以理解，在中美貿易戰的大環境下，且香港又是國際上金融中心之一，英國最先關注的不是歷史背景下的產物，而是現在正擺在眼前的經濟問題；到了後期，由於美國通過的香港人權與民主法案，一定程度上造成中美之間的關係有些微變化，故英國必須在這方面多做關注。正因如此，我們可以認為，英國對於反送中事件的立場，在基於整體經濟上的考量，會是中立偏反對的態度，因為事件愈演愈烈，就會越難收拾，而對於香港的經濟就會有重大的打擊，間接影響到其他各國的金融事業。

從上述的內容，可以發現各家新聞媒體在一開始都能保持著新聞中立的態度，但隨著時間的推移、事件的發展、客觀環境的變化，皆不同程度的影響到了新聞的內容以及立場的變化，這也可以說明，反送中事件，從一開始台港兩國政策之間，到後來台港中三方的關係，再延燒到國際之間的金融轉變，對於國際上的影響程度，越來越大，也越來越不容忽視。

TABLE VII. UDN 之 TF-IDF 每月前五之結果

TF-IDF					
月份	字詞	Score	月份	字詞	Score
四月	進會	0.259	八月	未經	0.340
	台方	0.247		批准	0.267
	會和策	0.194		集結	0.260
	女友	0.177		浩天	0.225
	陳同佳	0.172		香港眾志	0.194
五月	動議	0.341	九月	港澳辦	0.492
	林太	0.216		主任	0.332
	信任	0.196		張曉明	0.305
	票反	0.167		中聯辦	0.197
	該動議	0.167		中共	0.190
六月	韓特	0.342	十月	武警	0.430
	英國	0.258		灣體育	0.374
	更顯	0.256		美聯社	0.315
	不移	0.244		深圳	0.253
	抗爭	0.219		演習	0.249
七月	美國	0.389	十一月	催淚	0.338
	黑手	0.249		吸入	0.269
	作品	0.222		健康	0,251
	美方	0.195		焚燒	0,216
	指控	0.187		評估	0.169

TABLE VIII. 香港 01 之 TF-IDF 每月前十之結果

TF-IDF					
月份	字詞	Score	月份	字詞	Score
四月	台灣	0.287	八月	香港	0.357
	陳同佳	0.225		中英聯合聲明	0.280
	Shoulder	0.225		G7	0.201
	去年	0.167		峰會	0.201
	修例	0.112		耿爽	0.187
五月	台灣	0.556	九月	香港	0.437
	香港	0.310		一國兩制	0.369
	中共	0.275		一國	0.179
	失敗	0.206		社會	0.152
	何俊仁	0.206		中央	0.141
六月	節目	0.314	十月	基本法	0.282
	夜問	0.259		人權	0.276
	打權	0.259		香港	0.262
	環時	0.228		自由	0.248
	台獨	0.217		一國兩制	0.223
七月	楊光	0.342	十一月	韓國瑜	0.368
	香港	0.277		台灣	0.308
	堅決	0.258		遺忘	0.204
	暴力犯罪	0.205		中華民國	0.172
	支持	0.183		四大	0.167

TABLE IX. BBC 之 TF-IDF 每月前五之結果

TF-IDF					
月份	字詞	Score	月份	字詞	Score
四月	中國	0.221	八月	新加坡	0.612
	海上	0.201		香港	0.260
	肖揚	0.201		本土意識	0.160
	舉行	0.201		管制	0.137
	一帶一路	0.201		對象	0.128
五月	談判	0.505	九月	港交所	0.500
	貿易	0.278		交易	0.366
	特朗普	0.239		交易所	0.296
	股市	0.208		倫交所	0.218
	團隊	0.208		倫敦	0.171
六月	許穎婷	0.397	十月	革命	0.504
	香港	0.338		顏色	0.350
	內地	0.210		變革	0.194
	宏碁	0.188		香港	0.169
	學生	0.179		中國	0.164

TF-IDF					
月份	字詞	Score	月份	字詞	Score
七月	元朗	0.415	十一月	美國	0.356
	居民	0.382		香港	0.246
	新界	0.238		選舉	0.239
	香港	0.208		香港人權與民主法案	0.234
	港英政府	0.137		法案	0.171

C. Model

此次實驗會將 NEWS_NET2 模型和樸素貝葉斯模型進行比較，由於此次資料集是來自三間不同的媒體，考量到不同媒體用字遣詞的不同，因此會在 NEWS_NET2 的結構下訓練三個模型，同時和三個樸素貝葉斯模型進行比較，以下羅列模型概述。

- NEWS_NET2 : 是我們本次研究提出的模型。
- Naïve Bayesian Classifier : 樸素貝葉斯模型利用機率論預測樣本類別的機率算法。經由計算輸出機率最高的類別最為預測結果。是本次研究的比較基線模型。

1) 模型效果比較 – Validation Accuracy

表十顯示 NEWS_NET2 最好效果下的模型參數，可以看到三個訓練集下 Embedding Size 皆為 150，實驗時，即使將 Embedding Size 再往上調，模型效果和 Embedding Size 是 150 得到一樣的準確率，顯示 150 為 Embedding Size 最好模型的上限。在此設計 NEWS_NET2 模型時，設置兩層隱藏層的 Hidden Size 為一樣的，而測試三個媒體的訓練集時，發現 Hidden Size 大概都在 150 內做調整會得到較好的效果，若再往上調，模型效果並沒有更好。Learning Rate 的部分，可以看到有 0.1 和 0.01，實驗的時候同時還有 5e-4 在做測試，實驗結果表明在香港 01 和 BBC 中文的訓練資料下，Learning Rate 為 0.1 會得到較好的效果。可以看到有 Dropout Rate，這裡的 Dropout Rate 是為了避免模型 Overfitting 所設置的參數，Dropout Rate 越高代表隱藏層會有越多神經元被丟棄，避免模型過度學習，在 NEWS_NET2 模型中，每經過一層隱藏層，都會有部分神經元被丟棄，其中可以看到 BBC 中文的 Dropout Rate 高達 0.8，代表每經過一層隱藏層只有 20%的神經元進入下一層。為讓準確率達到最高和收斂準確率，Epochs 統一設置為小於等於 150。

TABLE X. NEWS_NET2 最好模型參數

參數	NEWS_NET2		
	香港 01	UDN	BBC 中文
Embedding Size	150	150	150
Output Size	30	20	20
Hidden Size	130	100	150
Learning Rate	0.01	0.1	0.01
Dropout Rate	0.1	0.2	0.8
Epochs	150	150	100

表十一為 NEWS_NET2 模型和 Naive Bayesian Classifier 在三間媒體的相同 Validation Dataset 的 Accuracy 比較，可以看到在香港 01 和 UDN 的 Validation Dataset 下，NEWS_NET2 的模型效果都比 Naive Bayesian Classifier 的效果差，而唯獨在 BBC 中文的 Validation Dataset 的時候，NEWS_NET2 的模型效果比較好，接下來會透過在 Test 隨機抽樣的方式來驗證模型的實際效果。

TABLE XI. NEWS_NET2 VS. NAÏVE BAYESIAN CLASSIFIER

新聞媒體	Model	Validation Accuracy
香港 01	NEWS_NET2	0.5030
	Naive Bayesian Classifier	0.5749
UDN	NEWS_NET2	0.3807
	Naive Bayesian Classifier	0.5596
BBC 中文	NEWS_NET2	0.4375
	Naive Bayesian Classifier	0.3541

2) 模型效果比較 - 隨機抽樣

為了比較我們所提出的模型與 Naïve Bayesian Classifier 兩者之間的效果對於此議題的預測何者較優，我們從每家媒體的 Test 資料中隨機抽取了 40 筆資料，以人工的方式標記立場，再去看我們所標記的立場與模型所預測的立場是否相同，以此來推定兩個模型對我們所探討的議題，何者較為適用。

從表十二我們可以看到，在 UDN 以及 BBC 中文的結果顯示，我們所提出的模型預測出來的結果跟人工所標記的結果較相符，而在在預測結果上其實兩者相當接近，故可以說我們所提出的模型並不會比 Naïve Bayesian Classifier 差。

TABLE XII. 隨機抽樣模型預測結果比較

新聞媒體	預測結果比較		
	預測結果是否正確	NEWS_NET2	Naïve Bayesian
香港 01	正確	22	21
	不正確	18	19
UDN	正確	24	28
	不正確	16	11
BBC 中文	正確	26	24
	不正確	14	16

表十三為在三間媒體各隨機抽樣 40 筆中正確答案和 NEWS_NET2 模型和 Naïve Bayesian Classifier 預測結果的實際數據。

可以看到在香港 01 的資料集中，NEWS_NET2 預測的類別最多為中立，而實際答案為反對則一筆都沒有預測到，回去看正確答案為中立的資料，觀察為什麼 NEWS_NET2 有 5 筆沒有預測為中立，會發現這 5 筆中有 1 筆關於「世界人道主義日」被標記為「支持」，而有 1 筆關於「撐警大會」、「警方逮補」被標記為「反對」，推測由於人工標記會判斷反對和支持視角的比例，來決定該新聞的立場，但是機器卻不一定能夠了解這部分，所以有可能是看到「關鍵字」判斷立場。

可以看到在 UDN 這個資料集中，在正確答案為反對時，NEWS_NET2 能夠預測出反對的個數遠勝過 Naïve Bayesian Classifier，而在正確答案為中立時和支持時，NEWS_NET2 的預測正確個數沒有比 Naïve Bayesian Classifier，在此三個數據集下，唯獨香港 01 的數據集下，NEWS_NET2 的預測立場有反對，回去看新聞內容，會發現這 8 篇文章內容主要有關「中國」、「中國國務院」、「港府」等機關的新聞。

在 BBC 中文的資料集中，可以看到 NEWS_NET2 在實際答案為支持的情況下，預測結果的支持的個數比 Naïve Bayesian Classifier 高，查看預測數據會發現 NEWS_NET2 在「佔中三子」的相關文章中預測立場為「支持」。

TABLE XIII. 隨機抽樣模型預測結果

新聞媒體	Model	正確答案		
		反對	中立	支持
香港 01	NEWS_NET2	0	16	6
	Naïve Bayesian	3	12	6
	實際正確個數	**6**	**21**	**13**
UDN	NEWS_NET2	7	9	8
	Naïve Bayesian	0	17	11
	實際正確個數	**8**	**20**	**12**
BBC 中文	NEWS_NET2	0	16	10
	Naïve Bayesian	0	18	6
	實際正確個數	**3**	**23**	**14**

表十四是從三間媒體中各挑一筆新聞出來比較模型的預測結果，可以看到在 HK01「一國兩制」、「止暴制亂」的反送中新聞中，兩模型都猜錯答案，而在 UDN「抗爭導致經濟下滑」的新聞中，Naïve Bayesian Classifier 預測為反對，而 NEWS_NET2 則預測為中立，在 BBC 中文「日本天皇退位」的新聞中，Naïve Bayesian Classifier 預測為支持，而 NEWS_NET2 則預測為中立。

TABLE XIV. 模型預測結果實例

新聞媒體	文章摘要	NEWS_NET2	Naïve Bayesian	正確答案
香港 01	約三十名愛港之聲成員和市民下午前往灣仔警總，高舉「保家衛港大丈夫、止暴制亂真英雄」的橫額，並不時高叫「阿 sir 加油」、「大丈夫」、「香港全賴有你」等口號。高達斌指，支持政府依法施政，並堅決支持「一國兩制」、「港人治港」。	支持	中立	反對
UDN	一件情殺案，導致犯罪引渡條例的修改，將七年以上罪犯的引渡地擴及於台灣、中國大陸。引起香港反送中的大規模抗爭，這兩天更演變成癱瘓捷運、機場的靜坐示威，香港觀光旅遊受挫、經濟成長恐下滑，更可能動搖其亞洲金融中心的地位。	中立	反對	中立
BBC 中文	日本新天皇德仁 5 月 1 日即位。日本告別此前持續約 30 年的「平成」時代，改年號為「令和」。雖然中日關係近年有起有伏，	中立	支持	中立

新聞媒體	文章摘要	NEWS_NET2	Naïve Bayesian	正確答案
	但不論是在中國內地、香港還是台灣，都對這次皇權交替報以極大關注。新華社報道稱，習近平致電德仁天皇，表示「中日兩國一衣帶水，友好交往歷史源遠流長。雙方應該攜手努力，共促和平發展，共創兩國關係美好未來」蔡英文在推特賬號上表示，台灣與日本在平成年代成為「區域間最好的朋友」，希望在令和年代也能「持續當彼此最好的伙伴」。			

IV. Conclusions

在本次研究，我們先以 TF-IDF 重要詞來觀察香港 01、UDN 和 BBC 中文三間媒體在不同月份對反送中相關議題的用字遣詞，來推測不同媒體對於反送中議題的立場。我們也藉由標記每篇新聞的立場發現香港 01 和 BBC 中文對於反送中議題的立場以中立居多，而 UDN 則是支持的立場最多，顯示不同媒體對於該消息立場的不同。

接著我們提出了 NEWS_NET2 模型來預測新聞立場，希望能夠藉由學習不同媒體針對特定議題的用字遣詞來判斷該媒體發布此文章的立場，藉由這個模型，我們能夠預測該媒體所發佈的文章是什麼立場。

V. Workload

- 黃閔慈—做 PPT、TF-IDF & Sentiment Analysis 程式碼以及論文撰寫、BBC 網頁爬蟲
- 簡毓漩—論文撰寫、UDN & 香港 01 網頁爬蟲、Sentiment Analysis 程式碼測試

VI. Referances

[1] 深度學習激勵函數介紹 http://cvfiasd.pixnet.net/blog/post/275774124-%E6%B7%B1%E5%BA%A6%E5%AD%B8%E7%BF%92%E6%BF%80%E5%8B%B5%E5%87%BD%E6%95%B8%E4%BB%8B%E7%B4%B9

[2] 進入 NLP 世界的最佳橋樑：寫給所有人的自然語言處理與深度學習入門指南 https://leemeng.tw/shortest-path-to-the-nlp-world-a-gentle-guide-of-natural-language-processing-and-deep-learning-for-everyone.html

[3] 立場、意圖與價值• 語言分析與資料科學 https://lab-of-ontologies-language-proce.gitbook.io/ladsbook/

[4] 逃犯條例爭議：香港五大媒體如何選邊站 https://www.bbc.com/zhongwen/trad/chinese-news-48908246

[5] 『長短期記憶網路』(LSTM) 應用 -- 情緒分析(Sentiment Analysis) https://ithelp.ithome.com.tw/articles/10193924