元大證券計量交易部實習總結報告

見習單位：元大證券 計量交易部

見習主管：余光麒 資深副總經理

見習督導：蘇高毅 專業副總經理

曾盟雅 專業經理、林家豪學長

實 習 生：陳冠維 國立清華大學計量財務金融學系

專案主題:

運用NLP技術分析財經新聞

研究目的與成果

動機

希望能夠藉由NLP的技術，去給予每一則新聞一個情感分數。甚至針對個股去做一個情緒指標，以幫助交易策略開發。

研究成果

1. 《BERT分類模型》訓練完成17個主題
2. 《提升模型準確度》利用文本聚類、抽樣
3. 《建立語料庫》利用BERT問答擷取財報資料、關鍵狀態詞彙、建構情感分析所需之語料庫
4. 《新聞情緒分數》利用情感分析結合BERT問答、分類模型，給予每則新聞情感分數
5. 《個股情緒指標》針對個股，設計time decay的情緒分數
6. 《文本數據提取》提取新聞中的文本數字，轉換成數值資料並建立數據庫

一、BERT分類任務

BERT 分類任務

動機

為了從文本中提取非結構化的信息，我們希望利用語言模型(LM)來對財經新聞標題進行文本情緒的分類：正面、中立、負面。

研究方法 – BERT

- 選擇利用模型：BERT（上下文相關、提升至句子級別）
- 優點：
  BERT透過**預訓練**(大量數據)能學習語言的一些基本結構特徵，在下游任務進行精調時，有更好的泛化效果及訓練時間；並且與之前的預訓練模型相比，它捕捉到的是真正意義上的雙向文本信息。
- 缺點：
  處理商業、金融等動態環境的文本需要持續**手動標記資料**，並進行再訓練與驗證，較為耗時且成本昂貴；且手動標記資料也容易有不一致的偏差存在。

BERT 分類任務 – 多主題模型

《BERT分類模型》

- 訓練資料：依據不同主題，提取其對應的資料（**新聞標題**）進行訓練
  - 優點：其雜訊較單模型少，較易提升準確率
- 資料標籤類別(三類)：**正面、中立、負面**
- 已訓練完成模型：準確率皆可達90%以上，財報面模型準確率可逾95%

籌碼面

外資、熱錢、投信、法人、大戶、籌碼、降息、降準

需求面

需求、訂單、大單、急單、轉單、接單、單量、追單、缺貨、供應

生產面

生產、產能、產量、產線、產值、製造、製程、良率、投產

銷售面

銷售、出貨、銷貨、銷量、買氣、通路、市場、後市、後勢、展店、市占、市佔、

營運面

營運、業績、獲利、營利、盈利、收益、

財報面

營收、財報、財測、EPS、毛利、毛利率、淨利、純益、盈餘、殖利率

BERT分類任務 – 文本相似度計算聚類 《提升模型準確度》

- 利用BERT預訓練模型(distiluse-base-multilingual-cased) 計算標題嵌入(Embeddings)
- 將標題嵌入進行聚類 (Kmeans)，將標題分成若干組
- 在每個標題組別中，進行隨機抽樣若干個

1月營收年增6%，Q1營運估落底回升	1月營收年減24%，宅經濟為Q1營運再添柴火
2月營收月減14%，復工拚Q1營運贏上季	11月營收月減9% 看好音樂串流平台發展助攻營運
1月營收年減10% Q1營運不確定性較大	1月業績小減，大陸湖北客戶佔比不到1%，今年營運穩健走揚
1月營收年減11% 第1季營運不確定性較大	1月營收年增6%！疫情衝擊「但沒有客戶銷訂單」 Q1營運估落底逐季回升

- 以**進行聚類，取樣後的**資料去訓練表現得比未分群的結果好

  其泛化程度較高，且能大量減少手動標籤的時間、數量；

  在相當的資料量下，較容易訓練到各種句型的標題

BERT 分類任務 – 資料前處理架構

資料聚類、分群

- 計算標題向量（Embeddings）
- 利用K-means聚類分群
- **分群後暫不抽樣**

計算含台股名稱的新聞標題占比

- 若占比低於標準則只取台股資料，以減少訓練資料雜訊
- **測試集則以台股資料為主**

依照新聞發布時間

- 依照時間將資料分為訓練集(前)、測試集(後)
- **若不以時間區分，容易高估準確率**

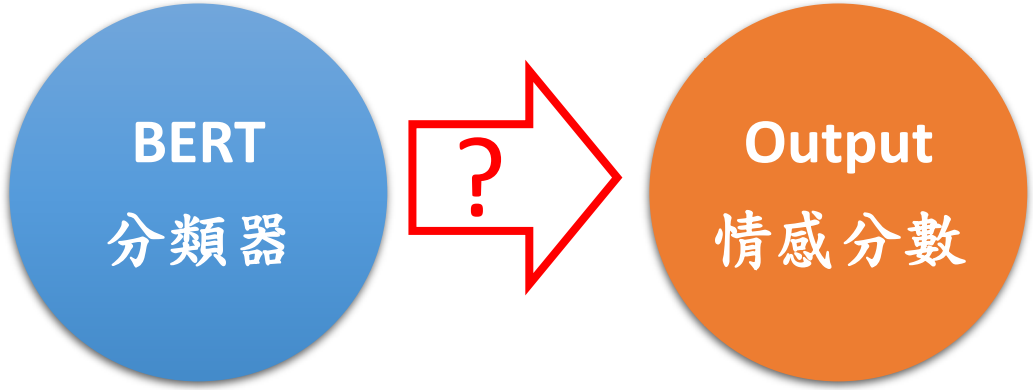
二、情感分析

BERT分類器的缺點

1. 無法體現正面、負面程度上的差異

2. 新聞出現一正一反的情形時，無法判別

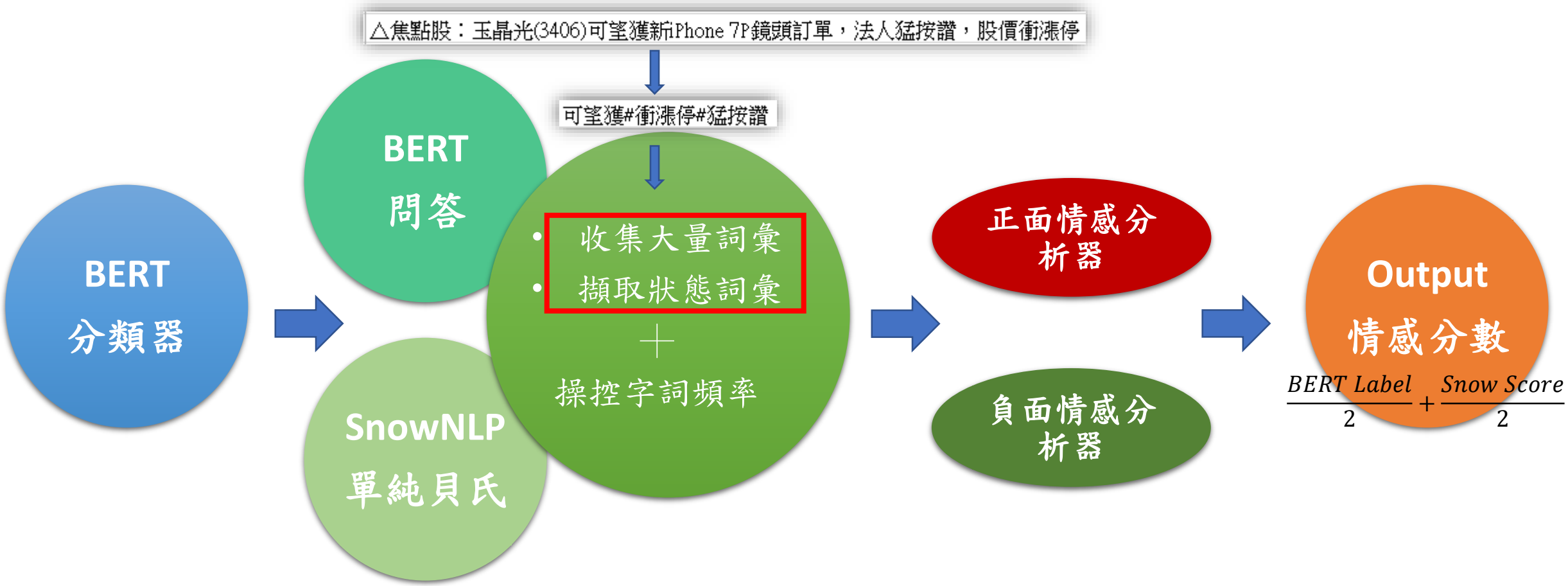
標題	類別
瑞祺電通(6416)**春燕來**，訂單能見度直達明年Q2，明年營運**大爆發**	**正面**
久陽訂單**增**拚**重返成長**	**正面**
外資鎖定加碼台積電、面板雙虎 反手賣超鴻海逾萬張	**中立**
台積電7月業績減，8月起5奈米效益顯現，營收重現成長動能	**中立**



- 可以透過情感分析去給正面、負面新聞一個分數
- 一正一反的中立新聞，暫時分數給0

情感分析建立流程圖

《新聞情緒分數》



SnowNLP 單純貝氏 《新聞情緒分數》

- Snow 情感分析訓練模型：**Naïve Bayes (單純貝氏)**

- Snow 優點：
  1. 可以自己丟入詞彙訓練，根據不同主題各自訓練一個情感分析器。
     此外訓練、預測花費時間非常短

  2. **統計理論輔佐，可以某種程度的去控制想要給特定詞彙的分數**

     Ex:

     假如我們想要給「滿載」這個詞彙高一點分數，可以去增加「非常正面」詞彙表裡「滿載」出現的次數，這樣子模型就會給他更高的分數

貝氏定理

《新聞情緒分數》

P（非常正面｜爆漲）= P（爆漲｜非常正面）x P（非常正面）/ P（爆漲）

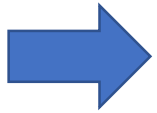
P（普通正面｜爆漲）= P（爆漲｜普通正面）x P（普通正面）/ P（爆漲）

P（爆漲｜非常正面）= 暴漲在非常正面詞彙表出現的次數 / 非常正面詞彙表的總數量

P（爆漲｜普通正面）= 暴漲在普通正面詞彙表出現的次數 / 普通正面詞彙表的總數量

非常正面詞彙表

暴漲, 暴漲, 暴漲, 暴漲, 暴漲, 滿載, 增加, 突破, 創新高, 月增, 年增, 成長……



P（非常正面｜爆漲）

SnowNLP – 情感分析器建構流程 《新聞情緒分數》

動機

認為主題新聞用語應有固定模式

抓取詞彙

抓取所有新聞的狀態詞彙
（BERT QA）

人工分類

- 非常正面
- 普通正面
- 非常負面
- 普通負面

調整字詞頻率

調控詞彙出現的頻率，來控制分數

訓練情感分析器、微調

- 用「非常正面、普通正面」等詞彙訓練出正面情感分析器
- 用「非常負面、普通負面」等詞彙訓練出負面情感分析器

SnowNLP – 情感分析器算分

《新聞情緒分數》

分數初步計算方法

- BERT分類器判斷為**正面**： $0.5 + \frac{Pos\ Snow\ Score}{2}$

- BERT分類器判斷為**負面**： $-0.5 - \frac{Neg\ Snow\ Score}{2}$

標題	類別	情感分數	總分
瑞祺電通(6416)**春燕來**，訂單能見度直達明年Q2，明年營運**大爆發**	正面	0.99	0.99
和大(1536)明年特斯拉訂單**增逾60%**，估全年營收75億元起跳**創新高**	正面	0.73	0.86
久陽訂單**增 拚重返成長**	正面	0.09	0.54
Google高階伺服器板爆材料瑕疵 傳訂單已轉向日本松電工 台燿**痛失大單**	負面	0.98	-0.99
訂單**能見度不佳**，外資**下修**大江目標價	負面	0.48	-0.74
〈大亞展望〉疫情干擾本業Q2訂單**能見度低** 營運**保守看**	負面	0.008	-0.50

三、BERT問答

BERT問答建立流程圖

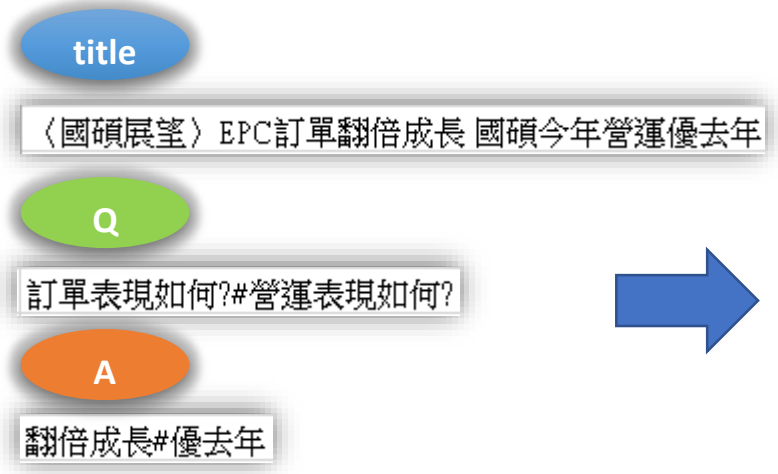
《建立語料庫》

動機

1. 收集大量狀態詞彙
2. 擷取標題狀態詞彙提升分數準確度

人工標籤

準備每則新聞所要問的問題以及答案



資料預處理

清理標題格式，提升模型準確度

訓練多主題QA模型

主題：營運、訂單、營收、需求、業績、EPS、出貨、股價、獲利、外資

人工標籤

《建立語料庫》

文章	問題	答案
▲個股：上半年獲利看俏＋現金減資45.44795%及配息1.5元，國揚盤中股價漲停	獲利表現如何?#股價表現如何?	看俏#漲停
△個股：LuLu上修第四季獲利預估，股價創新天價，儒鴻(1476)受惠，今年展望佳	獲利表現如何?#股價表現如何?	上修#創新天價
△個股：台驊上半年獲利可期，股價帶量突破大聯大收購價，盤中站上30元關卡	獲利表現如何?#股價表現如何?	可期#帶量突破
△個股：晶技(3042)Q1獲利佳，法人估Q2營收季增10-15%，早盤股價強攻漲停	獲利表現如何?#營收表現如何?#股價表現如何?	佳#季增#強攻漲停
△個股：板卡Q2淡季不淡，微星、技嘉、華擎4月業績攻高，股價勁揚	業績表現如何?#股價表現如何?	攻高#勁揚

- 每個主題須至少標籤約100筆的資料去訓練
- 此次模型含10個主題，共約800多筆資料

資料預處理

《建立語料庫》

- 將句子間的空格改為句號、在句子尾端加上句號，**讓模型能夠更好的抓取出狀態詞彙**。

清理前

《業績-其他電子》可成6月、Q2**營收**齊登同期高 Q3**動能**續看旺

清理後

《業績-其他電子》可成6月、Q2**營收**齊登同期高。Q3**動能**續看旺。

BERT問答訓練流程

《建立語料庫》

文章　　問題　　答案

答案位置 ➡ start: 9、end: 10

文章	問題	答案
▲個股：上半年獲利看俏+現金減資45.44795%及配息1.5元，國揚盤中股價漲停	獲利表現如何?#股價表現如何?	看俏#漲停
△個股：LuLu上修第四季獲利預估，股價創新天價，儒鴻(1476)受惠，今年展望佳	獲利表現如何?#股價表現如何?	上修#創新天價
△個股：台驊上半年獲利可期，股價帶量突破大聯大收購價，盤中站上30元關卡	獲利表現如何?#股價表現如何?	可期#帶量突破
△個股：晶技(3042)Q1獲利佳，法人估Q2營收季增10-15%，早盤股價強攻漲停	獲利表現如何?#營收表現如何?#股價表現如何?	佳#季增#強攻漲停
△個股：板卡Q2淡季不淡，微星、技嘉、華擎4月業績攻高，股價勁揚	業績表現如何?#股價表現如何?	攻高#勁揚

讀文章

將標題轉換為BERT看得懂的形式（Embeddings）

➡

接收問題

將問題轉換為BERT看得懂的形式（Embeddings）

➡

回答

BERT學習出答案在文章中的起始、結束位置

BERT問答模型訓練結果

《建立語料庫》

- 模型訓練 - 運算處理器：**Colab** TPU
- Server上預測速度：約 **30個問題 / 秒**
- 主題：
  - ■ **營運**、**訂單**、營收、需求、業績、EPS、出貨、股價、獲利、外資
  - ■ 測試集準確率： **0.90**

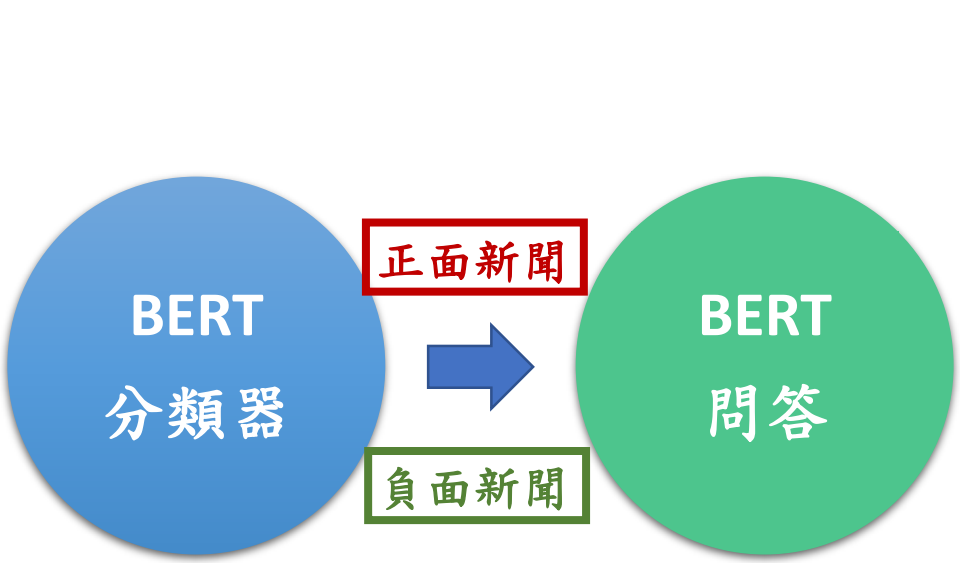
《備註》這裡的準確率代表：抓對了多少比例的字詞

Ex: 正確答案：遠優於預期；預測答案: 優於預期

準確率： 4/5 = 0.80

BERT、情感分析流程圖

《建立語料庫》

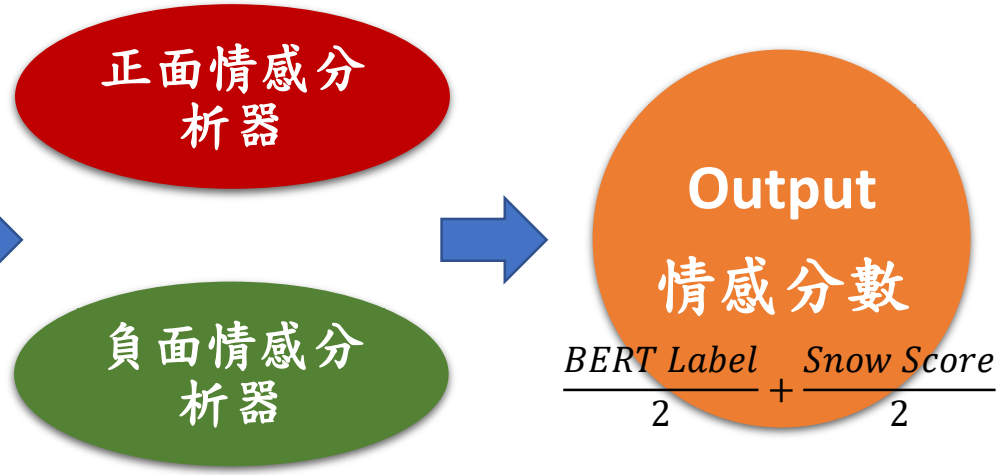


正面狀態詞彙

滿
增#逐月升
旺#續拚新高
正向看待
回升#優於去年同期
下單#續攻

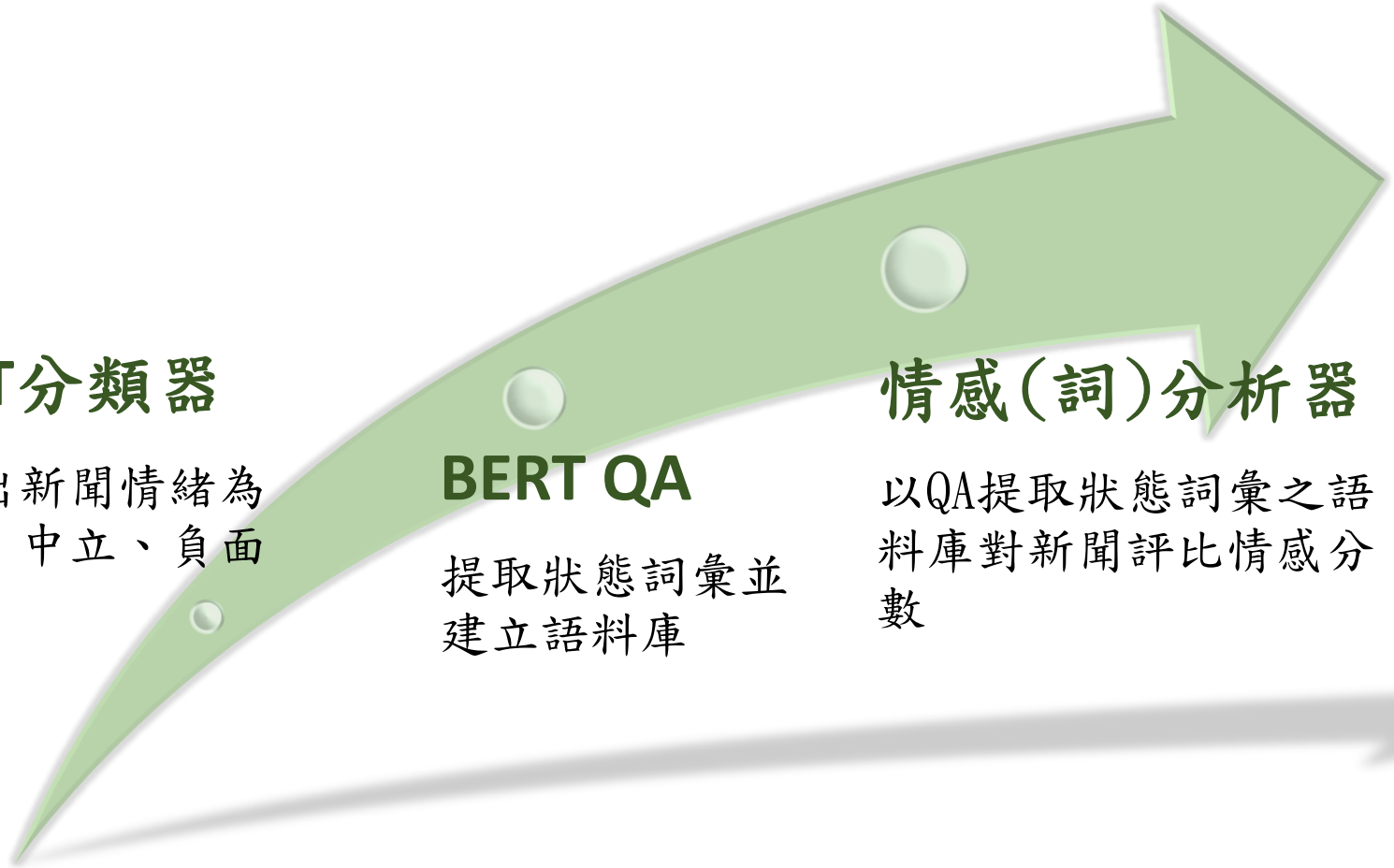
負面狀態詞彙

遞延#衰退
疫情波及#年減
能見度不明朗#跌
持平上季#遞延#仍衰退
衝擊有限
減、5G未甜#年減



四、建立個股情緒指標

個股情緒指標建立流程



BERT分類器

預測出新聞情緒為正面、中立、負面

BERT QA

提取狀態詞彙並建立語料庫

情感(詞)分析器

以QA提取狀態詞彙之語料庫對新聞評比情感分數

個股情緒指標

個股情緒指標的建立流程

《個股情緒指標》

- BERT分類器 & SnowNLP 情感分析器預測結果

times	titles	label	sentimental
2020-07-08 09:38:10	外資連六買台積電逾7.5萬張。短線市值大增6,871億。	正面(2)	0.724761

- 設定二天為一新聞影響的時間範圍，將以**秒**為基礎時間單位進行個股情緒指數計算
  - 提取過去二天內所有新聞，並以新聞發布時間減去現在時點，再除以一天的秒數，作為時間差的數值

$$\Delta time = \frac{time_{now} - time_{news}}{86400} \quad (秒/秒)$$

  - 接著利用右邊三式進行試驗，計算當下時點(1分鐘內)的情緒分值，離現在時點越遠的新聞分值越小；
    其中sentimental 為 SnowNLP 情緒分析器所給的情緒分值。

$$score = \sum_i sentimental_i \cdot \left(\frac{1}{1 + DecayRate \cdot \Delta time_i}\right)$$

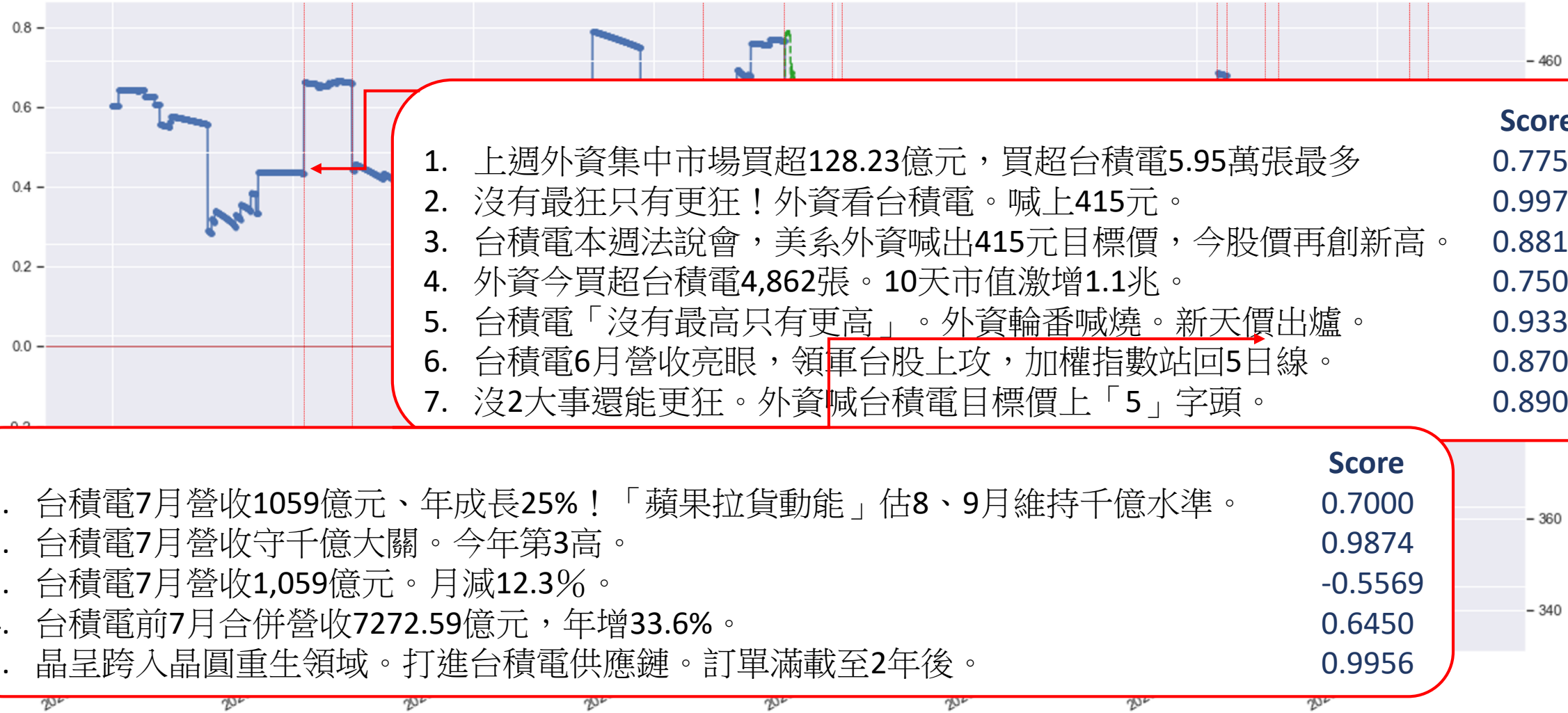
$$score = \sum_i sentimental_i \cdot \left(\sqrt{\left(1 - \frac{(\Delta time_i)^2}{DecayRate^2}\right)}\right)$$

$$score = \sum_i sentimental_i \cdot \left(ln\left(-\frac{\Delta time_i}{1.15} + e\right)\right)$$

    各個公式再取平均後為另外三組公式，總計計算六式。

個股情緒指標建立 – (2330)台積電

《個股情緒指標》



		Score
1.	上週外資集中市場買超128.23億元，買超台積電5.95萬張最多	0.7753
2.	沒有最狂只有更狂！外資看台積電。喊上415元。	0.9979
3.	台積電本週法說會，美系外資喊出415元目標價，今股價再創新高。	0.8818
4.	外資今買超台積電4,862張。10天市值激增1.1兆。	0.7500
5.	台積電「沒有最高只有更高」。外資輪番喊燒。新天價出爐。	0.9336
6.	台積電6月營收亮眼，領軍台股上攻，加權指數站回5日線。	0.8704
7.	沒2大事還能更狂。外資喊台積電目標價上「5」字頭。	0.8904

		Score
1.	台積電7月營收1059億元、年成長25%！「蘋果拉貨動能」估8、9月維持千億水準。	0.7000
2.	台積電7月營收守千億大關。今年第3高。	0.9874
3.	台積電7月營收1,059億元。月減12.3%。	-0.5569
4.	台積電前7月合併營收7272.59億元，年增33.6%。	0.6450
5.	晶呈跨入晶圓重生領域。打進台積電供應鏈。訂單滿載至2年後。	0.9956

五、提取文本數字

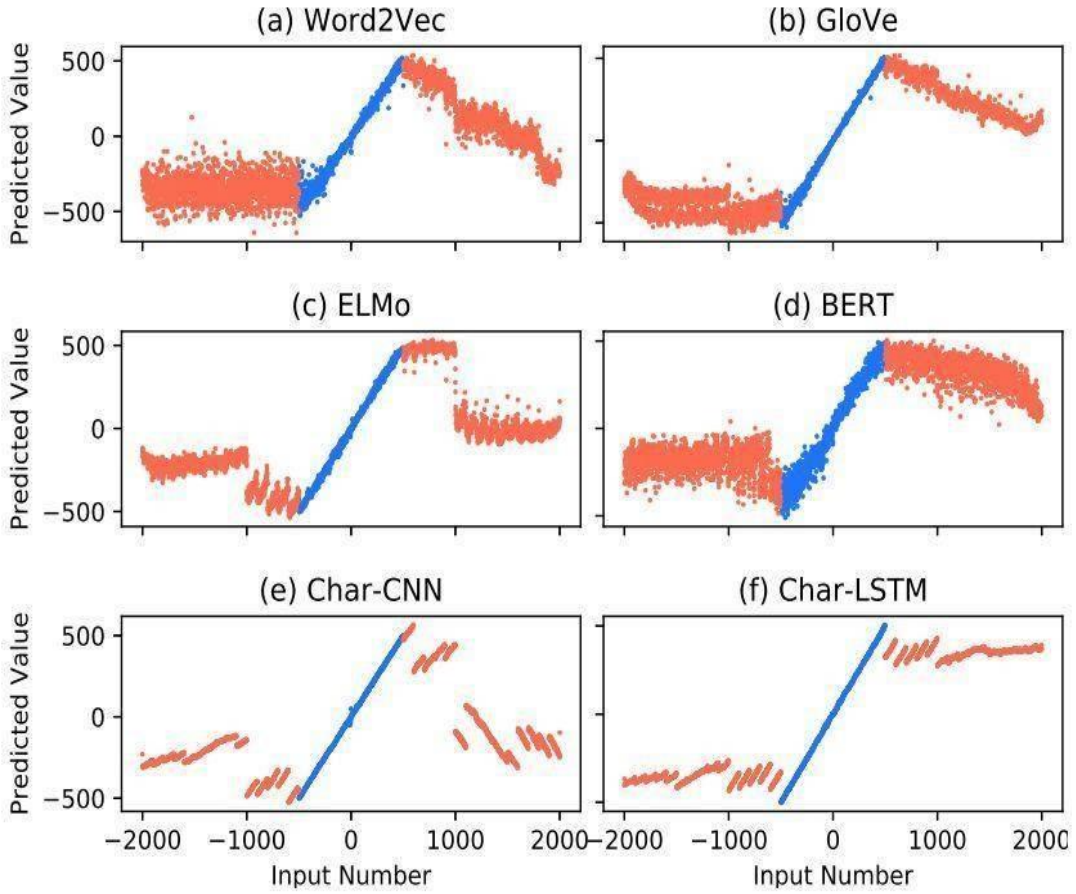
提取文本數字研究流程

動機

我們希望能爬取財經新聞中所提及的財報數據，並分析該文本中的數值大小與前值(同比、環比等)的比較；而BERT模型及其他語言模型(Language Model)目前沒有足夠的識數能力，且文本內的文本數字非純數值資料；所以將利用正規表達式(regular expression, regex)及其他NLP文本處理技巧提取數值資料

流程

1. 將文本數字轉乘數值資料
2. 提取新聞提及相關時點
3. 提取(時間；要素；數值)配對資料
   (e.g.2020/01；營收；16,085,000)



⇧ 語言模型(LM)判讀數字結果

提取文本數字、轉換為數值資料 《文本數據提取》

- 原資料：

times	titles
2020-02-10 13:46:37	台積電1月合併營收1,036.82 億 月增0.4%年增32.8%

- 轉換數字：
  - 1,036.82 億 -> 103682000000
  - 0.4% -> 0.004
  - 32.8% -> 0.328
- 提取年份：月份、季度、年份
  - 相對時點文字：去年、上月、上季等文本處理
- 結果：

標題	代號	公司	年份	時點	營收	MoM	YoY
台積電1月合併營收1,036.82 億 月增0.4%年增32.8%	2330	台積電	2020	01	103,682,000,000	0.004	0.328

前值比較與建立資料庫

《文本數據提取》

- 與前值比較同比、環比或顯示前值
- **未來應用：**

  將與前值比較結果計算一權值與情緒分析器所得之分數合併計算，從而得到較為完整的新聞情緒分值。

  （大部分公布財報資訊的新聞皆分析為中立情緒）

```
3    台積電Q1營收1036.83億元，EPS4.16元
Name: titles, dtype: object
{
    "pid": "2330",
    "year": "2020",
    "period": "Q1",
    "eps": {
        "2020Q1": "4.16",
        "2019Q4": "4.47",
        "2019Q3": "3.90"
    }
}
```

```
2    台積電11月合併營收1,036.82億 月增0.4%年增32.8%
Name: titles, dtype: object
{
    "pid": "2330",
    "year": "2020",
    "period": "11",
    "revenue": {
        "202011": "103,682,000,000",
        "201911": "-3.895%"
    }
}
```

- 建立個股財務資料庫(MongoDB)

	pid	company	year	period	revenue	revenue_MoM	revenue_YoY	revenue_QoQ	eps	grossMargin
0	2330	台積電	2019	06	NaN	NaN	0.3280	NaN	NaN	NaN
1	2330	台積電	2019	07	NaN	-0.0130	NaN	NaN	NaN	NaN
2	2330	台積電	2019	08	1.061176e+11	0.2520	0.1652	NaN	NaN	NaN
3	2330	台積電	2019	09	1.021700e+11	-0.0370	0.0760	NaN	NaN	NaN
4	2330	台積電	2019	10	1.060395e+11	0.0380	0.0440	NaN	NaN	NaN
5	2330	台積電	2019	11	1.078844e+11	0.0172	0.0965	NaN	NaN	NaN
6	2330	台積電	2019	12	1.033131e+11	-0.0423	0.1501	NaN	NaN	NaN
7	2330	台積電	2019	Q3	NaN	NaN	NaN	0.216	3.90	0.476
8	2330	台積電	2019	Q4	NaN	NaN	0.0370	NaN	4.47	0.502
9	2330	台積電	2019	Y	1.069985e+12	NaN	0.0370	NaN	13.32	NaN
10	2330	台積電	2020	01	1.036831e+11	0.0040	0.3280	NaN	NaN	NaN
11	2330	台積電	2020	02	9.339400e+10	-0.0990	0.5320	NaN	NaN	NaN
12	2330	台積電	2020	03	1.135200e+11	0.2150	0.4200	-0.020	NaN	NaN
13	2330	台積電	2020	04	9.600200e+10	-0.1540	0.2850	NaN	NaN	NaN
14	2330	台積電	2020	05	9.381900e+10	-0.0230	0.1660	NaN	NaN	NaN
15	2330	台積電	2020	06	1.208780e+11	0.2880	0.3790	NaN	NaN	NaN
16	2330	台積電	2020	H1	6.212960e+11	NaN	0.3520	NaN	NaN	NaN
17	2330	台積電	2020	Q1	3.105970e+11	NaN	0.4200	NaN	4.51	0.518
18	2330	台積電	2020	Q2	3.107000e+11	NaN	NaN	NaN	4.66	0.530

後續研究

1. 《BERT分類模型》

   嘗試其他BERT延伸模型，如**ALBERT**：參數量較BERT縮減約20倍，且在此模型出來也曾在數個NLP任務創造新的SOTA(State of the Art)；其預測速度也較BERT快速。

2. 《個股情緒指標》

   研究NLP-**事件抽取**並將之與股價進行比較，找出事件的有效性，再將情緒指標中無效事件移除，希望能得到較有效的個股情緒指標。

3. 《文本數據提取》

   研究提取財報數值資料後，如何將與前值比較的結果轉為個股情緒指標分值的加減項

   e.g. 台積電營收1,036.82 億 -> 提取後計算得年增率32.8% -> 情緒分值增加多少

謝謝聆聽

附錄

BERT 分類任務 – 標籤資料、半自動標籤

- 根據C-Money所給的**正面、負面詞彙**去生成多個人工樣本，並達成自動化標籤的效果

利多關鍵字	利空關鍵字
上升、上揚、上漲、大幅改善、不俗、不調降、升到、升高、升溫、止跌、牛市、加碼、可觀、...	力竭、下降、下修、下挫、下殺、下跌、下滑、不及、不利、不足、不振、不順、不穩、...

- 優點
  - 可以讓模型充分訓練到各種正面、負面詞彙，避免模型遇到沒有訓練過的詞彙就預測不出來
  - 可以大量**減少手動標籤**的時間，且可以讓模型學到更多資訊。

BERT 分類任務 – 模型訓練流程

取資料

```
python ./utils/data_retriever.py
```

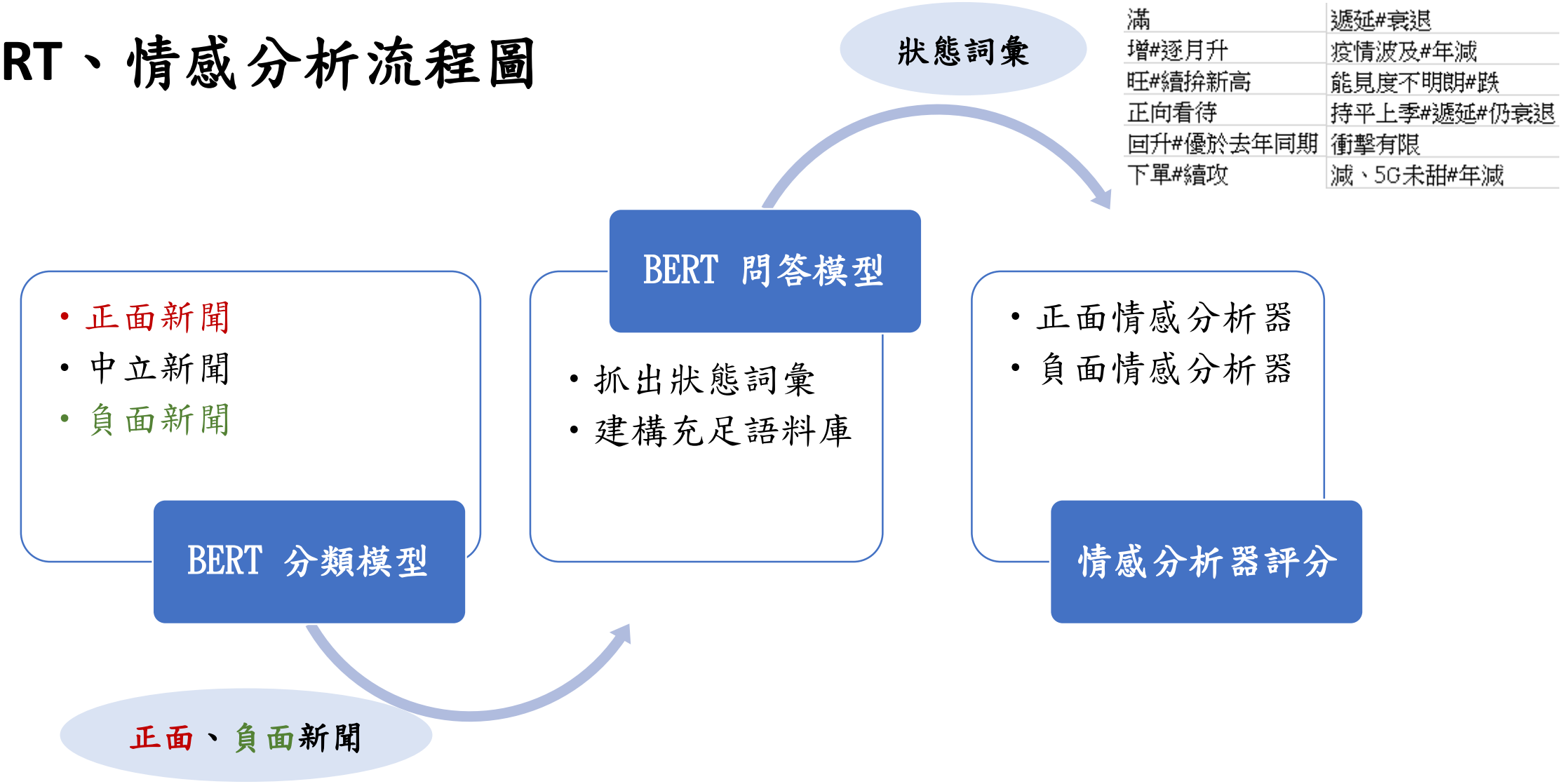
訓練模型

```
python run_bert_model.py \
    --topic operation \
    --subject titles \
    --dataset train \
    --val_dataset test \
    --data_path ./data/ \
    --model_version bert-base-chinese \
    --batch_size 8 \
    --epochs 30 \
    --patience 5 \
    --save_model_path ./model/
```

預測結果

```
python run_prediction.py \
    --topic operation \
    --subject titles \
    --dataset test \
    --data_path ./data/ \
    --model_path ./model/ \
    --train_dataset train \
    --save_results_path ./results/
```

BERT、情感分析流程圖



解釋詞彙的情感分析器

- 什麼是**詞彙**的情感分析器?

例子：模型可以**給予每一個詞彙一個情感分數**。舉例來說，模型看到「暴增」，可能會給他一個0.9的分數；看到「滿載」可能會給他一個0.93的分數。而同時看到「暴增、滿載」，可能會給他一個0.95的分數。

- 為何要詞彙的情感分析器，而不直接訓練句子的情感分析器?

原因1：避免問題發散，因此只去訓練一個新聞標題中的**關鍵狀態詞彙**

例子：伺服器、資料中心等訂單**太強**，信驊估Q2營運**彈升**，**上調**全年營收目標

原因2：比起訓練句子，單獨去訓練詞彙的情感分析器，能**更精確的控制想要給詞彙的分數**。

例子：以前一個例子來說，我們很難去控制說我們要給這一個標題多少分數。但是我們能夠一定程度的去判斷，我們應該要給「太強、彈升、上調」這三個詞彙分別多少分數。

傳統斷詞 VS BERT 問答

- 運用BERT QA模型抓出標題的狀態詞彙，並用這些狀態詞彙去做情感分析，以**增加情感分析器的精確度**

Ex: 3月疫情**爆發**，確診人數**大增** 大立光：訂單**不受影響**，出貨**正常**

- 上述的例子，如果我們用**斷詞的方法**去抓取狀態詞彙，就會連同「爆發、大增」一起抓進去。然而我們真正想抓取的只有「不受影響、正常」等含主題的狀態詞彙。