

Fintech-Text Mining and Machine Learning

# 智能新聞評分系統

第二組

2020/6/24

指導業師：詹益安

指導老師：蔡芸琇

本組組員：

東吳巨資系大四 劉品妤

台大國企所碩二 王昱達

台大經濟系大四 楊廣元

台大會計所碩二 呂明諺

# CONTENTS

- 01 現有問題描述與本專案優勢
- 02 專案流程圖
- 03 資料及樣態說明
- 04 成果展現與介紹
- 05 結論與小組分工

# 01

## 現有問題描述 & 本專案優勢

新聞重要性篩選

評分延遲

專家看法不一致

準確的新聞評分

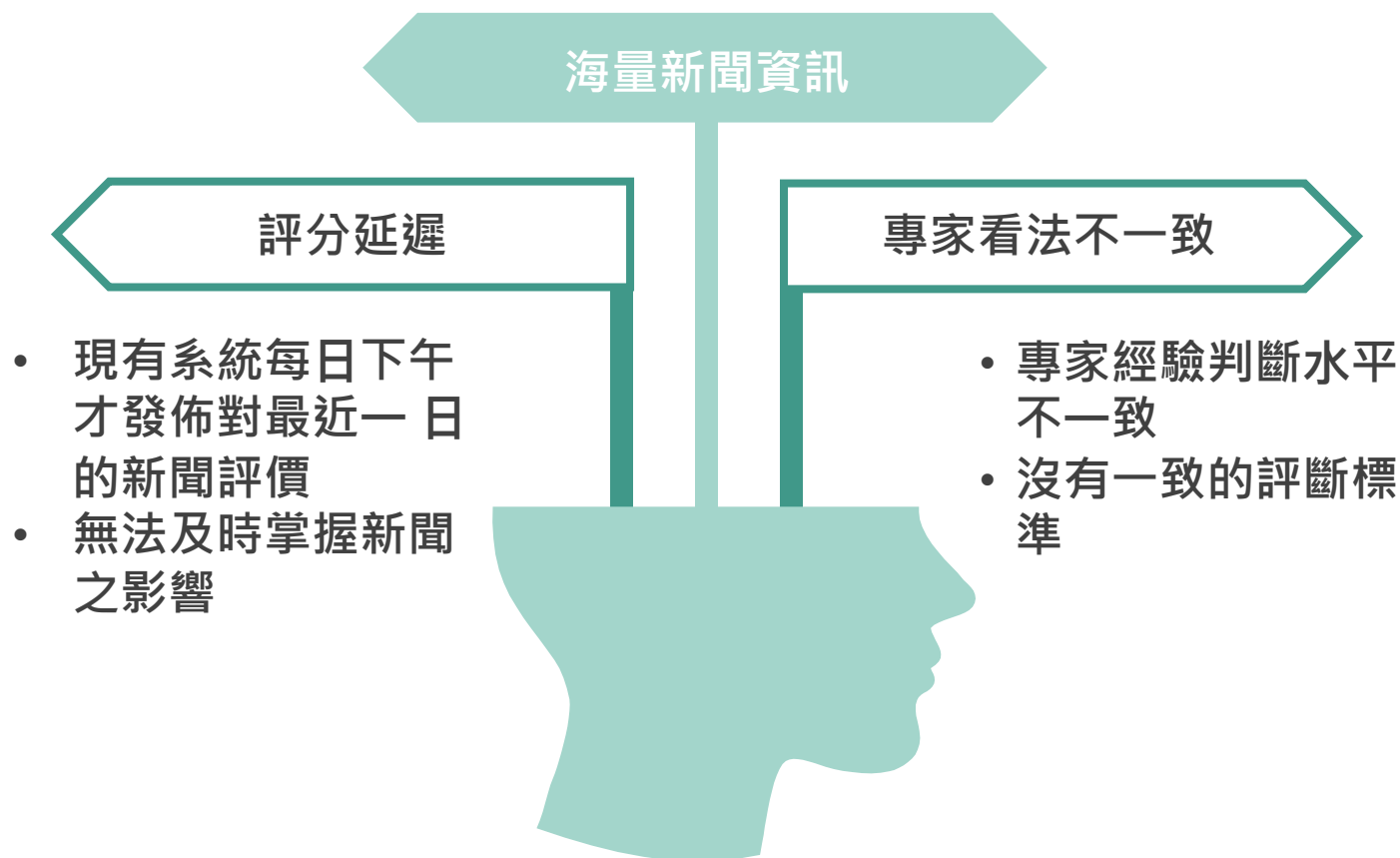
過濾不重要的訊息

提供即時資訊

可作為投資參考工具

# 看門狗評分機制現有痛點描述

- 訊息來源眾多
- 那些新聞重要?



# 本專案產品用途與競爭優勢

新聞評分

過濾訊息

109/06/23 錦明

BY [P314008DSA1](#) / ON [2020 年 6 月 23 日](#) / IN [3230 錦明](#) / [EDIT](#)

事件強度：-2

預期漲跌：下跌

發言時間：14:59:57

大事件類別：F\_市場交易

小事件類別：MT02\_董監異動

新聞內容：公告本公司董事會選任董事長與副董事長

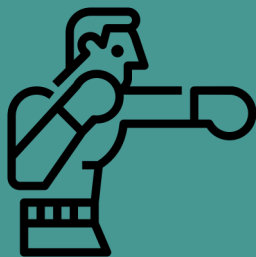
即時資訊

投資參考

## 優勢一 準確的新聞評分：

透過機器學習模型，避免多位專家評分看法不一致的偏誤

### 專家看法不一致



專家經驗判斷水平不一致



### 機器學習



利用機器學習參考各專家評分結果，降低評分變異程度

## 優勢二 過濾不重要的訊息： 建立篩選機制，讓使用者只看得到重要的新聞

### 海量新聞資訊



那些新聞重要？



### 重要新聞篩選機制



系統只展現重要的新聞  
內容及評分(-3, -2, 2, 3)

### 優勢三 提供即時資訊： 爬取最新新聞，即時更新新聞評分

#### 評分延遲

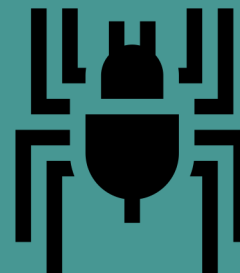


---

無法及時掌握新聞  
之影響



#### 新聞爬蟲



---

針對公開資訊觀測  
站，每兩分鐘爬一  
次，即時更新



## 優勢四 可作為投資參考工具： 透過預測預期漲跌幅指標，作為投資參考

### 大新聞V.S股價



---

大新聞不一定影響  
股價



### 股價預測



---

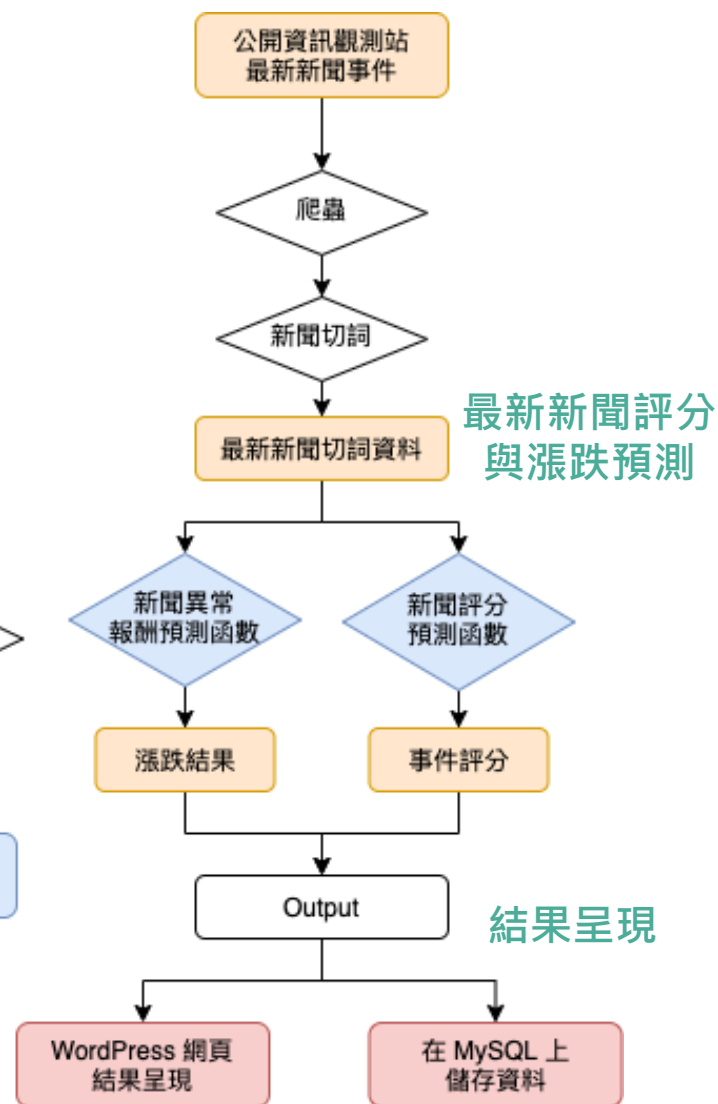
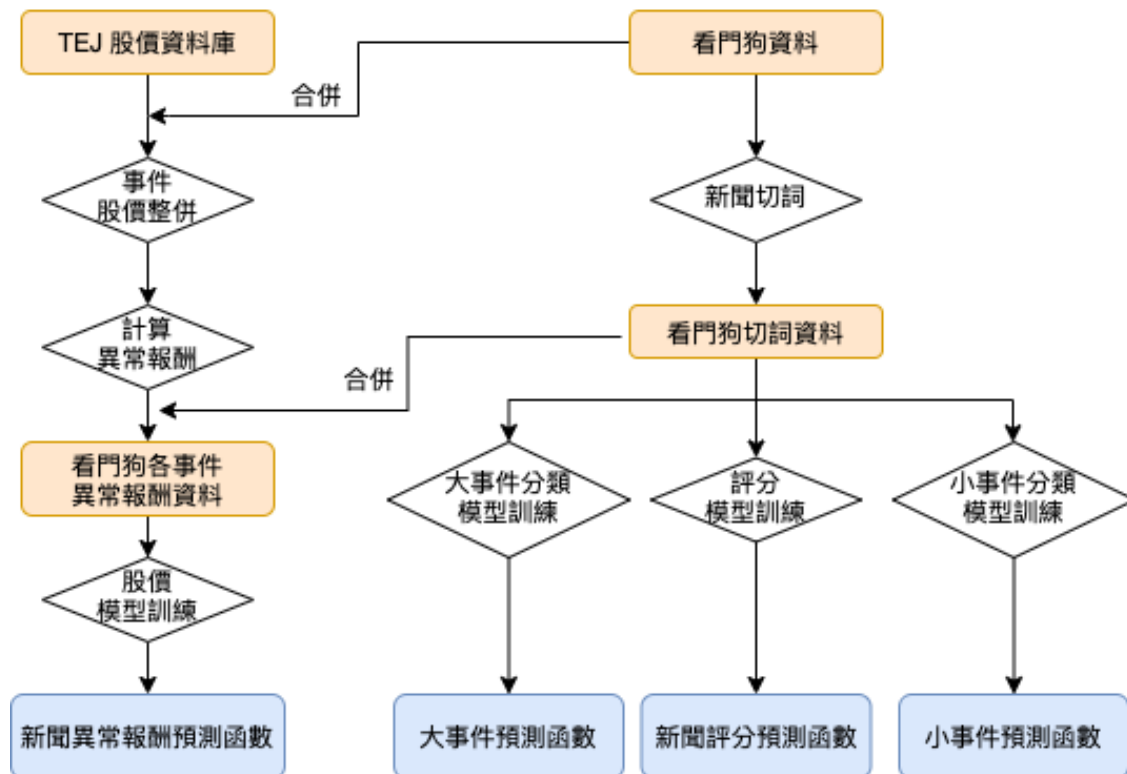
用事件分析法分析  
股價的異常報酬率

02

## 專案流程圖

# 專案流程圖

## 模型訓練過程



# 03

## 資料及樣態說明

# 資料集樣態說明

	A	B	C	D	E	F	G	H	I	J	K
1	個股代號	公司簡稱	事件日	TCRI(年/月)	事件強度	大事件類別	小事件類別	事件內容			
2	1218	泰山	20190101	6(2018/09)	0	M_經營層	MT06_高管異動	發言人林俐婉內部調動，由江巍峰接任。。			
3	1503	士電	20190101	4(2018/09)	0	M_經營層	MT06_高管異動	內部稽核主管林志強內部調動，由莊文清接任。。			
4	1504	東元	20190101	4(2018/09)	0	M_經營層	MT06_高管異動	會計主管藍俊雄內部調動，由林鴻名接任。。			
5	1709	和益	20190101	5(2018/09)	0	M_經營層	MT06_高管異動	內部稽核主管游本詮內部調動，由曾筱茜接任。。			
6	1721	三晃	20190101	7(2018/09)	0	M_經營層	MT06_高管異動	財務經理洪廷宜內部調動，由王婷渝接任。。			
7	1817	凱撒衛	20190101	6(2018/09)	0	M_經營層	MT06_高管異動	研發主管吳政峰內部調動，由朱清立接任。。			
8	2064	晉椿	20190101	7(2018/09)	-1	M_經營層	MT06_高管異動	總經理高進義離職，由陳馨接任。。發言人高進義離職，由			
9	2207	和泰車	20190101	4(2018/09)	0	M_經營層	MT02_董監異動	改派1董。董事大野勝仁(豐田自動車代表)卸任。董事長沼			
10	2330	台積電	20190101	1(2018/09)	-1	M_經營層	MT02_董監異動	辭任1董。獨立董事湯馬斯?延吉布斯卸任。			
11	2357	華碩	20190101	2(2018/09)	0	M_經營層	MT06_高管異動	總經理沈振來內部調動，由胡書賓接任。。			
12	2377	微星	20190101	3(2018/09)	0	M_經營層	MT06_高管異動	總經理徐祥內部調動，由江勝昌接任。。			
13	2442	新美齊	20190101	7(2018/09)	0	M_經營層	MO04_經營權轉讓疑慮	新美齊澄清報載提及本公司停業及隱射屬大同集團上市櫃公			
14	2724	富驛-KY	20190101	D(2018/09)	0	M_經營層	MT02_董監異動	改派1董。董事周威良(Furama Hotel International Management In			
15	2750	桃禧	20190101		0	M_經營層	MT06_高管異動	會計主管林慧茹內部調動，由楊崑岳接任。。			
16	2852	第一保	20190101		0	M_經營層	MT06_高管異動	財務經理李易致內部調動，由施冬森接任。。			
17	2888	新光金	20190101		0	M_經營層	MT06_高管異動	會計主管施貽昶內部調動，由呂雅茹接任。。			
18	3004	豐達科	20190101	5(2018/09)	0	M_經營層	MT06_高管異動	總經理邱智科內部調動，由林威村接任。。			
19	3167	大量	20190101	5(2018/09)	0	M_經營層	MT06_高管異動	研發主管宋漢釗內部調動，由商國強接任。。			
20	4104	佳醫	20190101	5(2018/09)	0	M_經營層	MT06_高管異動	總經理高省內部調動，由張明正接任。。			
21	4138	曜亞	20190101	6(2018/09)	0	M_經營層	MT06_高管異動	總經理傅若軒內部調動，由吳國龍接任。。			
22	4168	鰵聯	20190101	7(2018/09)	0	M_經營層	MT06_高管異動	總經理張東玄內部調動，由楊玫君接任。。			
23	4402	福大	20190101	D(2018/09)	0	M_經營層	MT06_高管異動	會計主管莊清揚內部調動，由何子龍接任。。			

## • 資料重要內容

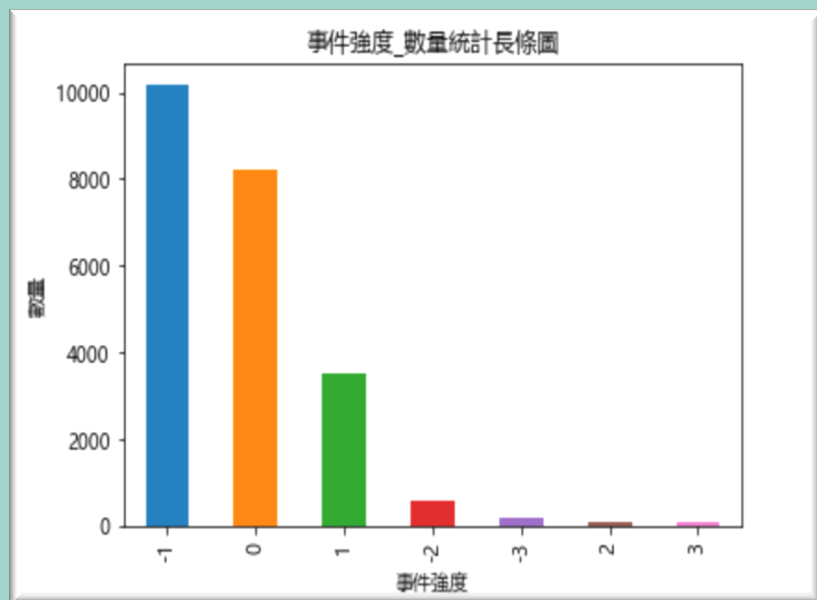
- A. 事件強度
- B. 大事件類別
- C. 小事件類別
- D. 事件內容

## • 資料樣本數

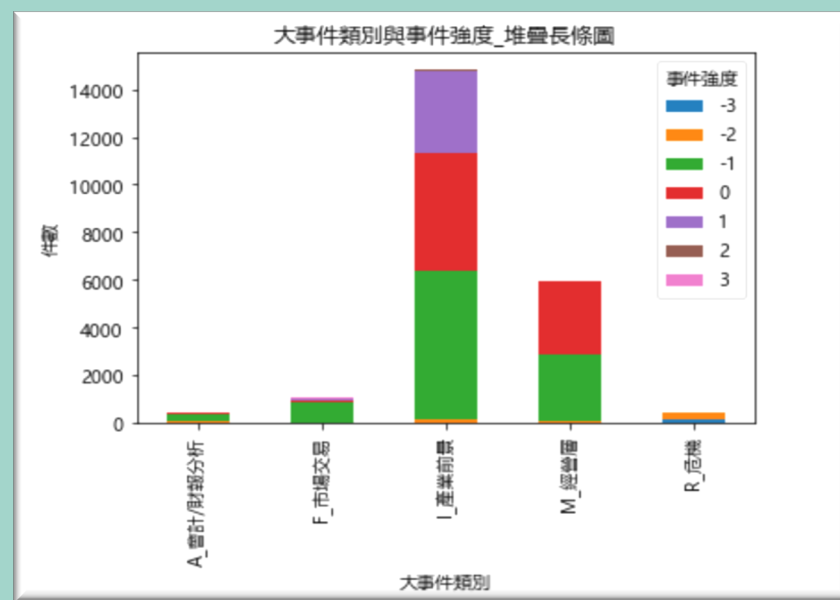
- A. 2019(1-12) : 23,703
- B. 2020(1-3) : 13,104

## 資料集樣態說明

事件強度本身樣本分布不均



大事件類別事件強度分布有所差異



# 04

## 成果展現與介紹

斷詞、爬蟲、新聞評分模型

股價預測模型、資料庫建立、網頁呈現

## 新聞資料爬蟲

- 本組採用 pandas 套件中的 read\_html 功能，對**公開資訊觀測站**的重大訊息主旨做爬蟲
- 設定成每 120 秒爬一次最新消息，將新增的新聞訊息加入資料庫
- 後續將新聞訊息做切詞處理後，放入模型跑重要性評分與股價漲跌預測

1:00~14:00進行系統維修，期間服務短暫中斷，不便之處敬請見諒

English

請輸入公司代號、簡稱，或報表關鍵字 搜尋

常用 營收 除權息 電子書 法說會 庫藏股 董監持股 獨立董事 董監酬金 ETF TDR

常用報表 基本資料 彙總報表 股東會及股利 公司治理 財務報表 重大訊息與公告 營運概況 投資專區 認購(售)權證 債券

重大訊息與公告

即時重大訊息

重大訊息綜合查詢

臺灣存託憑證收盤價彙

總表

法說會

公告查詢

券商對媒體轉載之澄清或說明

即時重大訊息

列印網頁 開新視窗 問題回報

市場別：全體公司

全體公司 上市公司 上櫃公司 興櫃公司 公開發行人公司

公司代號	公司簡稱	發言日期	發言時間	主旨	
4133	亞諾法	109/06/11	09:56:40	本公司簽訂COVID-19治療性抗體人類化合作協議	詳細資料
6679	鈺太	109/06/11	09:35:47	澄清媒體報導	詳細資料
3625	西勝	109/06/11	07:37:58	公告本公司成立第一屆審計委員會	詳細資料
3625	西勝	109/06/11	07:25:59	公告本公司109年度股東常會改選董監事當選名單	詳細資料
3073	天方能源	109/06/11	07:00:01	公告本公司名稱由「凱柏實業股份有限公司」更名為「天方能源科技股份有限公司」	詳細資料

P.S. 20200624 發現公開資訊觀測站每三分鐘才刷新一次最新資料，考慮程式所需的計算時間後，本組改設每 120 秒爬蟲一次



## 資料預處理 ( 斷詞 )

### 斷詞

- 詞是最小有意義且可以自由使用的語言單位
- 任何語言處理的系統都必須先能分辨文本中的詞才能進行進一步的處理
- 將一段中文切分出有「意義」的小單位 ( 詞 )

### 標的

- 本次報告須做處理的部分為新聞的內容

### 工具

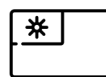
- 分別使用2種不同的斷詞系統做新聞斷詞

## Jieba vs Ckptagger



### Jieba

- Jieba 這個中文斷詞程式是由中國開發者所開發
- 可同時支援簡體與繁體的斷詞



### Ckptagger

- 中研院CKIP Lab中文詞知識庫小組開發之中文斷詞工具

TOOL	(WS) PREC	(WS) REC	(WS) F1
Ckptagger	97.49%	97.17%	97.33%
Jieba	90.51%	89.10%	89.80%

## 斷詞工具選擇

### Ckiptagger

- Ckiptagger 斷詞結果較準確
- 使用 Ckiptagger 斷詞後模型預測結果較佳
- 在切詞工具方面我們選擇 Ckiptagger

## 建立字典與數字序列

- 將文本切成一個個有意義的詞彙（Token）後，我們需要建立字典並將文本轉成數字序列
- 使用 Keras 套件的 Tokenizer，我們可以為斷詞後的詞彙建立一個字典，並將文本轉成數字序列
- 為了讓不同長度的新聞文件能轉換成相同長度的數字序列（以方便建立模型），我們需要將長度較短的新聞文件補零（Zero Padding）
- 程式碼的細節請見本組的 [github 連結](#) 及 Keras 的 [documentation](#)

## 模型建立

利用長短期記憶模型（LSTM）建立：

大事件類別分類器

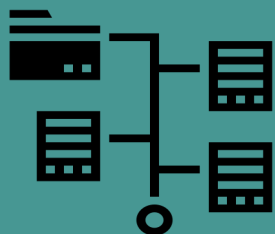
小事件類別分類器

事件強度分類器

股價異常報酬分類器

# 大事件分類器：資料分割與不平衡資料處理

## 大事件分類器



將新聞分類為以下五個大事件類別：

- 'A\_會計/財報分析'
- 'F\_市場交易'
- 'I\_產業前景'
- 'M\_經營層'
- 'R\_危機'

## 資料分割



- 所有資料的64%作為訓練資料
- 所有資料的16%作為驗證集
- 所有資料的20%作為測試集

## 不平衡資料處理



由於大事件類別的分布相當不平衡，所以我們使用了以下兩種方法來處理資料不平衡的問題

1. 使用Oversampling
2. 調整損失函數 ( loss function ) 的權重

# 大事件分類器：模型架構

Layer (type)	Output Shape	Param #
embedding_2 (Embedding)	(None, None, 128)	1280000
lstm_2 (LSTM)	(None, 16)	9280
dense_2 (Dense)	(None, 5)	85
Total params: 1,289,365		
Trainable params: 1,289,365		
Non-trainable params: 0		

Embedding layer 用來進行詞嵌入

LSTM layer 長短期記憶模型

Dense layer 作為此模型的output layer

## 大事件分類器：模型表現（在驗證集上）

Accuracy: 0.971

	A_會計/財 報分析	F_市場交易	I_產業前景	M_經營層	R_危機
Precision	0.672	0.995	0.979	0.959	0.888
Recall	0.741	0.974	0.982	0.960	0.879
F1 score	0.705	0.984	0.980	0.960	0.883



# 小事件分類器

## 小事件分類器



Created by Eucalyp

利用新聞中的文字資料，將新聞分類為以下15小事件類別：

- MT02\_董監異動
- MT06\_高管異動
- 經營層\_其他

- FS02\_股價暴跌或異常
- FS03\_其他市場交易議題
- 市場交易\_其他

- AF05\_財務警示
- AI01\_延遲公告
- 會計/財報分析\_其他

- RB01\_TCRI負向觀察
- RB02\_TCRI降等
- 危機\_其他

- IP01\_成本/產能變動或資本支出
- IS01\_營收變動或客戶/商品/通路策略
- 產業前景\_其他

# 小事件分類器：資料分割與不平衡資料處理

## 資料分割



- 所有資料的64%作為訓練資料
- 所有資料的16%作為驗證集
- 所有資料的20%作為測試集

## 不平衡資料處理



由於小事件類別的分布相當不平衡，所以我們使用了以下兩種方法來處理資料不平衡的問題

1. 使用Oversampling
2. 調整損失函數 ( loss function ) 的權重

# 小事件分類器：模型架構

Layer (type)	Output Shape	Param #
embedding_6 (Embedding)	(None, None, 128)	1280000
lstm_6 (LSTM)	(None, 64)	49408
dense_15 (Dense)	(None, 32)	2080
dense_16 (Dense)	(None, 32)	1056
dense_17 (Dense)	(None, 15)	495
Total params: 1,333,039		
Trainable params: 1,333,039		
Non-trainable params: 0		

Embedding layer 用來進行詞嵌入

LSTM layer 長短期記憶模型

Dense layer  
(3 dense layers): 進行小事件類別的分類

# 小事件類別分類器：模型表現（在驗證集上）

## Accuracy: 0.905

	AF05_財務 警示	AI01_延遲 公告	FS02_股價 暴跌或異常	FS03_其他 市場交易議 題	IP01_成本/ 產能變動或 資本支出	IS01_營收 變動或客戶 /商品/通路 策略	MT02_董 監異動
Precision	0.128	1	1	0.8	0.797	0.943	1
Recall	0.451	0.375	0.999	0.333	0.701	0.945	0.995
F1 score	0.191	0.545	0.999	0.47	0.746	0.944	0.997

	MT06_高 管異動	RB01_TC RI負向觀 察	RB02_TC RI降等	危機_其 他	市場交易 _其他	會計/財 報分析_ 其他	產業前景 _其他	經營層_ 其他
Precision	1	0.806	0.888	0.941	0.8	0.7	0.759	0.907
Recall	0.995	0.781	0.8	0.592	0.47	0.491	0.839	0.87
F1 score	0.997	0.793	0.842	0.727	0.592	0.577	0.797	0.888

# 事件強度分類器：資料分割與不平衡資料處理

## 事件強度分類器



Created by Euclyp

利用新聞中的文字資料，  
將新聞分類為以下七個事件強度：

- 3
- 2
- 1
- 0
- -1
- -2
- -3

## 資料分割



- 所有資料的64%作為訓練資料
- 所有資料的16%作為驗證集
- 所有資料的20%作為測試集

## 不平衡資料處理



由於事件強度的分布相當不平衡（極端事件：-3, -2, +2, +3 出現的頻率相對較少），  
所以我們使用了以下兩種方法來處理資料不平衡的問題

1. 使用Oversampling
2. 調整損失函數（loss function）的權重

## 事件強度分類器：模型架構

Layer (type)	Output Shape	Param #
=====		
embedding_3 (Embedding)	(None, None, 128)	1280000
-----		
lstm_3 (LSTM)	(None, 16)	9280
-----		
dense_3 (Dense)	(None, 7)	119
=====		
Total params: 1,289,399		
Trainable params: 1,289,399		
Non-trainable params: 0		

Embedding  
layer

用來進行詞嵌入

LSTM layer

長短期記憶模型

Dense layer

作為此模型的  
output layer

## 事件強度分類器：模型表現（在驗證集上）

accuracy: 0.878

	-3	-2	-1	0	1	2	3
precision	0.333	0.629	0.957	0.829	0.789	0.857	1.

recall:     **0.710**     **0.542**     0.941     0.843     0.770     **0.5**     **1.**

F1 score:	0.454	0.582	0.949	0.836	0.779	0.632	1.
-----------	-------	-------	-------	-------	-------	-------	----

# 股價預測核心方法論——事件研究法

## 何謂事件研究？

**事件研究法(Event Study)** 為研究結果之驗證方法，其起源於1960年代 Ball and Brown，及Fama, Fisher, Jensen and Roll（沈中華、李建然，2000），為近代會計及財務領域實證研究所廣泛運用之研究設計之一

## 事件研究的目的？

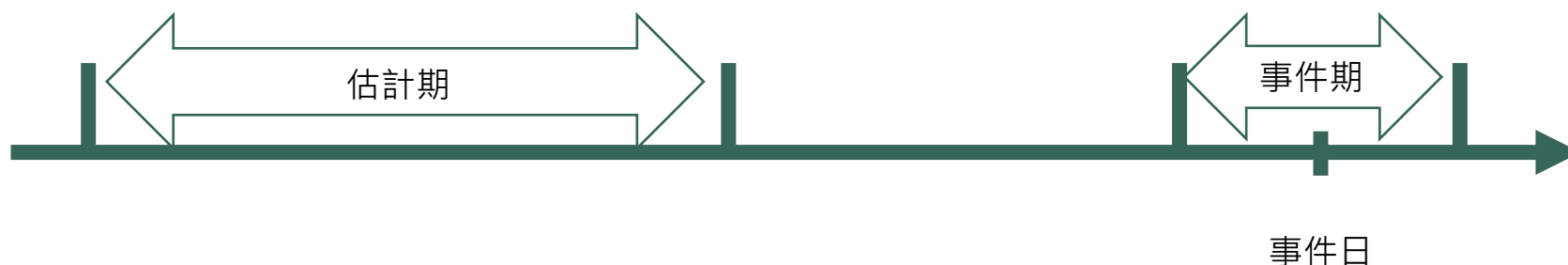
事件研究法(Event Study) 主要目的在於**利用統計方法檢定異常報酬狀況**，藉以明瞭特定事件是否**對公司股價造成影響**，並可以了解股價的波動與該事件是否相關

## 事件研究流程

1. 決定事件與事件日
2. 估計異常報酬率



## 事件日、事件期、估計期之定義



### 如何設定事件期、估計期長度

- 事件影響的區間應包括在事件期之內，如新聞發布之日。通常事件期間比發生日期（事件日）更寬廣一些，包括事件發生前後的一段時間。因為**事件發生後一段時間**的資訊能顯示**應變數（如盈利、股價）變化**的情況；而考察**事件發生後一段時間**的股價則有利於捕捉**事件前徵兆與事前洩漏資訊**所造成的影響。
- 估計期間或稱**估計窗口（estimation window）**的目的，是利用該期間的數據去估算在事件未出現情況下應變數之值，即**預期報酬率**。將預期報酬率與事件期間應變數變異後（即實際報酬率）相比較，變得出事件所帶來的異常報酬率。
- 一般而言，估計期選取要比事件期間長，本組採用年（250 個交易日）、季（60 個交易日）、月（20 個交易日）三個區間去估計異常報酬，而事件期則是事件日前後一天（明日收盤價 - 昨日收盤價）

## 異常報酬計算結果

- 本組採用TEJ股價資料，將所有看門狗中所有事件對應的股價資料合併起來，進而依序計算三種估計期的正常報酬與異常報酬
- 由於事件期長度定為兩日，因此所有的報酬率皆以兩日報酬率的形式去做計算
- 根據初步的模型預測結果發現，事件期定為一個月的準確率較高，因此後續的股價預測模型皆採用「一個月事件期」的報酬率資料

個股代號	公司簡稱	事件日	TCR(年/月)	事件強度	收盤價	明日收盤價 / 昨日收盤價	明日收盤價 / 昨日收盤價 - 1	兩日報酬率(%)	前5 ~ 245日平均兩日報酬率(%)，年平均正常報酬(%)	前5 ~ 65日平均兩日報酬率(%)，季平均正常報酬(%)	前5 ~ 25日平均兩日報酬率(%)，月平均正常報酬(%)	前5 ~ 245日(年平均)兩日異常報酬(%)	前5 ~ 65日(季平均)兩日異常報酬(%)	前5 ~ 25日(月平均)兩日異常報酬(%)	大事件類別	小事件類別	事件內容	切詞詞數	切詞結果
1218	泰山	20190101	6(2018/09)	0	19	0.972973	-0.02703	-2.7027	0.072008802	-0.126391428	0.91151213	-2.774711504	-2.576311275	-3.614214833	M_經營層	MT06_高	發言人林	19	['發言人', '林']
1503	士電	20190101	4(2018/09)	0	40.95	0.99754	-0.00246	-0.246	-0.000782104	-0.459070858	-0.042079001	-0.245220356	0.213068398	-0.203923459	M_經營層	MT06_高	內部稽核	22	['內部', '稽']
1504	東元	20190101	4(2018/09)	0	17.45	0.994302	-0.0057	-0.5698	-0.373847067	-0.699126047	0.436006766	-0.195953503	0.129325477	-1.005807335	M_經營層	MT06_高	會計主管	20	['會計', '主']
1709	和益	20190101	5(2018/09)	0	14.8	1.003367	0.003367	0.3367	-0.097104514	-0.364537588	-0.160869687	0.433804851	0.701237924	0.497570024	M_經營層	MT06_高	內部稽核	22	['內部', '稽']
1721	三晃	20190101	7(2018/09)	0	10.05	1	0	0	-0.275157057	-0.026064579	0.557660856	0.275157057	0.026064579	-0.557660856	M_經營層	MT06_高	財務經理	20	['財務', '經']
1817	凱撒衛	20190101	6(2018/09)	0	36.95	0.989276	-0.01072	-1.07239	-0.15792161	-0.081807145	0.3658829	-0.914464449	-0.990578914	-1.438268959	M_經營層	MT06_高	研發主管	20	['研發', '主']
2064	晉椿	20190101	7(2018/09)	-1	15.55	1.01634	0.01634	1.633987	-0.069460762	-0.219389167	-0.009464759	1.70344769	1.853376095	1.643451687	M_經營層	MT06_高	總經理高	32	['總經理', '高']
2207	和泰車	20190101	4(2018/09)	0	255.5	0.988	-0.012	-1.2	-0.262075048	-0.048626627	1.179844775	-0.937924952	-1.151373373	-2.379844775	M_經營層	MT02_董	改派1董	41	['改派', '1', '董']
2330	台積電	20190101	1(2018/09)	-1	225.5	0.984305	-0.0157	-1.56951	-0.024201137	-0.551439256	0.027877375	-1.545305589	-1.01806747	-1.597384101	M_經營層	MT02_董	辭任1董	20	['辭任', '1', '董']
2357	華碩	20190101	2(2018/09)	0	201.5	1.0175	0.0175	1.75	-0.240915578	-0.822803805	-0.693052721	1.990915578	2.572803805	2.443052721	M_經營層	MT06_高	總經理沈	19	['總經理', '沈']
2377	微星	20190101	3(2018/09)	0	76.4	1.005222	0.005222	0.522193	-0.026185798	-0.34696121	0.924262083	0.548379009	0.869154422	-0.402068872	M_經營層	MT06_高	總經理徐	18	['總經理', '徐']
2442	新美齊	20190101	7(2018/09)	0	12	1.0125	0.0125	1.25	0.460998352	-0.435840484	-0.448641991	0.789001648	1.685840484	1.698641991	M_經營層	MO04_經	新美齊澄	536	['新美齊', '澄']

## 模型建立—股價異常報酬分類器

### 股價異常報酬分類器：模型架構

- 若利用機器學習做股價迴歸預測，效果其實並不理想，因此改用二分法（亦即只預測漲跌）
- 預測股價的異常報酬為正值或負值
- 將「預測股價異常報酬」的問題視為二分類問題（正值 or 負值）

Layer (type)	Output Shape	Param #
embedding_4 (Embedding)	(None, None, 128)	1280000
lstm_4 (LSTM)	(None, 64)	49408
dense_13 (Dense)	(None, 32)	2080
dense_14 (Dense)	(None, 32)	1056
dense_15 (Dense)	(None, 32)	1056
dense_16 (Dense)	(None, 1)	33
Total params: 1,333,633		
Trainable params: 1,333,633		
Non-trainable params: 0		

### 資料分割狀況

訓練資料 Training set	驗證集 Validation set	測試集 Testing set
64%	16%	20%

## 模型架構與表現

### 模型架構

- Embedding layer : 用來進行詞嵌入 ( word embedding )
- LSTM layer : 長短期記憶模型
- Dense layers (4 dense layers) : 進行股價異常報酬的分類

### 模型表現

- Accuracy : 0.596

事件強度類別	負	正
Precision	0.582	0.619
Recall	<b>0.716</b>	0.474
F1 score	0.642	0.537

( 為了維持易讀性，將所有數字取四捨五入到小數點後三位數。 )

## 模型架構與表現

### 模型結果

- 若將評分與異常報酬做迴歸分析， $R^2$ 只有0.1，代表平均來說，新聞事件的變動只能解釋10%的股價變動，解釋力不足
- 若只預測上漲與下跌（也就是只分兩個類別），Accuracy Rate 只有 0.596，還有進步空間

### 結果解釋

- 新聞事件對股價影響的反應時間極短，可能事件前後幾小時內股價就已經反應完畢，而本組採用的兩日區間過長，造成解釋力不足的結果（但受限於沒有 Intraday 的股價資料，只能這麼做）

# 建立資料庫，存取最新新聞斷詞與評分預測結果



Google Sheets



	Google Sheets	MySQL
與Python 連接 方式	gsread 套件	pymysql 套件
容量	有容量限制，不適合放大量數據	無容量限制，適合放大量數據
存取量	每 100 秒有資料存取量限制，不適合作為存取頻繁的資料庫	無資料存取量限制，適合作為存取頻繁的資料庫
費用	免費試用期一年，之後要收費	開源軟體，免費使用

比較過後，本組採用 **MySQL** 作為本組的資料庫工具

## 建立資料庫，存取最新新聞斷詞與評分預測結果

- 在將資料輸入上傳到網頁的同時，我們會將下列資料存放在資料庫中，若未來公司需要使用即可直接存取

### 各項資料儲存欄位

股票代碼	公司名稱	發生日期	發生時間	預測漲跌狀態
事件強度評分	大事件 類別預測	小事件 類別預測	新聞內容	新聞切詞結果

### MySQL 資料儲存狀況

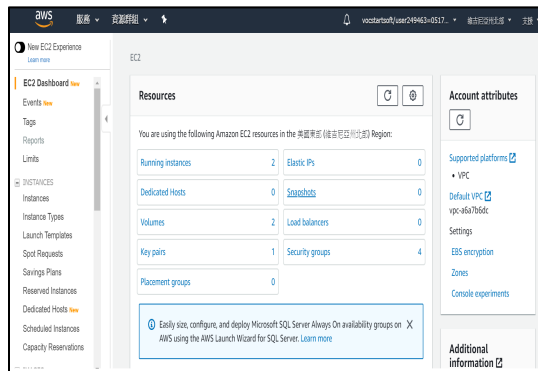
Result Grid											
Filter Rows: <input type="text" value="Search"/>											
Edit:    Export/Import:  											
id	Stock_id	Company_name	Date	Time	Intensity	Direction	Big_event_type	Small_event_type	News_content	News_cut	
1	2371	大同	109/06/23	22:38:22	0	上漲	I_產業前景	產業前景_其他	代子公司中華映管股份有限公司公告一〇九年股...	代子公司 中華 映管 股份 有限	
2	8929	富堡	109/06/23	22:37:54	-1	下跌	F_市場交易	市場交易_其他	公告本公司第三屆薪資報酬委員會委員任期屆滿	公告 本 公司 第三 屆 薪 資 報 酬	
3	1260	富味鄉	109/06/23	22:04:23	-1	下跌	I_產業前景	AF05_財務警示	公告本公司收到臺北關通知繳納罰鍰及稅款公文	公告 本 公司 收到 臺北關 通知	
4	1260	富味鄉	109/06/23	22:03:29	0	上漲	I_產業前景	產業前景_其他	補充公告本公司與富鼎產業科技股份有限公司間...	補充 公告 本 公司 與 富 鼎 產 業	
5	5203	訊連	109/06/23	21:31:27	0	上漲	M_經營層	產業前景_其他	公告本公司董事會決議委任新任薪酬委員	公告 本 公司 董事 會 決 議 委 任	
6	5203	訊連	109/06/23	21:30:48	0	上漲	I_產業前景	AF05_財務警示	公告本公司設立審計委員會	公告 本 公司 設 立 審 計 委 員 會	
7	5203	訊連	109/06/23	21:29:39	0	下跌	M_經營層	MT02_董監異動	本公司全體新任董事推選黃肇雄董事續任董事長	本 公 司 全 體 新 任 董 事 推 選 黃	
8	5203	訊連	109/06/23	21:29:17	-2	上漲	M_經營層	MT02_董監異動	本公司股東會通過提請股東常會解除本公司新任...	本 公 司 股 東 會 通 過 提 請 股 東	
9	6645	金萬林	109/06/23	21:28:37	0	下跌	I_產業前景	產業前景_其他	本公司向衛福部食藥署申請所開發新型冠狀病毒...	本 公 司 向 衛 福 部 食 藥 署 申 請	
10	5203	訊連	109/06/23	21:23:17	-1	下跌	M_經營層	MT02_董監異動	2020股東常會全面改選董事及獨立董事當選名單	2020 股 東 常 會 全 面 改 選 董 事	
11	5203	訊連	109/06/23	21:22:53	0	下跌	M_經營層	MT02_董監異動	訊連科技股份有限公司2020年股東常會決議事項	訊 連 科 技 股 份 有 限 公 司 2020	
12	4191	法德藥	109/06/23	20:40:54	0	下跌	I_產業前景	產業前景_其他	代子公司公告中國藥品(琥珀酸美托洛爾緩釋片)...	代 子 公 司 公 告 中 國 藥 品 琥 珀	
13	4431	敏成	109/06/23	20:29:02	0	上漲	A_會計/財報分析	市場交易_其他	公告本公司109年股東常會重要決議事項	告 告 本 公 司 109 年 股 東 常 會	



## 透過 WordPress 網頁來呈現我們的預測結果

- 本組採用 AWS 與 WordPress 架構個股新聞評分系統網站，並利用 wordpress-xmlrpc 套件完成自動發文的功能
- 我們預期將使用爬蟲爬取最新新聞內容後，經過我們的評分與股價預測系統模型預測，將重要的新聞（評分絕對值大於1的新聞）呈現在我們的網站中

AWS EC2控制台



發文更新畫面



手機版頁面





# 05

## 結論與小組分工

## 結論

1. 透過長短期記憶模型，我們建立了事件強度分類器、大事件類別分類器、小事件類別分類器。這些分類器對於看門狗系統的新聞內容有很好的分類能力
2. 透過長短期記憶模型，我們能從新聞事件中取得未來股價變動的一些資訊

## 未來展望

1. 使用的訓練資料侷限於看門狗系統的新聞內容。對於其它來源的新聞，模型的分類能力尚有改善空間。未來若能使用更多來源的新聞資料進行訓練，便能進一步提升此模型的適用範圍
2. 股價預測模型的表現尚有改善空間。未來若能使用更為細緻的股價資料（ex: Intraday data），便能進一步提升此模型的表現

## 小組分工

劉品妤	資料 EDA 處理、AWS 與 WordPress 網頁製作、爬蟲最新新聞資料、製作爬蟲函數、影片處理與嵌入字幕
王昱達	股價資料整理、MySQL 資料庫建立、網頁串接、主程式串接、簡報整理、影片錄音
楊廣元	新聞評分、大小事件類別與股價預測模型的建立與參數調整、製作所有機器學習的模型函數、影片錄音
呂明諺	看門狗資料清洗、斷詞與進一步的優化、製作斷詞函數、簡報整理、影片錄音

Trello 連結:  
<https://reurl.cc/Mvozzv>

Github 連結:  
<https://reurl.cc/E7vQQk>

成果網站連結:  
<http://ec2-52-87-157-212.compute-1.amazonaws.com/>

感謝聆聽 敬請指教