

新聞情緒與市場反應

第十二組

N124020003 林廷祐

N124020009 陳俐錚

N124020013 胡馨予

N124020002 陳威良

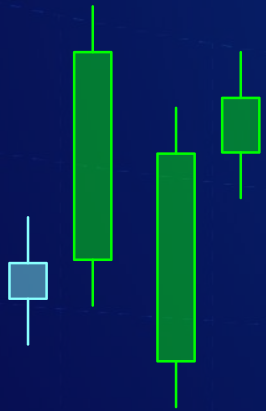
N124020008 朱浚騰

N124020010 張世安

目的與動機

我們這組希望透過自然語言處理與情緒分析技術，系統性評估新聞媒體報導與股市表現間的關聯性。我們特別關注報紙標題與內容所傳達的情緒，**如何在短期內影響投資者決策及市場走勢**，進而建立一套輔助性的市場情緒指標系統。





01

分析議題簡介



分析資訊

我們的研究聚焦於利用新聞報導中的情緒分析來預測股市的未來走勢。為了實現這一目標，我們特別挑選了報導中包含「分析」、「趨勢」、「展望」這些關鍵字的文章，因為這些詞彙通常反映了市場專家和經濟學家對未來走勢的預測和評估。

分析期間: 2024/5/1~2024/12/31

資料來源: 聯合日報-stock(股票)

使用工具: TarFlow 工作流程系統

關鍵字: 「分析」、「趨勢」、「展望」

聯合新聞爬蟲 (82)

參數設定

任務結果

選擇看板

important(重要)

sport(運動)

global(全球)

social(社會)

produce(產程)

stock(股市)

life(生活)

搜尋關鍵字

分析

趨勢

展望

排除關鍵字

以換行區隔，e.g.

雪山動物園

獅子

...

搜尋起始日期

2024/05/01

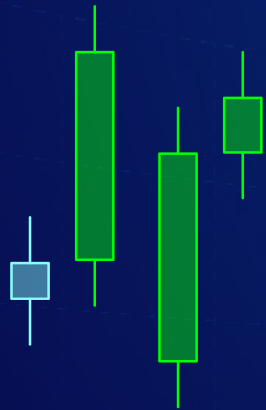
搜尋結束日期

2024/12/31

儲存更改

工作流程





02

資料清理說明



替換字串

Fed>>聯準會

NVIDIA>>輝達

Nvidia>>輝達

AI>>人工智慧

\n\n>>*

\n>>*

Sent from JPTT on my \w+>>

((http|ftp|https):/)(([a-zA-Z0-9\._-]+\.[a-zA-Z]{2,6})|([0-9]{1,3}\.[0-9]{1,3}))*(:[0-9]{1,4})*(/[a-zA-Z0-9&%_\./~]*)?>>

nt.>>

nt.>>

聯合新聞爬蟲 (82)

參數設定

任務結果

選擇看板 *

important(要聞)

sport(運動)

global(全球)

social(社會)

produce(產經)

stock(股市)

life(生活)

搜尋關鍵字 ❶

分析

趨勢

展望

排除關鍵字 ❷

以換行區隔，e.g.

雪山動物園

獅子

...

搜尋起始日期

2024/05/01

搜尋結束日期

2024/12/31

儲存更改

替換字串



組織 or 公司名稱

Fed = 聯準會 NVIDIA / Nvidia = 輝達

於Fed升息前水準

NVIDIA) H100 GPU



專有名詞

AI = 人工智慧

新AI科技巨頭



其他網址或是符號

\n\n>>*

\n>>*

Sent from JPTT on my \w+>>

((http|ftp|https):/)(([a-zA-Z0-9\._-]+\.[a-zA-Z]{2,6})|([0-9]{1,3}\

. [0-9]{1,3}))(:[0-9]{1,4})*(/[a-zA-Z0-9&%_\./~]*)?>>

nt.>>

nt>>

中文斷詞

俄烏戰爭 500	第一金 500	證券 500
人工智慧 500	台幣 500	交易 500
法說會 500	歐元 500	外資 500
中國信託 500	人民幣 500	法人 500
工業革命 500	經濟 500	成交量 500
投資級 500	金融 500	買超 500
非投資級 500	大盤 500	賣超 500
中國大陸 500	股市 500	利空 500
新興亞股 500	投資 500	利多 500
最新 500	上漲 500	獲利 500
集保戶 500	下跌 500	損失 500
高股息 500	漲幅 500	風險 500
軟著陸 500	跌幅 500	報酬 500
美元 5000	市值 500	分析師 500
美金 5000	行情 500	預測 500
財報 500	趨勢 500	展望 500
護國神山 500	安聯投信 500	指數 500

中文斷詞 (9)

參數設定

Input - 7

任務結果

選擇處理欄位 *

result

定義詞彙 ①

俄烏戰爭 500
人工智慧 500
法說會 500
中國信託 500
工業革命 500

選取字典 ①

請選擇

儲存更改

中文斷詞



名稱

中國, 大陸,

安聯, 投信,

中國, 信託,

第, 一金,



專有名詞

人工, 智慧,

高, 股息,

新興, 亞股,

非, 投資級,



習慣用語

軟, 著陸,

護國神, 山,

清除停用字

指出
表示
CMoney
eSports
PGIM
Jennison
Salesforce
MAGA
周四
上周
今年
去年
明年

ky
目前
近期
預計
預期
累積
上半年
下半年
主要
提供
認為
包括
預估

清除停用詞 (10)

參數設定

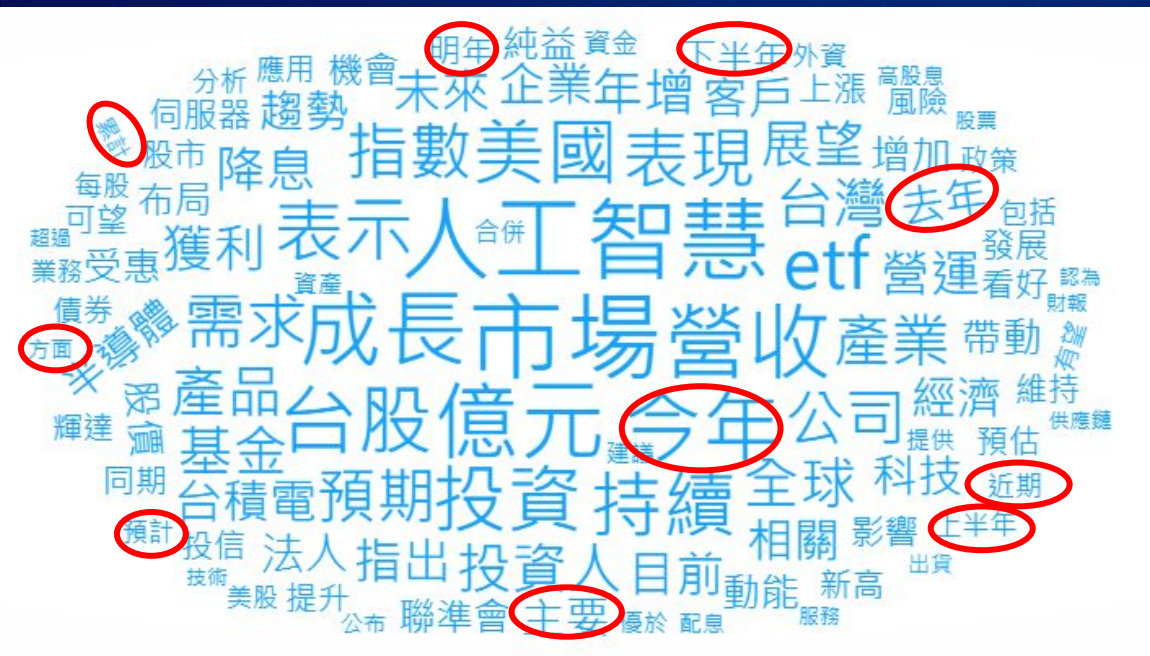
Input - 9

任務結果

語言 *	使用預設停止詞
Chinese	是
是否清除單字元 ⓘ	是否轉為小寫英文
是	是
清除英文字母 *	清除數字 *
否	是
清除換行符號 *	清除特殊標點符號 *
是	是
清除html tag *	自定義停止詞
是	指出 表示 CMoney eSports PGIM

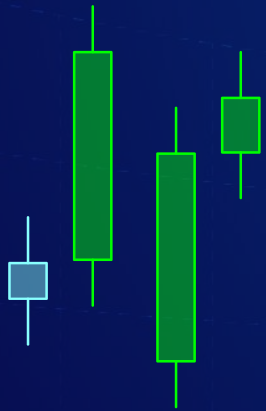
儲存更改

消除停用字



左圖為一開始的文字雲，因無設定停用字詞，文字雲內會有許多無意義的詞彙，這些詞彙對理解報導內的情緒本質沒有貢獻。

經過多次的篩選，我們可以將被紅圈圈起來的詞語，例如「今年」、「上半年」、「近期」等詞語，與判斷報導的正負面較無關聯，將這些詞語加入停用詞。



03

LIWC 情緒分析



LIWC 情緒分析

LIWC 情緒分析 (91)

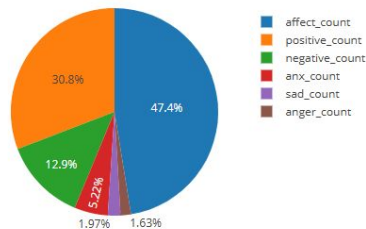
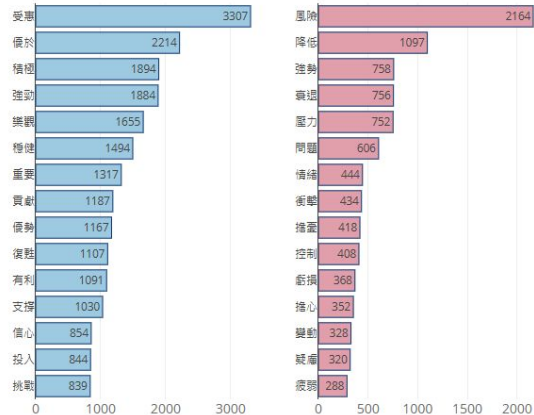
參數設定

Input - 10

任務結果

情緒分佈

情緒貢獻長條圖



全螢幕瀏覽

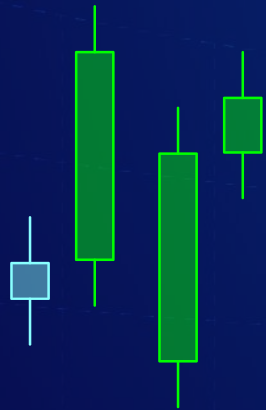
點我下載完整CSV資料

點我下載完整Rdata

點我下載完整json資料

LIWC 情緒分析

system_id ^	positive_count ^	positive_words ^	negative_count ^	negative_words ^	anger_count ^	anger_words ^	affect_count ^	affect_words ^	sad_count ^	sad_words ^	anx_count ^	anx_words ^	sentime ^
1	5	[突出, 信心, 良好, 不錯, 受惠]	4	[緊張, 攻擊, 凌亂, 遲到]	1	[攻擊]	10	[緊張, 突出, 攻擊, 信心, 氣氛, 凌亂, 良好, 遲到, 不錯, 受惠]	0	[]	1	[緊張]	
2	7	[吸引, 貢獻, 優於, 放鬆, 支撐, 有助, 樂觀]	5	[干擾, 缺乏, 疑慮, 壓抑, 對抗]	1	[對抗]	12	[干擾, 缺乏, 吸引, 貢獻, 疑慮, 優於, 壓抑, 對抗, 放鬆, 支撐, 有助, 樂觀]	1	[缺乏]	1	[疑慮]	
3	6	[值得, 最佳, 復甦, 支撐, 穩健, 保護]	3	[悲觀, 風險, 戰爭]	1	[戰爭]	10	[悲觀, 值得, 分散, 風險, 最佳, 復甦, 支撐, 戰爭, 穩健, 保護]	1	[悲觀]	1	[風險]	
4	4	[希望, 復甦, 支撐, 穩健]	6	[衝突, 擔憂, 沉重, 風險, 壓力, 強勢]	2	[衝突, 強勢]	10	[衝突, 擔憂, 沉重, 風險, 壓力, 希望, 強勢, 復甦, 支撐, 穩健]	0	[]	3	[擔憂, 風險, 壓力]	
5	1	[重要]	1	[壓力]	0	[]	2	[壓力, 重要]	0	[]	1	[壓力]	
6	2	[強勁, 充沛]	2	[降低, 風險]	0	[]	4	[強勁, 降低, 風險, 充沛]	1	[降低]	1	[風險]	



04



進階字典法情緒分析

新聞報導情緒分析

進階字典法 情緒分析 (63) 參數有做更動，建議重新執行

參數設定Input - 10任務結果

選取字典 ⓘ
NTUSD

定義正面詞彙 ⓘ
反彈

是否使用否定詞 * ⓘ
是

移除情緒詞 ⓘ
用功
活著
功德
奇妙
一拳

定義負面詞彙 ⓘ
留意
注意
謹慎
考慮

是否使用加強詞 * ⓘ
是

將新聞報導資料使用進階字典法進行情緒分析，設定移除情詞用功、活著、功德、奇妙、一拳

情緒正面字詞資料篩選

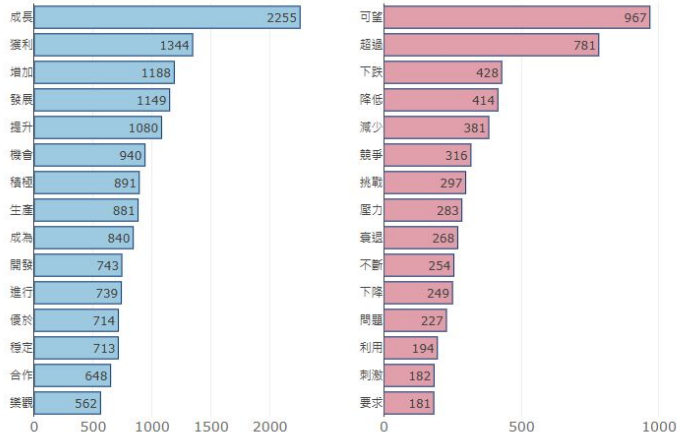
正面字詞資料篩選 (65)

參數設定

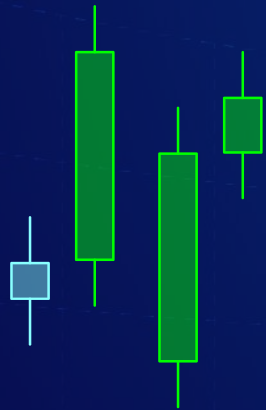
Input - 63

視覺化

情緒貢獻長條圖



使用進階字典法進行情緒分析結果，在視覺化結果中確認是否有需要移除字詞，並使用資料篩選功能保留正面情詞



05



Lexicon Based 情緒分析

Lexicon Based 情緒分析

Lexicon Bases分析會將所有重複出現的詞保留且計分

system_id	positive_count	positive_words	negative_count	negative_words	sentiment_value
1	11	[穩健, 挑戰, 吸引, 值得, 保護, 吸引, 值得, 創造, 不錯, 享受, 優勢]	2	[風險, 壓力]	9
2	6	[創造, 良好, 優勢, 創造, 優勢, 支撐]	1	[控制]	5
3	5	[冠軍, 激勵, 獲勝, 強勁, 支持]	2	[降低, 干擾]	3
4	8	[受惠, 受惠, 復甦, 支持, 有助, 樂觀, 受惠, 強勁]	1	[風險]	7
5	4	[驚喜, 希望, 智能, 強勁]	0	[]	4
6	6	[簡單, 聰明, 擁抱, 靈活, 創造, 強勁]	2	[猛烈, 擊敗]	4
7	8	[把握, 改善, 溫和, 成功, 正面, 受惠, 優於, 優於]	10	[擔心, 降低, 風險, 降低, 風險, 風險, 風險, 動盪, 風險]	-2
8	8	[強勁, 樂觀, 復甦, 復甦, 受惠, 受惠, 有助, 支撐]	0	[]	8
9	7	[突出, 主動, 最佳, 突出, 支撐, 受惠, 適時]	2	[風險, 降低]	5
10	7	[創造, 穩健, 支撐, 樂觀, 健康, 主動, 熱絡]	1	[對抗]	6

每日報導分析

將Lexicon Bases分析結果與爬蟲資訊合併
並依sentment_value分為正面(>0), 中立(=0), 負面(<0)
再分別使用分群匯總計算每天的報導數量
最後將正面與負面的資料合併



每日報導分析

最終得到聯合日報每天的情緒報導數量

artDate	sentiment_value_x	正面	sentiment_value_y	負面
2024-04-20		7		3
2024-04-21		9		2
2024-04-22		31		15
2024-04-23		29		10
2024-04-24		44		6
2024-04-25		30		7
2024-04-26		25		6
2024-04-27		11		2
2024-04-28		8		1

每日報導分析

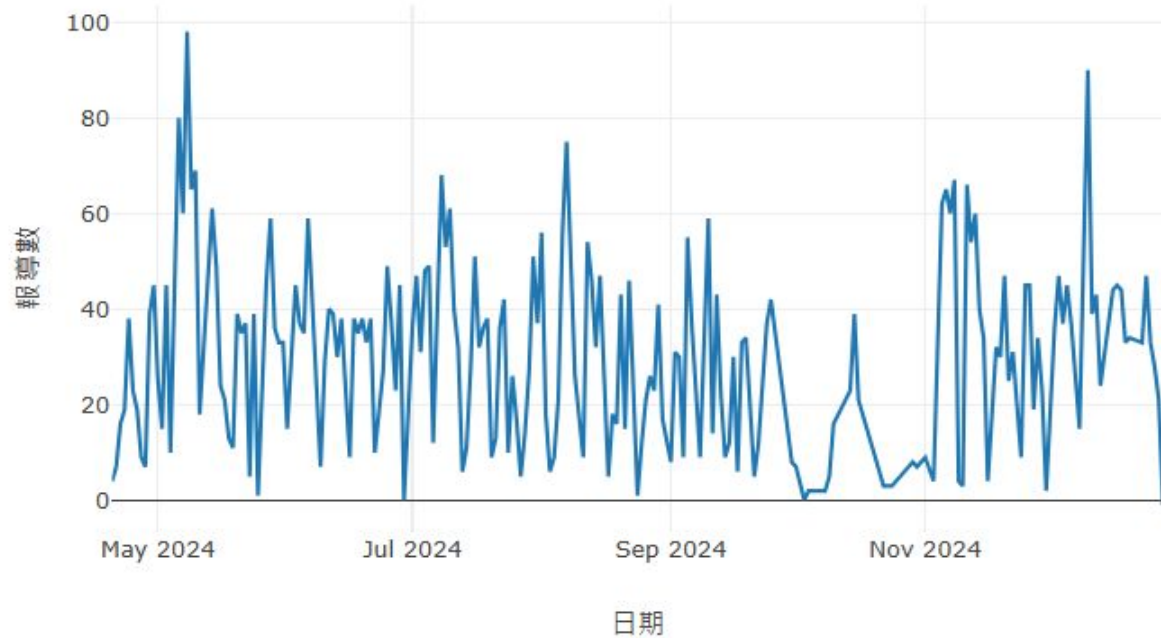
新增一個欄位, 得出每日正面報導-負面報導的數量

欄位一	計算方式	欄位二	新增的欄位名稱
sentiment_value_x	減	sentiment_value_y	Daily_count

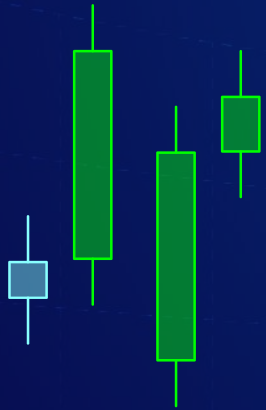
artDate	sentiment_value_x	sentiment_value_y	Daily_count
2024-04-20	7	3	4
2024-04-21	9	2	7
2024-04-22	31	15	16
2024-04-23	29	10	19
2024-04-24	44	6	38
2024-04-25	30	7	23
2024-04-26	25	6	19
2024-04-27	11	2	9

每日報導分析

每日報導(正面-負面)



將每天計算的結果用折線圖呈現
在年底才有負面>正面的狀況



06

困難與解決方式



困難與解決方式

在研究過程中我們遇到了幾項挑戰：

- 斷詞處理問題：「億」、「美元」等特定詞彙無法按照預期被分割。由於這不影響研究的核心目的，我們選擇保留原有處理方式。
- 情緒詞典的適用性：金融新聞使用的專業術語與一般情緒詞典有所差異，需要調整詞典內容以提高分析準確度。
- 新聞來源多樣性：不同媒體對同一事件的報導角度與用詞存在差異，我們透過納入多家媒體的報導以平衡這種差異。
- 噪音過濾：剔除與研究目標無關的新聞內容，保留具有市場影響力的報導。

