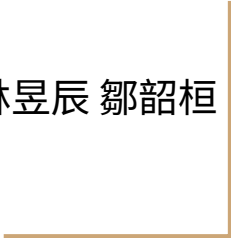




大數據與商業分析 期中專題

第十八組

吳家俊 李昊軒 許柏威 汪晁安 林昱辰 鄒韶桓



Contents

1. 篩選看漲及看跌文章
2. 透過關鍵字建構向量空間
3. 從分類到預測
4. 移動回測
5. 綜合投票結果

篩選看漲及看跌文章

資料使用說明

MEDIATEK

使用文章資料

敘述

- 文章
 - 平台：新聞、Dcard、Mobile01、PTT
 - 內容：title、content
- 時間
 - Requirement 1&2 : 2023.1~2023.12
 - Requirement 3 : 2022.6~2024.2 (全部)

選擇原因

- 使用老師提供的所有平台和時間以確保數據收集的準確性與完整度
- 文章內容及標題都有可能含有重要訊息

選擇公司標的

- 公司名稱：聯發科
- 股票代碼：2454

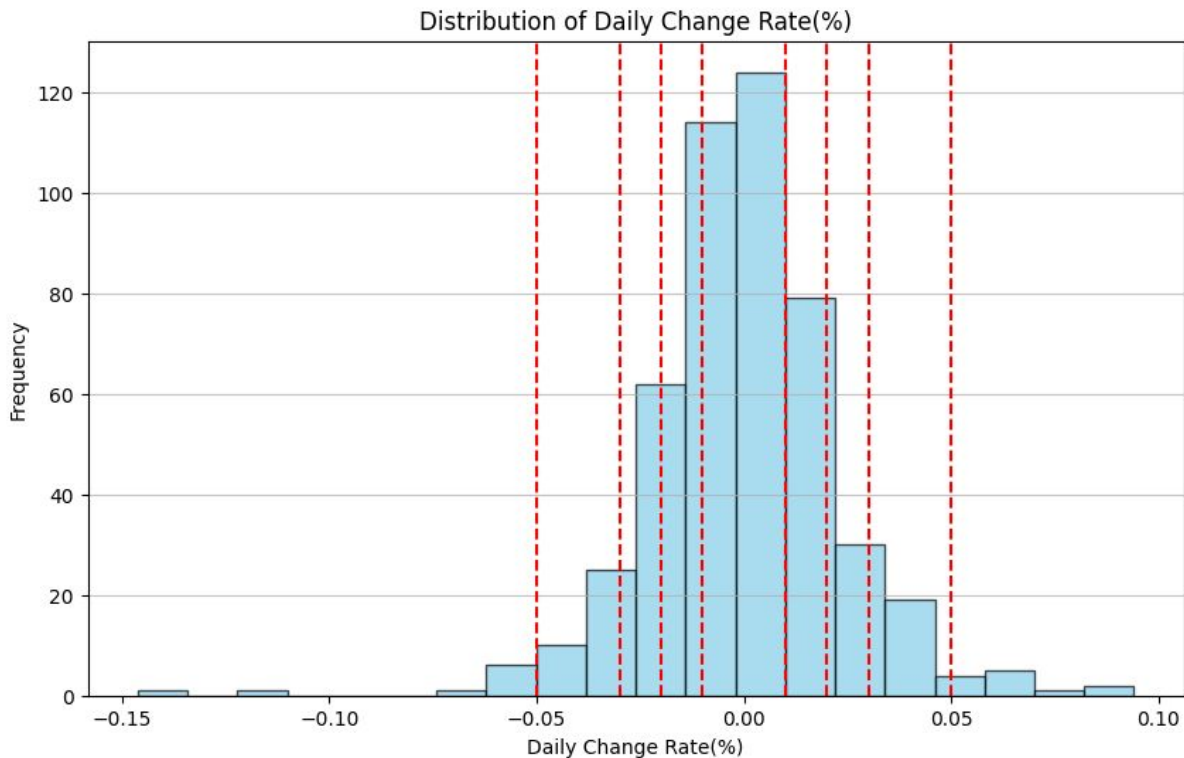
- 討論聲量高
- 有明顯漲跌幅

參數設定

- $n : 10$
- 漲跌幅度判定閾值 $\sigma : 2\%$

- 本組討論結果

參數選擇說明 ($\sigma = 2\%$)



- ± 0.05 的數據占總數的 3.72%。
- ± 0.03 的數據占總數的 16.12%。
- ± 0.02 的數據占總數的 27.07%。
- ± 0.01 範圍的占總數的 58.26%

本組認為將漲跌幅的判定閾值設定為2%
能更好地凸顯文章中的漲跌特性，同時又
不至於使得訓練模型的資料量過少。

文章標記方式

透過檢視未來10天的漲跌情況，將文章標記為看漲或看跌 (閾值 $\pm 2\%$)

title	content	label	s_name	post_time_update
聯發科、台積電利多消息	Rog Phone真的越做越帥，可惜老了不太打電動	漲	Mobile01	2022-10-05
聯發科、台積電利多消息	台積電技術未來不敢說目前是第一流的，因此一些晶片製造商對他恨的牙癢癢的，代工了不起啊，沒錯...	漲	Mobile01	2022-10-05
聯發科、台積電利多消息	不聽美國的話想重演日本半導體的下場？	漲	Mobile01	2022-10-05
聯發科、台積電利多消息	NaN	漲	Mobile01	2022-10-05
聯發科、台積電利多消息	天璣9000+的手機台灣很少看到	漲	Mobile01	2022-10-05
台股今天也太猛了吧	今天大盤都漲權值股，包含台積電、聯發科鴻海與壽險金控，下跌家數很多是正常的，畢竟這波下跌，除...	跌	Mobile01	2022-10-06
國慶行情反彈是逃命，還是空轉多呢？	未來一、二週算是安全的，只是目前多數均線往下走，震盪大迫容易短套就是了，週選擇權價平和都掉...	跌	Mobile01	2022-10-06
別人恐懼我們貪婪，抄底的時候到了！！	樓主還好嗎？樓主在4月28日發文「現在大家都在怕，是最佳買點的時候」2022年4月28日...	持平	Mobile01	2022-10-12
日月光，多？（兼論大盤及個人投資流水帳）	因為溢價購買造成的商譽總額超過兩百四十億元美金啊，商譽攤銷沒有標準答案，AMD管理階層選擇提...	持平	Mobile01	2022-10-17
美國晶片戰禁令。半導體&台積電已死（塊陶）	老美怎麼會坐視被掐脖子，沒聽聯發科說嗎，蘋果硬撐了兩年拒絕跟聯發科採購，只為了培養另一競爭者...	持平	Mobile01	2022-10-17
工程師比醫師好賺多了	有認識聯發科副理也不想做了	持平	Mobile01	2022-10-18
工程師比醫師好賺多了	我兒子高中&大學的同學，不是讀醫(牙)學系就是頂大電資，畢業後沒有人一入行就輕鬆破五百萬，除...	持平	Mobile01	2022-10-18
美國晶片戰禁令。半導體&台積電已死（塊陶）	「那天，公司禁止所有人去中國出差」一份139頁公文，讓台積電蒸發兆元市值...晶片戰爭全解析...	持平	Mobile01	2022-10-20
友達 2.5 元OK以迫嗎？	「那天，公司禁止所有人去中國出差」一份139頁公文，讓台積電蒸發兆元市值...晶片戰爭全解析...	持平	Mobile01	2022-10-20
友達 2.5 元OK以迫嗎？	三星、台積電3奈米諸事不順！魏哲家鼓勵台積電員工多休假 陳玉燭 / 新竹 2022-10-25 ...	跌	Mobile01	2022-10-25
台灣人均GDP首超日韓 居東亞之冠	公司有知名度比較可信，能不能說說30年前台積電起薪多少？聯發科起薪多少？現在台積電起薪多少？...	跌	Mobile01	2022-10-25
台積電的誠信與公司治理，是否出現問題？	笑死，那太立光、聯發科等一堆股票跌到腰斬再腰斬，是否都要出來個說法？	跌	Mobile01	2022-10-25
台灣人均GDP首超日韓 居東亞之冠	公司有知名度比較可信，能不能說說30年前台積電起薪多少？聯發科起薪多少？現在台積電起薪多少？...	漲	Mobile01	2022-10-26
台積電的誠信與公司治理，是否出現問題？	笑死...不然在這行業玩股票贏的機率比較高的原因 不就在這嗎 我之前買聯發科股價的方式就是...	漲	Mobile01	2022-10-26
台積電是不是快完蛋了？	現今台股已經到了資金可以控制大盤指數的地步，他耍你幾漲點就是幾點，跌多少點就跌多少點，可以精...	漲	Mobile01	2022-10-26

透過關鍵字建構向量空間

關鍵字選擇方法說明

Step1：文章斷詞

- 斷詞套件：Monpa
- 停用詞：stopwords_zh.txt
- 正則表達式去除中文字元以外的字元

看漲文章

看漲文章

Step2：挑選關鍵字

- 計算每篇文章各斷詞TF-IDF，並計算出chi-square或LLR（likelihood ratio）值

新聞

Dcard

Mobile01

Ptt

Step3：特徵篩選

選取chi-square/LLR值
前500/1000高的字

選取chi-square
前1000個字

取出漲跌各自最高的200
字並去重複；
使用chi2選出最後150字

選取chi-square/LLR
前1000字/1000字

建構向量空間 —— 新聞

- **Yahoo新聞 —— 股市匯市**

- 使用monpa對每一篇文章進行斷詞
- 利用chi square進行特徵篩選取出前500字

- **Yahoo股市 —— 財經新聞**

- 使用short_sentence和monpa對每一篇文章進行斷句斷詞
- 將標記好的文章集使用Likelihood Ratio Selection進行特徵挑選
 - 計算每篇文章各斷詞的df以及tf，以計算其LLR值
 - 用以判斷各斷詞是否對分辨類別起作用
 - LLR值越高，代表該斷詞有更高機率可以分辨文章的類別
- 取LLR值最高的前1000字詞作為向量空間

建構向量空間 —— Dcard

- 資料集:Dcard文章與評論(2022.3 ~ 2024.2)
- 將資料標為看漲與看跌
- 清理非中文字詞後，對資料集的title與content用monpa進行斷詞與斷句
- 利用chi-square選擇前1000字建立向量空間

建構向量空間 —— Mobile01

- 資料集:Dcard文章與評論(2022.3 ~ 2024.2)
- 將資料標為看漲與看跌
- 清理非中文字詞後，對資料集的 title 與 content 用 monpa 進行斷詞與斷句
- 取出漲跌各自 LLR 最高的 200 字並去重複
- 再使用 chi2 選出最後 150 字

建構向量空間 —— ptt

- 資料選取範圍：PTT股票版文章內容
- 使用monpa斷詞套件
 - 將文章句子進行格式化處理
 - 用short_sentence將過長的文章進行切割
- 將兩批漲跌文章的tokens合起來建立向量空間
 - 同時篩選ptt特有的stopwords(原文、記者、連結等等)
 - 使用LLR特徵篩選，保留前1000的字詞

從分類到預測

從分類到預測(新聞)

- 新聞資料集 — 股市匯市

- 使用資料

- 訓練集：2023.01~2023.09

- 測試集：2023.10~2023.12

- 模型與預測結果

- SVC, 準確率：0.73

	預測為漲	預測為跌
真實為漲	2	52
真實為跌	0	144

從分類到預測(新聞)

- 新聞資料集 — 財經新聞

- 將資料集按時間分成訓練集和測試集

- 訓練集範圍：2023.01 ~ 2023.09；標記為漲的共 104 篇、標記為跌的共 40 篇

- 測試集範圍：2023.10 ~ 2023.12；標記為漲的共 77 篇、標記為跌的共 39 篇

- 選取的訓練模型與預測準確率：

GDBoost	真實為漲	真實為跌
預測為漲	68	25
預測為跌	9	14

Random Forest	真實為漲	真實為跌
預測為漲	77	39
預測為跌	0	0

SVC	真實為漲	真實為跌
預測為漲	67	26
預測為跌	10	13

MLP	真實為漲	真實為跌
預測為漲	71	24
預測為跌	6	15

分類模型	準確率
GDBoost	0.707
SVC	0.689
Random Forest	0.663
MLP	0.741

從分類到預測(Dcard)

- trainSet 範圍: 2022.3.1 ~ 2023.6.30，標記為漲的共 161 篇、標記為跌的共 396 篇
testSet 範圍: 2023.7.1 ~ 2024.2.29，標記為漲的共 188 篇、標記為跌的共 58 篇
- 使用模型及準確率：

Decision Tree

	預測為漲	預測為跌
真實為漲	6	61
真實為跌	29	150

GDBoost

	預測為漲	預測為跌
真實為漲	29	38
真實為跌	83	96

BernoulliNB

	預測為漲	預測為跌
真實為漲	48	19
真實為跌	131	48

MLP

	預測為漲	預測為跌
真實為漲	45	22
真實為跌	115	64

分類模型	準確率
BernoulliNB	0.377
GDBoost	0.506
Decision Tree	0.649
MLP	0.454

從分類到預測(Mobile01)

- 將兩批看漲與看跌的文章合起來隨機打散
 - 標記為漲的共 945 篇、標記為跌的共 900 篇
 - 使用train_test_split將文章集拆成訓練資料(80%)與測試資料(20%)
- 使用chi-square特徵篩選出前150字詞
- 預測結果：

分類模型	BernoulliNB	SVM	DecisionTree	KNN
準確率	0.56	0.57	0.55	0.52

SVM

	預測為漲	預測為跌
真實為漲	55	134
真實為跌	25	155

從分類到預測(PTT)

- 將兩批看漲與看跌的文章合起來隨機打散
 - 標記為漲的共 741篇、標記為跌的共 609篇
 - 使用train_test_split將文章集拆成訓練資料(80%)與測試資料(20%)
- 使用chi-square特徵篩選出前1000字詞
- 預測結果：

GDBoost

準確率：0.60

	預測為漲	預測為跌
真實為漲	55	68
真實為跌	39	108

SVC

準確率：0.58

	預測為漲	預測為跌
真實為漲	37	86
真實為跌	27	120

移動回測

移動回測

Step. 1 設定起始日期

Step. 2 建立結果儲存空間



Step. 3 迴圈
(訓練集3個月、測試集1個月)

```
start_date = datetime.date(2022, 6, 1)
end_date = datetime.date(2024, 1, 31)
current_date = start_date
Score = []
result = []
Date = []
while current_date <= end_date - datetime.timedelta(days=26):
    date_lst = []
    train_startDate = current_date - datetime.timedelta(days=90)
    train_endDate = current_date - datetime.timedelta(days=1)
    print(f"Train: {train_startDate} ~ {train_endDate}")
    test_startDate = current_date
    if current_date.strftime("%m") in ['01','03','05','07','08','10','12']:
        test_endDate = test_startDate + datetime.timedelta(days=30)
    elif current_date.strftime("%m") in ['04','06','09','11']:
        test_endDate = test_startDate + datetime.timedelta(days=29)
    else:
        test_endDate = test_startDate + datetime.timedelta(days=27)
    print(f"Test: {test_startDate} ~ {test_endDate}")
```

印出每一輪迴圈的日期



```
Train: 2022-03-03 ~ 2022-05-31
Test: 2022-06-01 ~ 2022-06-30
```

分類投票

分類投票

- Ensemble Voting - Hard voting
- 結合多個預測模型解決分類問題
- 統計看哪個類別得到的票數最多

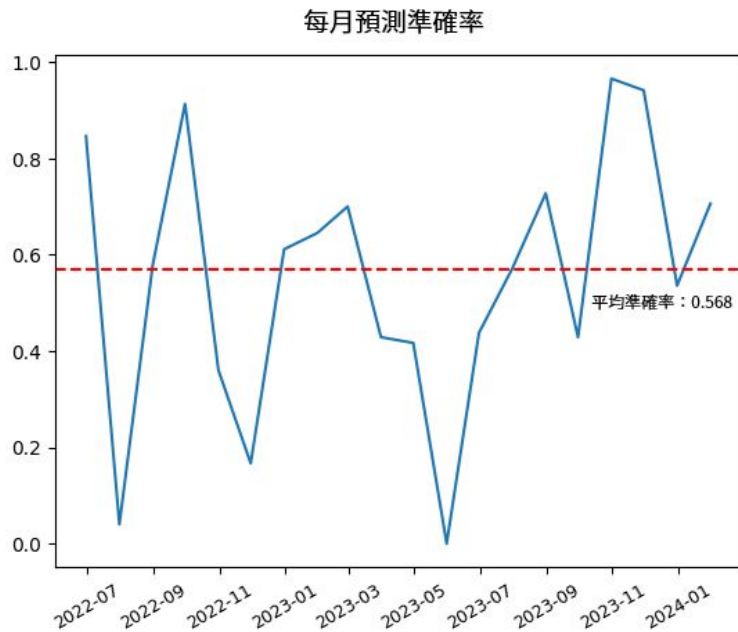
Hard Voting	Class A	Class B	Class C
Sub-Model 1	60%	30%	10%
Sub-Model 2	0%	90%	10%
Sub-Model 3	20%	20%	60%
Sub-Model 4	20%	40%	40%
Sub-Model 5	60%	30%	10%
Sub-Model 6	60%	40%	0%
	3	1	1

資料移動回測結果

- 共 471 天有漲或跌
 - 標記為漲的共 282 天、標記為跌的共 189 天
- 混淆矩陣 (Confusion Matrix) :

	真實為漲	真實為跌
預測為漲	161	83
預測為跌	94	72

- 出手率：87.0%
- 準確率：56.8%



Thanks

簡報影片連結：
<https://youtu.be/hVtYZ8DEyQI>