基於財經字典與分析指標的神經網路預測股價趨勢

Predicting Stock Price Trend Using Neural Network Based on Financial Lexicon and Technical Indicators

Mar. 2022

指導教授：張哲誠

研究生： 潘亮晴

Outline

- 緒論
  - 研究動機與背景
  - 研究目的
- 相關文獻探討
- 實驗架構與流程
- 實驗結果與分析
- 結論

Motivation (1/3)

- 台灣在2021年股票成交量高達近15億張股票

- 基本面分析
  - 宏觀經濟分析 ex. GDP、CPI
  - 行業分析　　ex. 產業現況
  - 公司分析　　ex. 財務狀況

- 技術分析領域分為資金流、原始數據、趨勢、動量、交易量、週期和波動性

Motivation (2/3)

- 情緒代表各種市場參與者的行為，市場情緒將會反映股價
  - （1）新聞透漏的基本面資訊影響投資者
  - （2）新聞引發公眾情緒，影響投資者投資決策

- 最具爭議的理論—有效市場假說（Efficient-market hypothesis）
  - 在任何時候，股票的市場價格都包含有關該股票的所有資訊
  - 價格變化是不可預測的

Motivation (3/3)

- 金融市場在某種程度上是可預測的
- 本研究認為量化後的資訊具有參考價值
- 以台積電作為研究對象，探討切割資料集的方法是否能提升模型準確率

Related Work (1/2)

- 在機器學習技術出現之前，線性統計技術提供了一種分析和預測股票的方法

- 多元線性迴歸模型(Multiple Linear Regression, MLR)
  - 擬合係數為$\mathrm{w} = w_1, w_2, \ldots, w_n$的線性模型，藉由線性近似去最小化觀察目標與預測目標之間的誤差平方和

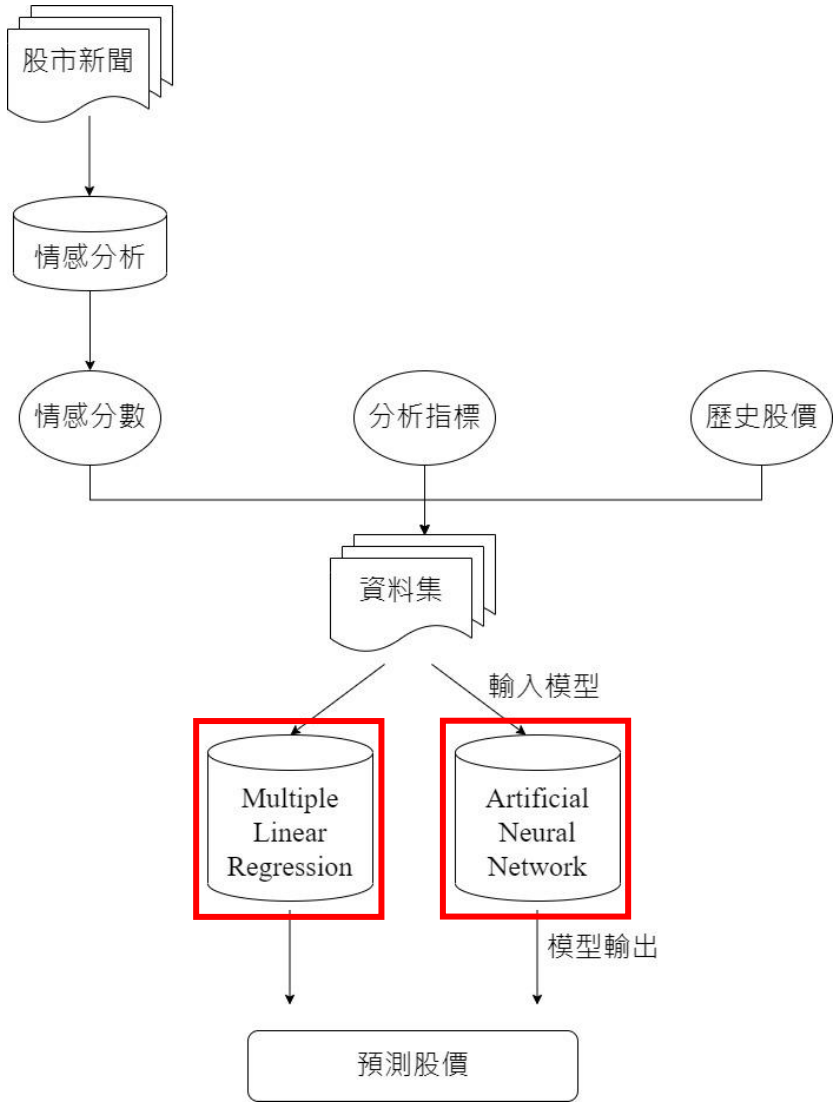
$$\mathrm{y} = w_1 x_1 + w_2 x_2 + \cdots + w_n x_n + b$$

x_i為輸入變量，w_i為係數，b為殘差，y為輸出變量。

Related Work (2/2)

- 股價預測中的機器學習任務大致分為監督學習和無監督學習
- 使用深度人工神經網路（Artificial Neural Network, ANN）進行多變數分析已成為金融市場分析中佔主導地位和流行的分析工具
- 特點是迭代速度快、學習精度高、能夠處理非線性關係數據

Architecture (1/15)
實驗架構



Architecture (2/15)
資料蒐集—股市新聞與分析指標

FugleAPI

2017年1月1日~2021年12月31日

濾除重複

基本面	計算期間	分析指標
營收	月	營收
EPS	季	每股盈餘
利潤比率	季	毛利率
		營業利益率
		稅後淨利率
ROE及ROA	季	股東權益報酬率
		資產報酬率
成長能力	季	每股盈餘季成長率
		毛利季成長率
		營業利益淨成長率
		稅後淨利淨成長率
經營能力	季	應收帳款週轉率
		存貨週轉率
		不動產及設備週轉率
		總資產週轉率
償債能力	季	流動比率
		速動比率
		利息保障倍數

表 分析指標

Architecture (3/15)
資料蒐集—歷史股價

- 使用Python套件Finmind下載2017年1月1日至2021年12月31日的歷史股價資料。

欄位	欄位敘述
date	當日日期
stock_id	股票代號
Trading_Volume	成交股數
Trading_money	成交金額
open	開盤價
max	最高價
min	最低價
close	收盤價
spread	漲跌價差
Trading_turnover	成交筆數

表 歷史股價

Architecture (4/15)
資料預處理—斷詞&工具比較

- 斷詞正確才能正確瞭解中文語句的語意
  - Ex.下個月|可|獲|利多|少？
    盤勢|呈現|利多|走向

- 結巴(jieba)
  - 斷詞快速
  - 內建簡體中文字典
  - 需要自行建立使用者字典提升精準度

- CKIP tagger→採用
  - 中研院研發
  - 斷詞較慢
  - 斷詞準確
  - 可使用自建字典強迫分詞

Architecture (5/15)
資料預處理—斷詞

- 採用CKIPTagger，其在繁體中文斷詞上表現優良。

文本內容	台積電即將在19日召開法人說明會，由於大立光法說釋出利多訊息，市場關注台積電對下半年科技業景氣看法及蘋果(Apple)新機拉貨力道，加上台積電現金股息即將發放，外資領到股息後會回頭買股還是匯出，也關乎指數未來發展方向。
斷詞後	台積電 即將 在 19日 召開 法人 說明會 ， 由於 大立光 法說釋出 利多 訊息 ， 市場 關注 台積電 對 下 半 年 科技業 景氣 看法 及 蘋果 (Apple) 新機 拉貨 力道 ， 加上 台積電 現金 股息 即將 發放 ， 外資 領到 股息 後 會 回頭 買股 還是匯出 ， 也 關乎 指數 未來 發展 方向 。

表 斷詞前後對照表

Architecture (6/15)
資料預處理—斷詞

- 取得台股所有股票列表，將台股總覽加入CLIPTagger使用者自建字典中。

產業類別	股票代碼	股票名稱	交易所	發行日期
ETF	0050	元大台灣50	twse	2021-10-05
ETF	0051	元大中型100	twse	2021-10-05
ETF	0052	富邦科技	twse	2021-10-05

表 台股總覽範例

Architecture (7/15)
資料預處理—文句特徵提取

- 從富果網取得的新聞有以下情況：
  (1) 新聞只有提到台積電
  (2) 新聞討論台灣股市整體現況
  (3) 新聞提及眾多股票公司

- 新聞斷句
  - 開頭到句末標點符號「。、？、！」視為一個句子
  - 分別存入對應的股票公司檔案中。

新聞

台積電
台積電ADR 30日隨著美國科技股回穩，台積電今跳空開高2.5元為228.5元，盤中穩步走揚至230元之上，終場收復231元季線關卡，最高觸及233.5元。

康那香、美德醫療-DR 、恆大
近日台灣飽受空汙之苦，國人防疫意識興起，康那香(9919) 、美德醫療-DR (9103) 、恆大 (1325) 挾著不織布、口罩題材持續發燒，股價逆勢揚升，康那香漲幅達2%，美德醫大漲5.7%，恆大收漲1.5%。

圖 文句特徵提取範例

Architecture (8/15)
資料預處理—文句特徵提取

- 關鍵字分類
  - 與新聞斷句同理，將斷句分別存入對應的關鍵字檔案中。

- 新聞兩大分類：(1)台積電相關新聞 (2)大盤相關新聞。

新聞分類	句子具有正向情感則台積電股價上漲之關鍵字	句子具有正向情感則台積電股價下跌之關鍵字
台積電相關新聞	半導體、電子、晶圓、台積電、奈米	三星、英特爾
大盤相關新聞	台股、大盤、外資、投信、自營商、法人、加權指數、台灣、景氣、美股、美國	

表 新聞分類

Architecture (9/15)
資料預處理—情感字典

- 透過帶有情感的詞語計算出整個文本帶有正向情緒或是負向情緒
- 字典中的詞語量與精確度越高，越能準確判別一個文本的情感

正面詞	負面詞
熱錢湧入	訂單流失
擴大市佔	商譽受損
法人看好	收賄弊案
財報亮眼	資金出逃
銷售一空	擦鞋童

表 自建字典範例

字典	正面詞數量	負面詞數量	總數
NTUSD	2,812	8,276	11,088
FinDict	412	237	649
總計	3,159	8,472	11,631

表 自建字典詞數量

Architecture (10/15)

資料預處理—情感分數計算

- 每比對到一個斷詞與正向情感特徵詞語一致，則計分1分；
- 每比對到一個斷詞語負向情感特徵詞語一致，則計分(-1)分。
- 最後對當日的所有文本後將分數加總。

表 情感分數計分範例

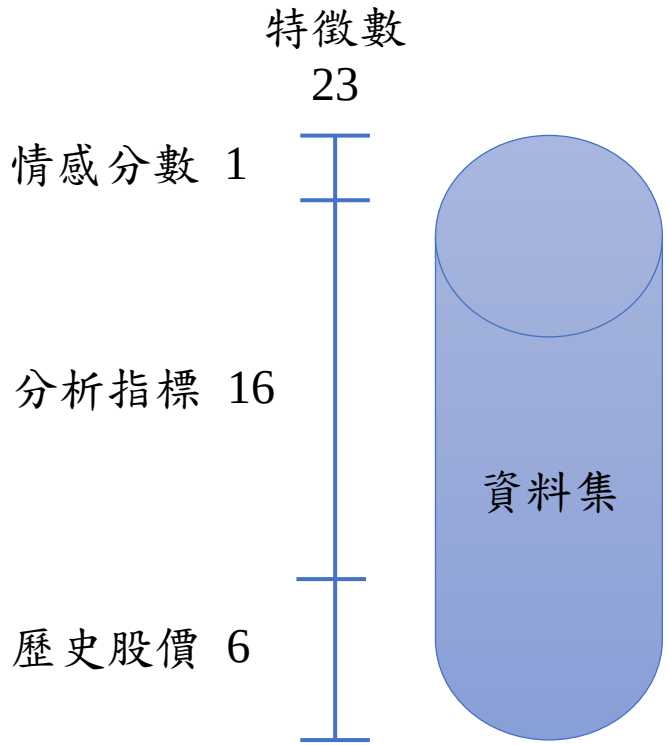
日期	2017-02-13	2017-02-13	總計
文句斷詞	在 台幣升值 態勢 ， 熱錢湧入 明顯 下 ， 台股 今日 再 收上 9700 關卡 壓力 ， 目前 台積電 正在 區間 上緣 位置 ， 一旦 上 攻 突破 前 高 ， 台股 將 持續 加速 上漲 幅度 。	台北 晶圓 代工 大廠 台積電 (2330) 將 在 本周 二 (14 日) 舉行 董事會 ， 可望 公布 股利 政策 ， 市場 預期 ， 台積電 今年 現金股利 將 從 去年 6 元 起跳 ， 上'看 8 元 ， 利多 帶動 下 台積電 今 (13) 日 股價 走強 ， 漲幅 逾 2% ， 站穩 多頭 均線 之上 。	
情感比對	0 1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 1 0 1 0 0 0 1 1 0 0 0	
情感分數	5	8	13

Architecture (11/15)
資料預處理—機器學習資料集

資料與資料標籤日期調整

預測交易日之標籤	預測交易日之特徵資料
2021-12-24(五) 收盤價	23日歷史股價 23日新聞情感分數 11月分析指標 第三季分析指標
2021-12-27(一) 收盤價	24日歷史股價 26日新聞情感分數 11月分析指標 第三季分析指標

表 預測交易日資料標籤範例



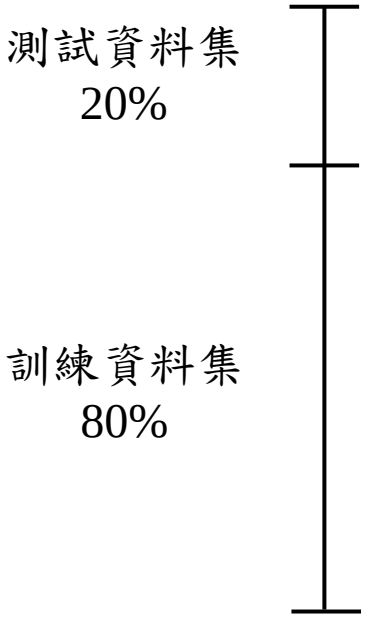
2017-01-01~2021-12-31
共1223筆資料

Min-max normalization

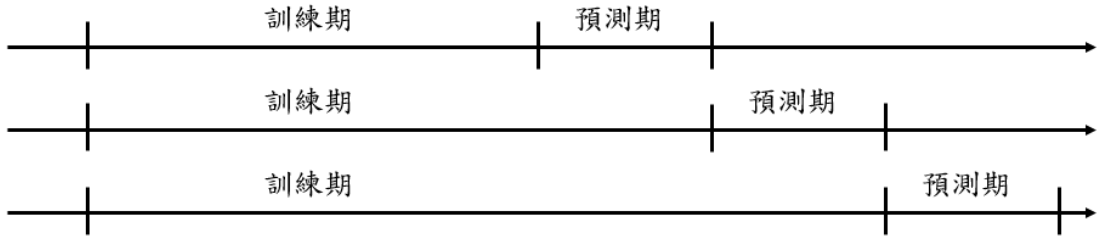
Architecture (12/15)

資料預處理—機器學習資料集切割法

1. 傳統法



2. 滑動視窗法



期數編號	訓練資料集	訓練資料筆數	測試資料集	測試資料筆數
1	2017/01/03 ~ 2020/12/30	979	2021/01/03 ~ 2021/01/17	10
2	2017/01/03 ~ 2021/01/14	989	2021/01/17 ~ 2021/01/31	10
3	2017/01/03 ~ 2021/01/28	999	2021/01/31 ~ 2021/02/23	10

Architecture (13/15)
實驗方法

模型類別	特徵	資料集	模型架構
傳統法-MLR	情感分數 分析指標 歷史股價資訊	傳統法	MLR
傳統法-ANN		傳統法	ANN
視窗法-ANN		滑動視窗法	ANN

Architecture (14/15)
成果評估方法— 評估指標

- 使用均方根誤差(Root Mean Squared Error, RMSE)機器學習模型評估指標，評估預測成果。

- RMSE可以透過預測資料與真實資料間的差距評估預測效能

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{m}(y_i - \hat{y_i})^2}$$

- 其中m為預測資料總數，y_i為真實資料，即真實收盤價；$\hat{y_i}$為預測資料，即預測收盤價。

Architecture (15/15)
成果評估方法— 模擬投資

- 投資策略：

	預測為漲	預測為跌
持有股票	不做投資動作	賣出股票
未持有股票	買進股票	不做投資動作

- 績效報表：

總損益	交易總次數	交易產生總費用	勝率
最大損失	最大獲利	投資報酬率	平均交易報酬率

Experiment Results (1/6)
情感分數

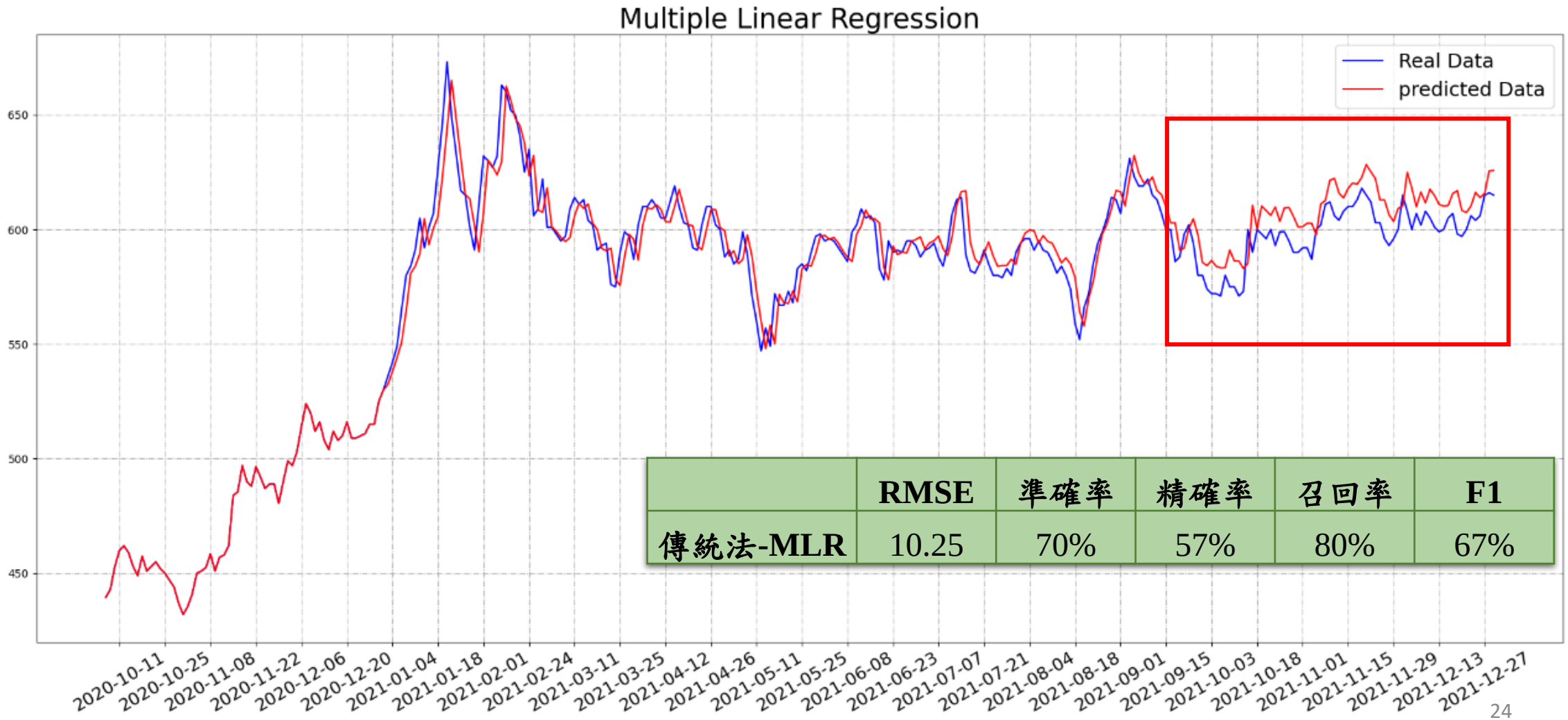
- 字典法預測成效：

	準確率	精確率	召回率	F1
字典法	57%	96%	57%	72%

- 區間準確率：

情感分數區間	準確率	出現次數
600~699	66.67%	3
500~599	66.67%	3
400~499	80.00%	5
300~399	50.00%	20
200~299	67.95%	79
100~199	54.96%	242
0~99	56.83%	815
-99~-1	45.10%	56
總計		1223

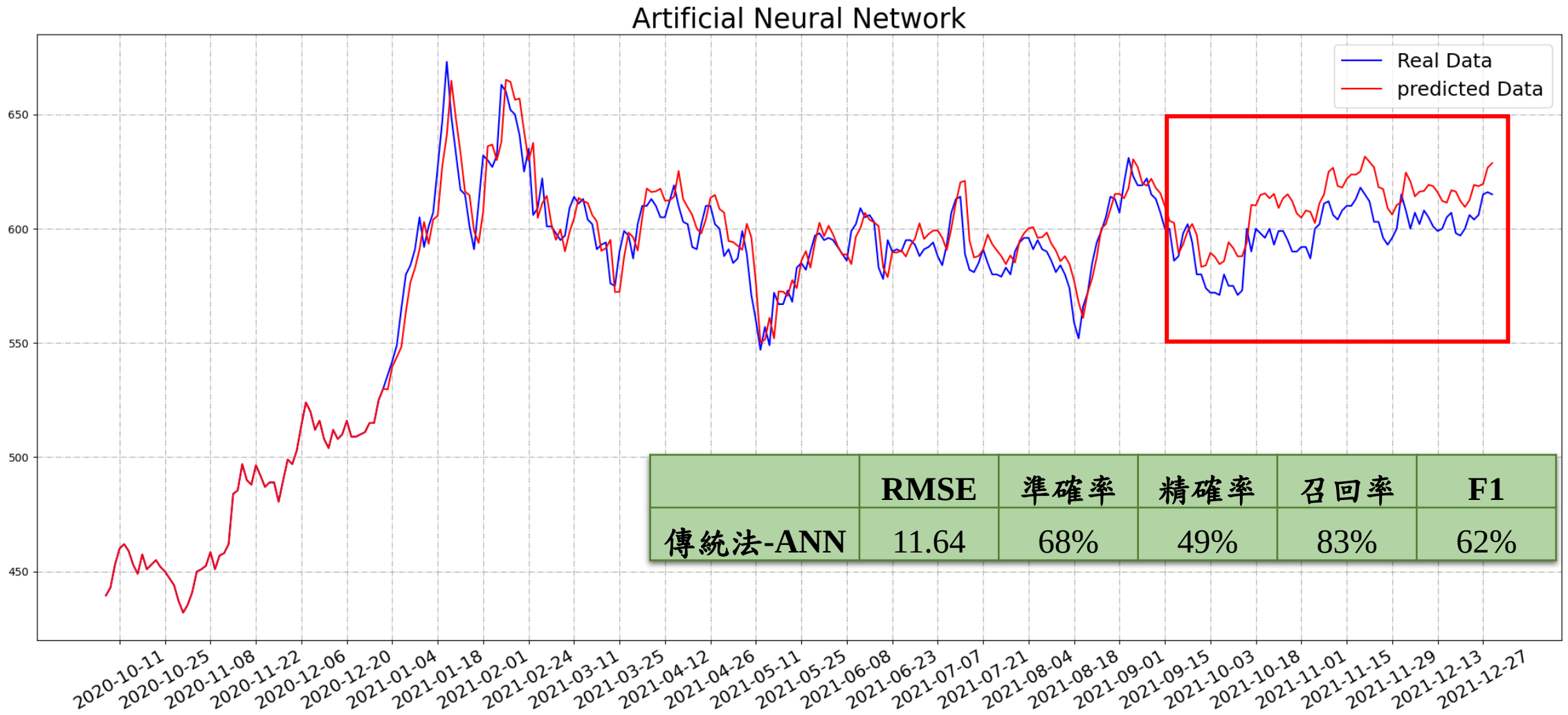
Experiment Results (2/6)
傳統法-MLR



	RMSE	準確率	精確率	召回率	F1
傳統法-MLR	10.25	70%	57%	80%	67%

Experiment Results (3/6)

傳統法-ANN



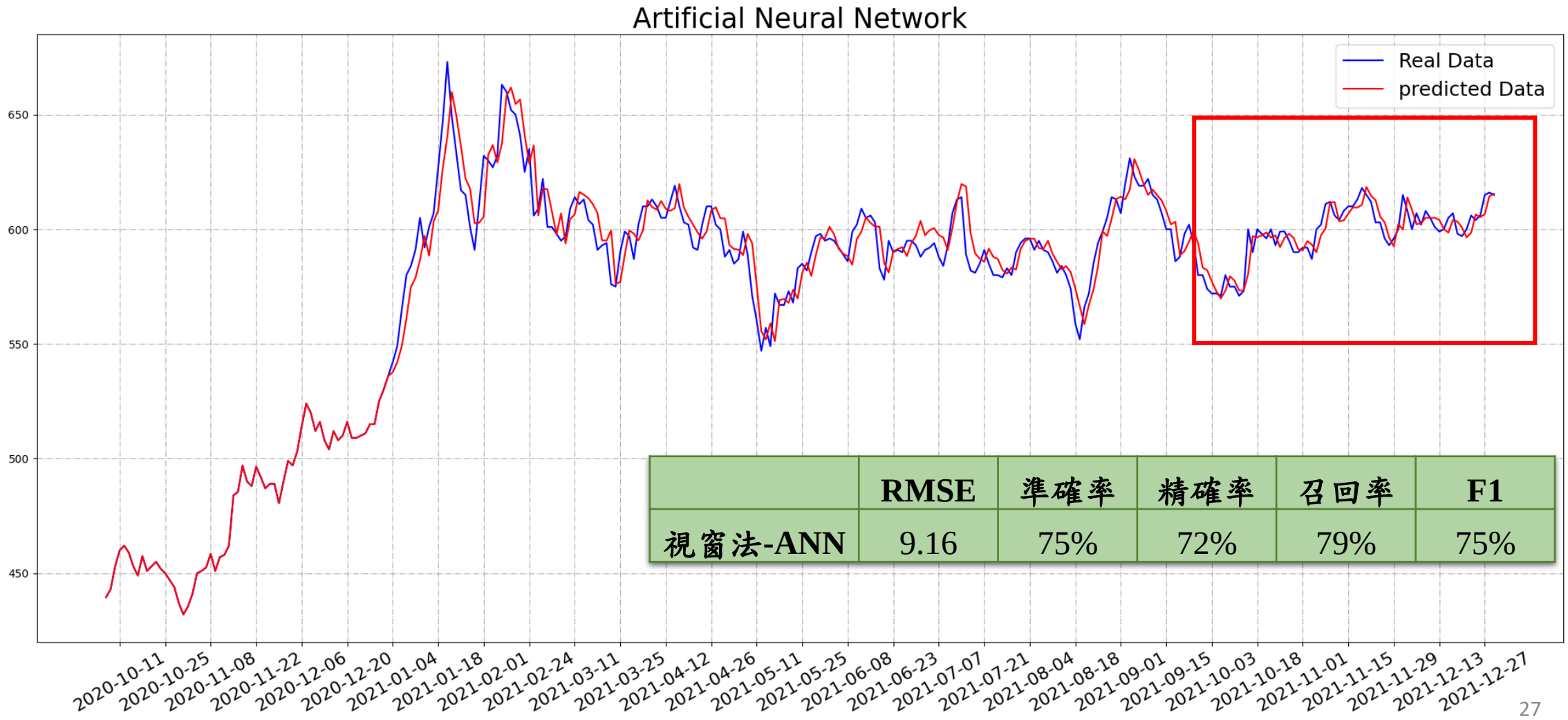
	RMSE	準確率	精確率	召回率	F1
傳統法-ANN	11.64	68%	49%	83%	62%

Experiment Results (4/6)
傳統法-模擬投資

	傳統法-MLR	傳統法-ANN
年份	2021年	2021年
總損益	114,837	(-103,851)
交易總次數	15	19
交易產生總費用	(-24,985)	(-32,329)
勝率	66.67%	26.32%
最大損失	(-11,982)	(-28,958)
最大獲利	33,952	50,927
投資報酬率	18.79%	(-16.38%)
平均交易報酬率	1.25%	(-0.86%)

Experiment Results (5/6)

滑動視窗法-ANN



	RMSE	準確率	精確率	召回率	F1
視窗法-ANN	9.16	75%	72%	79%	75%

Experiment Results (6/6)
滑動視窗法-模擬投資

	滑動視窗法-ANN	傳統法-ANN
年份	2021年	2021年
總損益	(-339,651)	(-103,851)
交易總次數	53	19
交易產生總費用	(-89,355)	(-32,329)
勝率	20.75%	26.32%
最大損失	(-46,933)	(-28,958)
最大獲利	33,952	50,927
投資報酬率	(-51.15%)	(-16.38%)
平均交易報酬率	(-0.97%)	(-0.86%)

Conclusion (1/2)

- 模型評估：
  - 最佳預測成效模型：視窗法-ANN
  - 最佳模擬投資模型：傳統法-MLR

- 以視窗法切割資料集能確實提高模型準確率

- 視窗法-ANN的模擬投資表現不佳
  - 情感分數特徵表現不佳，影響模型
  - 投資策略不適用此模型

Conclusion (2/2)

- 研究限制：
  - 投資標的必須有足夠樣本數的新聞報導

- 研究貢獻：
  - 自己建立的字典
  - 國內少有這種使用多特徵輸入、中文情感分析的研究

Thank You