以社群網路預測投資人行為

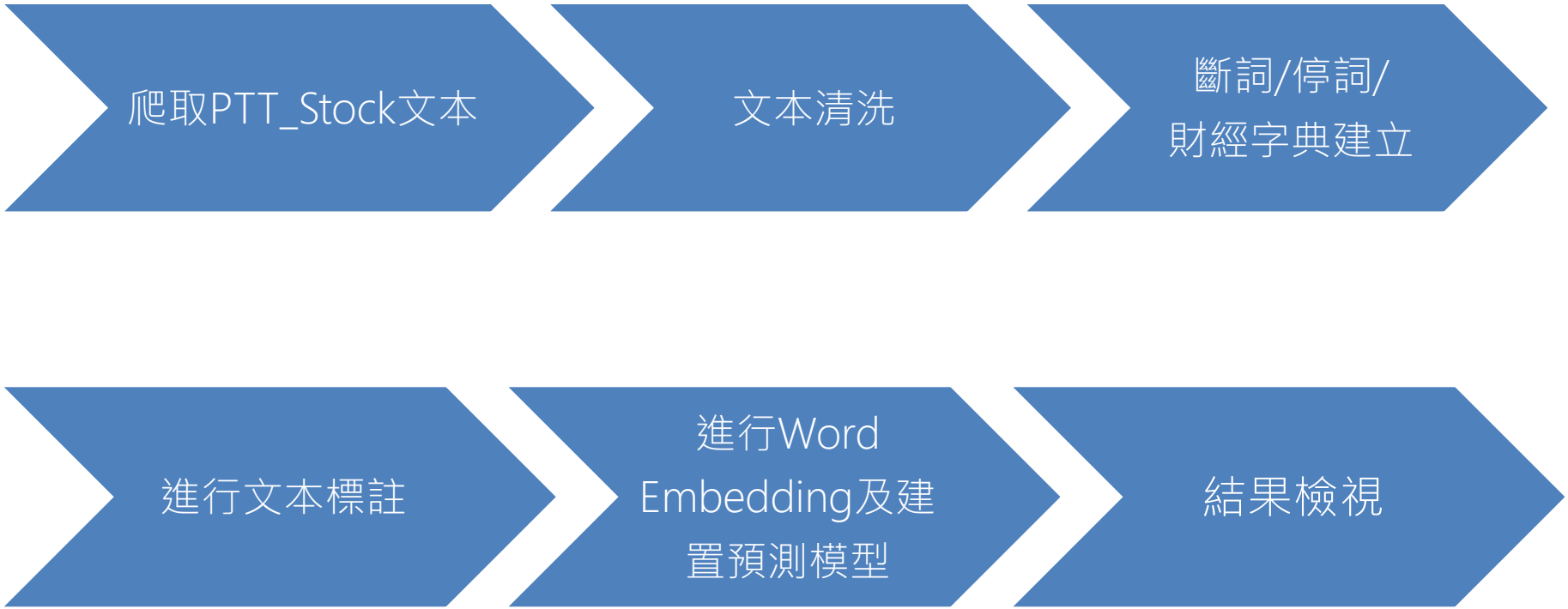
組長：余柏叡(臺大財金碩一)
組員：趙翊茹(臺大財金碩二)
張庭寧(臺大工科碩一)
林庭樂(政大統計碩一)
黃俊穎(臺大電子碩一)

專案說明

- 目標：
  以社群網路之輿情，對投資人行為(基金之申購與贖回量)做預測

- 社群網路：PTT Stock版

專案流程


爬取PTT_Stock文本
文本清洗
斷詞/停詞/
財經字典建立
進行文本標註
進行Word
Embedding及建
置預測模型
結果檢視

流程一：PTT爬蟲

- 說明：
  我們使用python BeautifulSoup套件進行網頁讀取。因各篇文章之網址並無絕對規律，所以首先必須先到看版區抓取各篇文章之連結。連結存取後才進入所有連結爬文，文檔除了文章內容外也紀錄發文時間、發文者、回覆、推噓數、回文者、回覆時間等。把下載的檔案轉成json檔方便建立資料庫。

流程二：文本清洗

- 說明：
  因PTT為社群網路論壇，故常有不帶表投資人情緒之文本如公告、水桶文等，故需將此類文本進行清洗。所幸PTT股版有固定發文格式，僅需寫好特定過濾條件即可完成清洗

流程三：斷詞/停詞/財經字典

- 斷詞：以Jieba為主，雖CKIP之斷詞準確率較高，但因文本數量龐大，速度差距甚遠，故採用Jieba再輔以人工建立財經字典改善斷詞精準度。

- 停詞：以網路公開之停詞，再以人工方式進行增添或刪除

- 財經字典：透過人工方式以隨機抽樣檢視建立

  財經字典新增數目：888詞
  停詞表新增數目：2756詞

流程四：文本標註

- 說明：
  文章內容以人工方式合力進行標註，回覆留言以推噓標註，並加入過去專案已標註之新聞文本標題

- 總標註數：44025
  Pos:15032
  Neg:5207
  Neu:23786

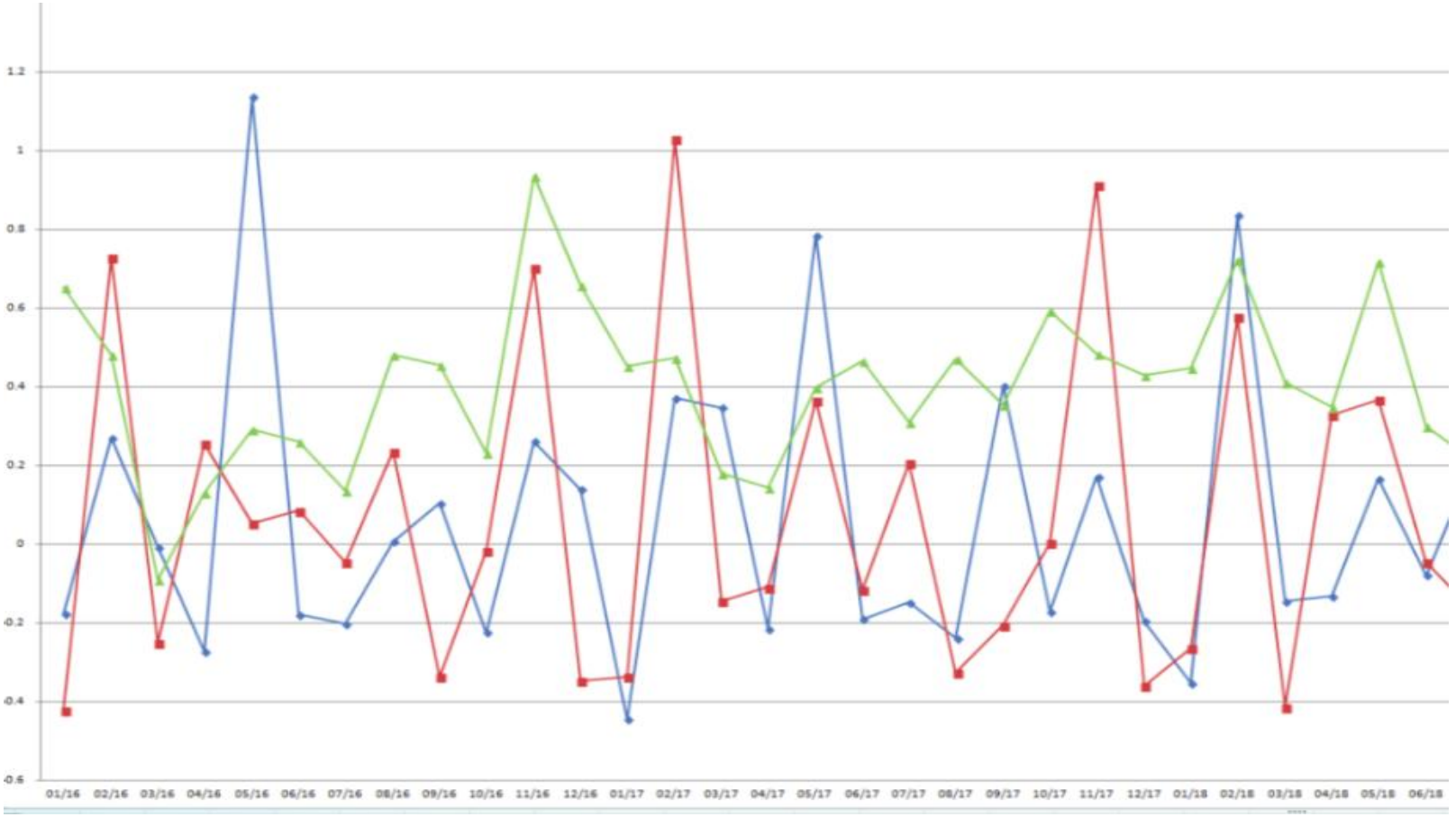
流程五：建置模型

- Word_Embedding : word2vec
- Model :

Architecture	Test accuracy	Train accuracy
LSTM DENSE	0.628(+)	0.84336
CNN LSTM	0.5955(+)	0.8909
CNN MLP	0.5945(+)	0.9997(overfitting)
MLP	0.581(+)	0.9998(overfitting)
LSTM DENSE TIME (previous)	0.4804	0.6092
CNN LSTM time (previous)	0.5698	0.7239
CNN MLP time (previous)	0.4804	0.9336
MLP time (previous)	0.4916	0.9318

流程六：情緒指標建置

- 紅：贖回變動率
- 藍：申購變動率
- 綠：預測情緒指標



改進與未來完成目標

- 運用BERT、XLNet等新的預訓練模型進行指標建置，並與現有模型比較
- 加入其他金融數據作為變數，輔以增強申購贖回預測能力
- 增加詞庫，及增強情緒指標判斷準確度
- 縮小預測週期，觀察預測效果有無改進