
結合情感詞與階層式關注網路辨識股市文章之維度型情感

作者：何彥南、吳政隆

學校：東吳大學巨量資料管理學院

研討會：TCSE 2020 第16屆 台灣軟體工程研討會



目錄

1. 緒論
2. 研究方法
3. 實驗結果
4. 結論與未來展望

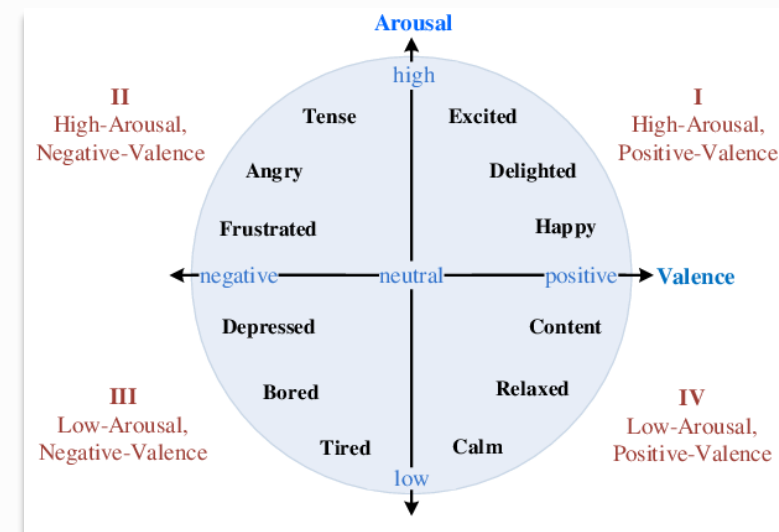
緒論

研究背景與動機 研究目的 文獻探討



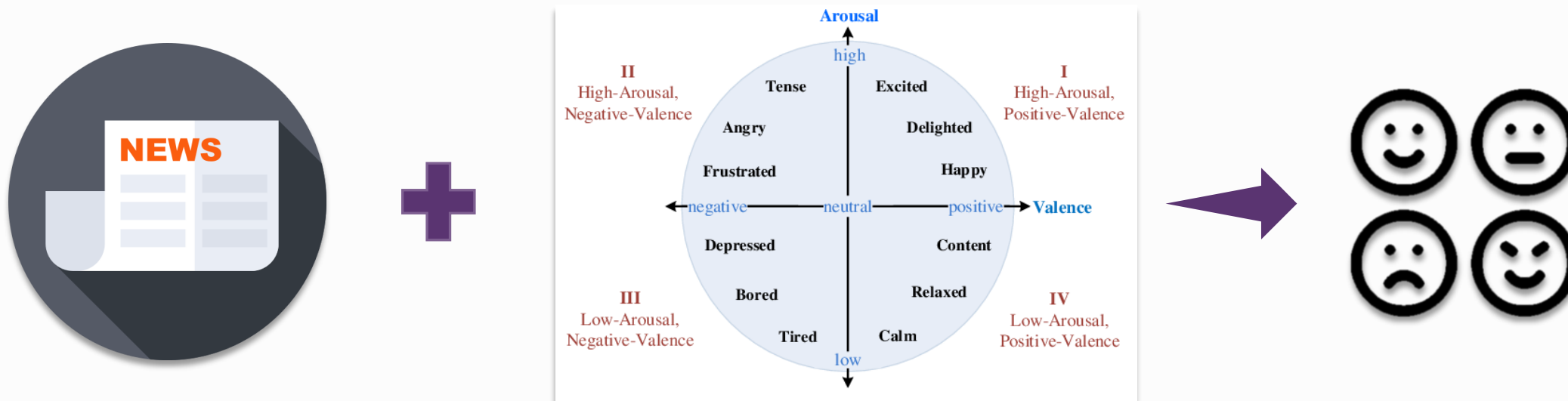
研究背景與動機

- 在股票的投資上
 - 非結構的資料中抽取有用的資訊。
 - 相較於傳統機器學習方法，深度學習導入文本上效果顯著。
- 股市文章的情緒上
 - 以往的研究多是以股市正負面情緒為主。
 - 將 VA 模型的概念導入股市新聞。
 - ✓ 可以進一步知道情緒的強烈程度。
 - ✓ 可以做為股市新聞的情緒的量化指標。



研究目的

- 以 **VA 情感維度** 的架構去標記股市新聞資料
- 透過 **HAN 深度學習模型** 與基於外部情感詞典建立的 **SAN 情感網路模型** 去建立分類模型。
- 此模型可以**辨別股市新聞屬於哪個VA維度**。



文獻探討

股市新聞影響

- ✓ 探討1992 ~ 2017年相關研究，針對社交網路、網路新聞和投資者行為做探討，發現**網路上資訊的擴散**與**股市趨勢**是有很大的相關。
(Shweta Agarwal, 2019)
- ✓ 新聞文章所包含的信息，會改變**投資者的認知**，並影響他們的**投資決策**，影響力會隨文章內容而異。
(Qing Li, 2014)

情感詞典

- ✓ 透過**情感詞典**對股市新聞做**極性分析**，以此預測股價。
(Xiaodong Li, 2014)
- ✓ 結合 VA 的概念建立**中文VA維度型情感詞典(CVAW)**。
(Liang-Chih Yu, 2017)

情緒指標的重要性

- ✓ 使用**綜合新聞情緒指數(ANSI)**，用於投資組合的價值評估，發現新聞中的**情緒指標的反應是很重要的特徵**。
(Yu-Chen Wei, 2019)

文本上的情緒捕捉

- ✓ 提出 **UAM** 模型，主要使用 **topic model**，分析使用者所發的短文，進行**情感分類**。
(Wen Long, 2019)
- ✓ 從社群文章中抽取**情感特徵**和 **LDA 特徵**，使用 **RNN-boost** 模型預測股市指數的變動。
(Weiling Chen, 2018)
- ✓ 提出 **HAN** 模型，可以讓模型學習文章整體的架構。
(Zichao Yang, 2016)

研究方法

文章收集、標記與分類 VA模型的建構 文章分類訓練

A solid purple horizontal bar at the bottom of the slide.

文章的收集

- 網站: 鉅亨網
- 資料: 新聞標題與內文
- 時間: 2019年 1~4月
- 資料筆數: 3588



文章的標記

- 採 5 人獨立標記。
- 各新聞分別標記 Arousal 和 Valence 的數值，數值為1~5。
 - Arousal：為**激動程度**，越大代表越激動。
 - Valence：為**正負面程度**，以 3 為中心，越大則越正面，越小則越負面。
- 最終標記值以 5 人中刪除最大與最小值後剩下的3人做平均。

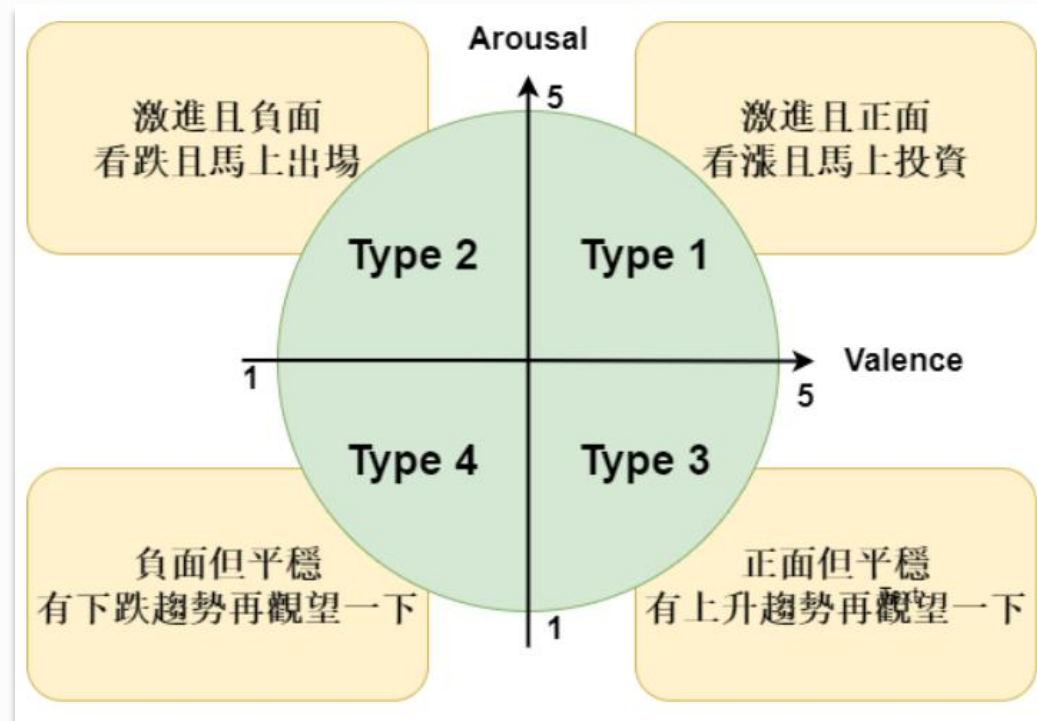
範例

生技產業在經歷前幾年的風霜之後，在今年看起來**漸有起色**，包許多生技大廠接連在臨床實驗與藥證取得下傳出捷報。法人預期，今年下半年開始生技看起來**還算活潑**，明年第 1 季生技產業將如何表現，**值得多加留意**。

Arousal: 2 Valence: 4

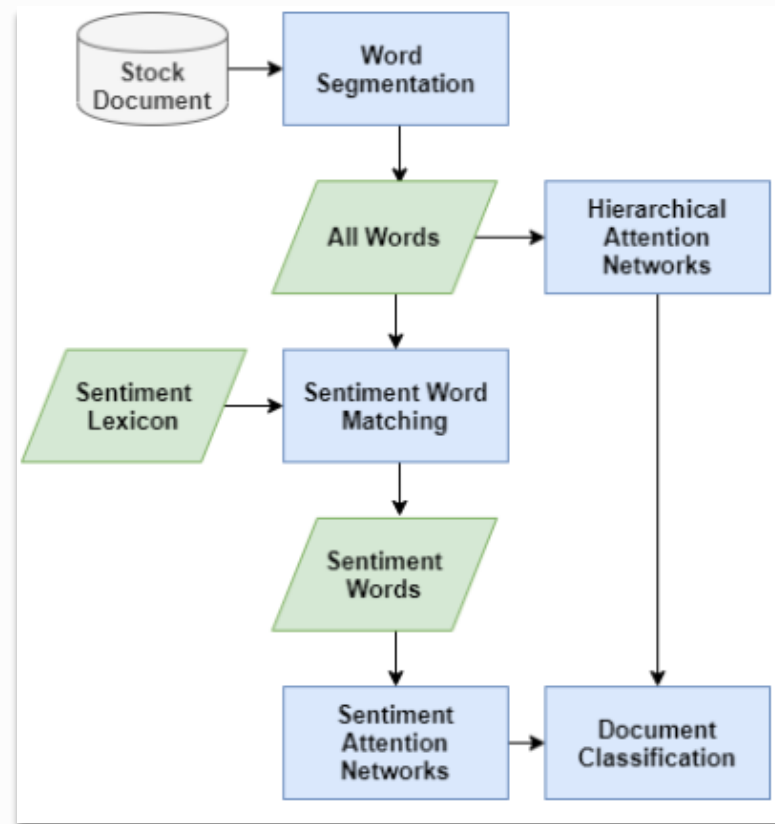
文章的分類

- 在類別的定義上，本研究以 3 為基準分成以下 4 類：
 - Type 1：激進且正面，該新聞表示看漲且會讓人想馬上投資。
(Arousal ≥ 3 · Valence ≥ 3)
 - Type 2：激進且負面，該新聞表示看跌且會讓人想馬上出場。
(Arousal ≥ 3 · Valence < 3)
 - Type 3：正面但平穩，該新聞表示會有上升趨勢但是讓人想先觀望一下。
(Arousal < 3 · Valence ≥ 3)
 - Type 4：負面但平穩，該新聞表示會有下跌趨勢但是讓人想先觀望一下。
(Arousal < 3 · Valence < 3)



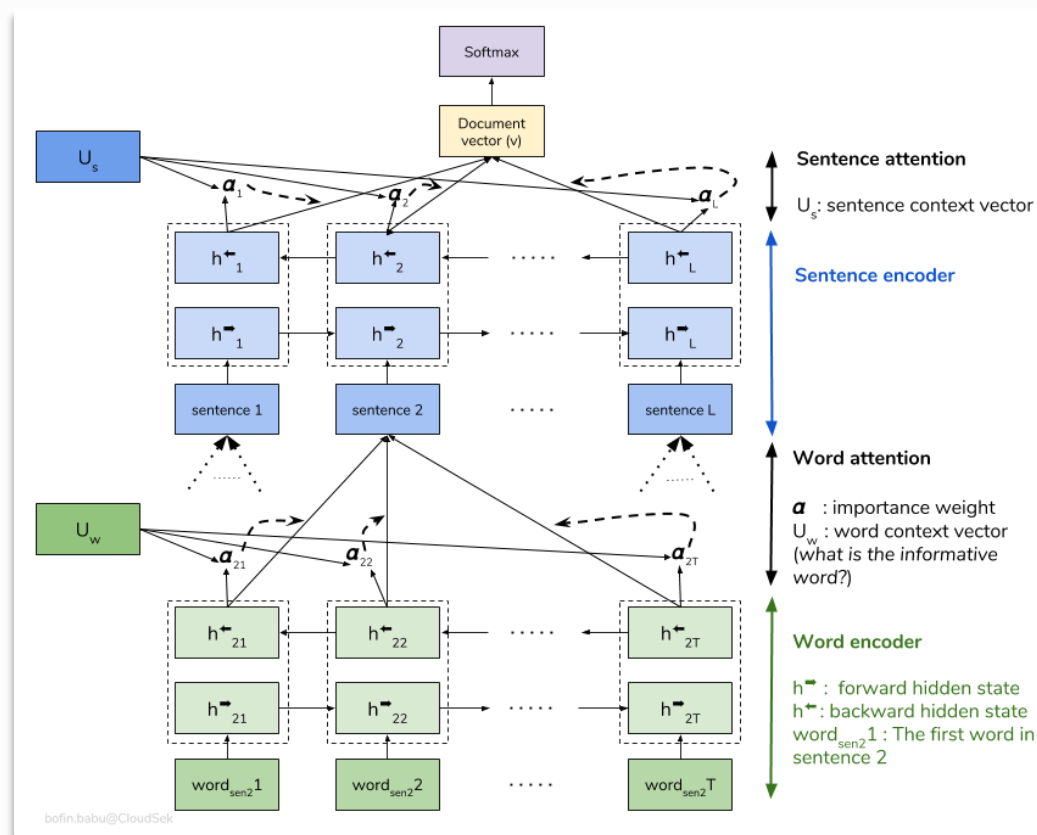
VA 模型的建構 – 主要架構

- VA 模型主要由兩個部分組成:
 - HAN: 輸入是原始文章斷詞和斷句後，學習新聞整體的資訊。
 - SAN: 輸入是透過情感詞典挑出各個文章中的正面詞與負面詞，讓模型專注於學習文章中情感詞的部分。
- 使用 jieba 斷詞
- 以 one-hot encoding 的方式編碼
- 外部情感詞典使用台灣大學(NTUSD)和元智大學(CVAW)



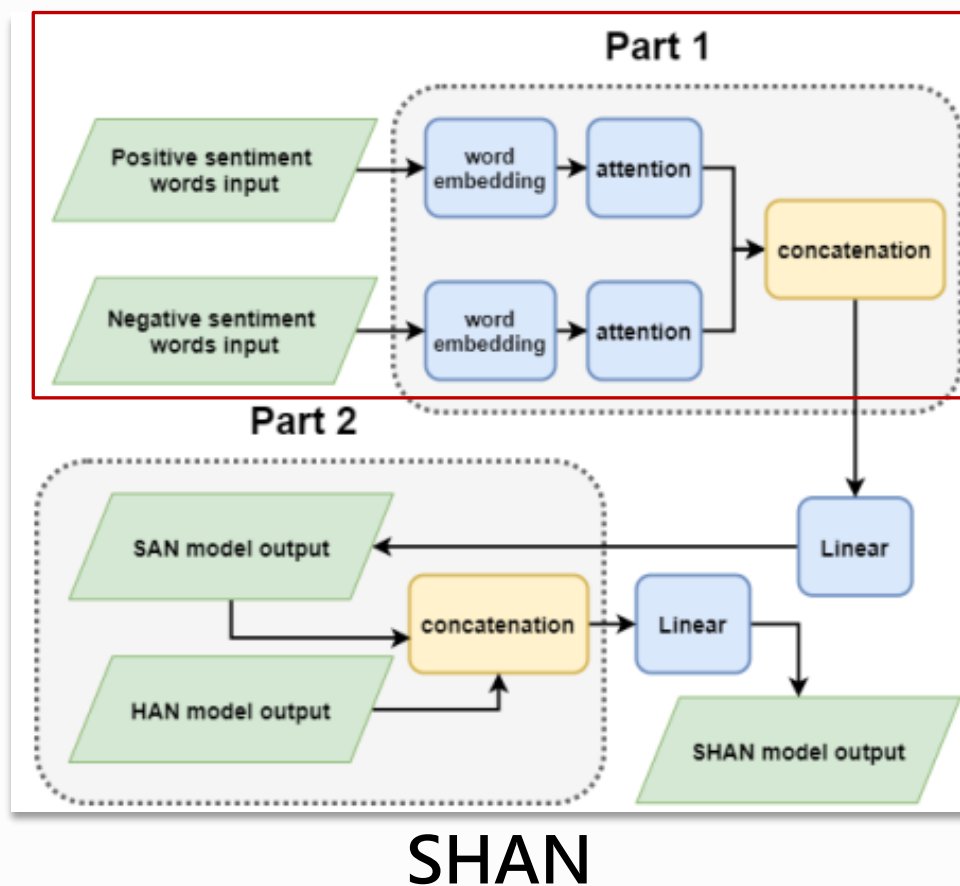
VA 模型的建構 – HAN

- HAN 模型由句子和單詞兩個層次組成，透過對這兩個層次去模擬人在讀文章時接收訊息的一個過程。
- 主要對單詞和句子的處理，機制是一樣的，將編碼過的序列，透過 RNN 遞歸神經網路學習，最後在加入一層 attention 層次讓模型可以去調整要比較關注在哪個地方。

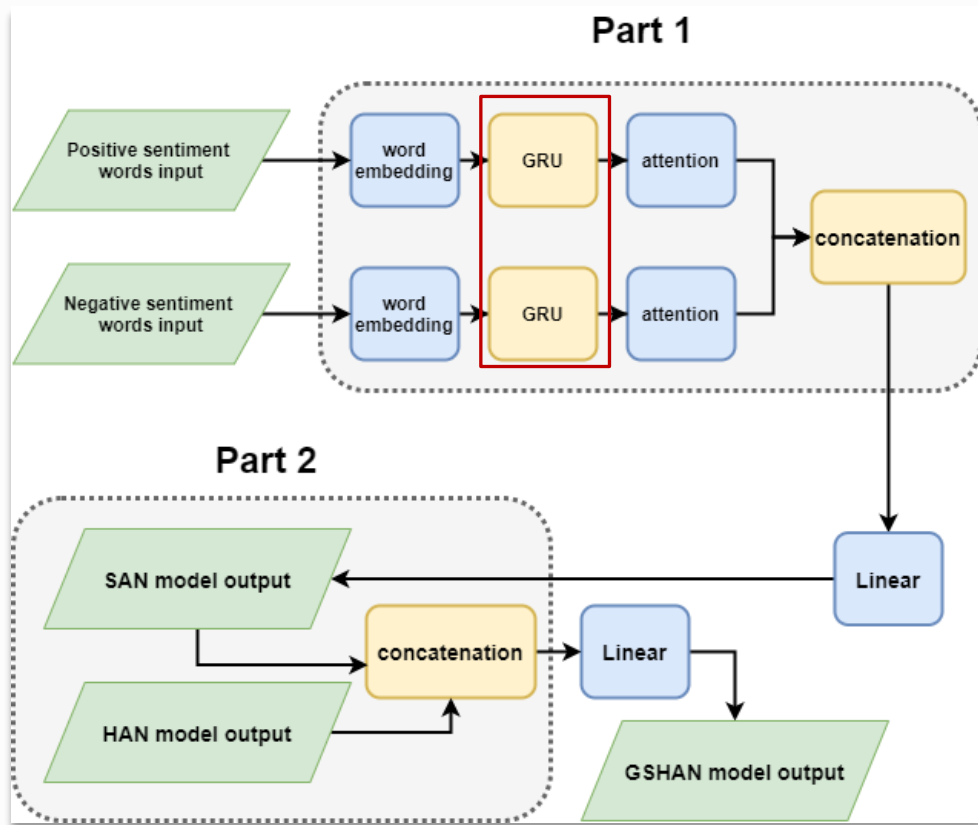


VA 模型的建構 – SAN

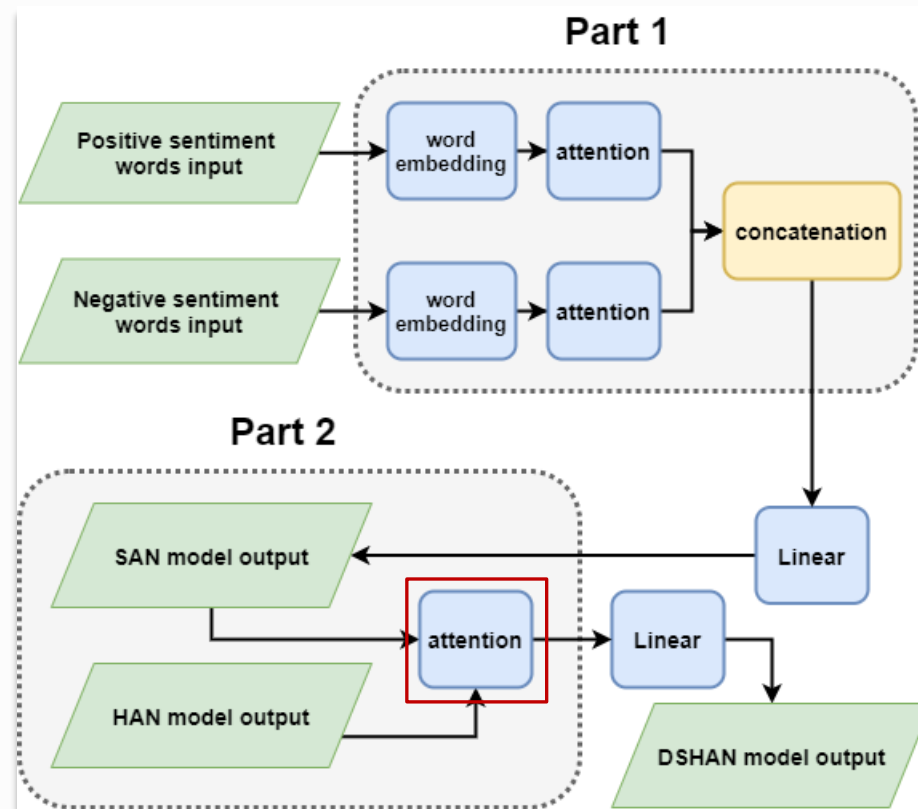
- 本研究提出 SHAN 模型是基於HAN模型上，在額外加入 SAN 情感網路模型，此模型分為兩個 Part:
 - Part 1 : SAN 模型主要的輸入為該文章中所有的正面和負面情感詞，經過編碼後，各自進入關注層，在將兩個序列直接併起來，經過一層線性層，就是 SAN 模型的結果。
 - Part 2 : SAN 模型的輸出與原先 HAN 模型的輸出併起來，經過一層線性層為 SHAN 最終的輸出結果。



VA 模型的建構 – GSHAN & DSHAN



GSHAN



DSHAN

文章分類訓練

- 最後每篇文章會得到一個 1×4 向量 Y ，本模型的預測目標為多類別分類。為了預測四種分類，本研究採用 **Cross Entropy** 計算訓練誤差，其誤差公式如下：

$$loss(Y, class) = -\log\left(\frac{\exp(Y[class])}{\sum_j \exp(Y[j])}\right)$$

- Y 表示第 d 篇的股市文章的類別標籤。而在收斂部分，根據誤差 $loss$ 反饋傳遞計算梯度(Gradient)，是一種**梯度下降(Gradient Descent)**的收斂方式，並以隨機最佳化的 **Adam** 最佳化器進行網路參數收斂。
- 以 **Pytorch** 作為主要架構模型的工具。

實驗結果

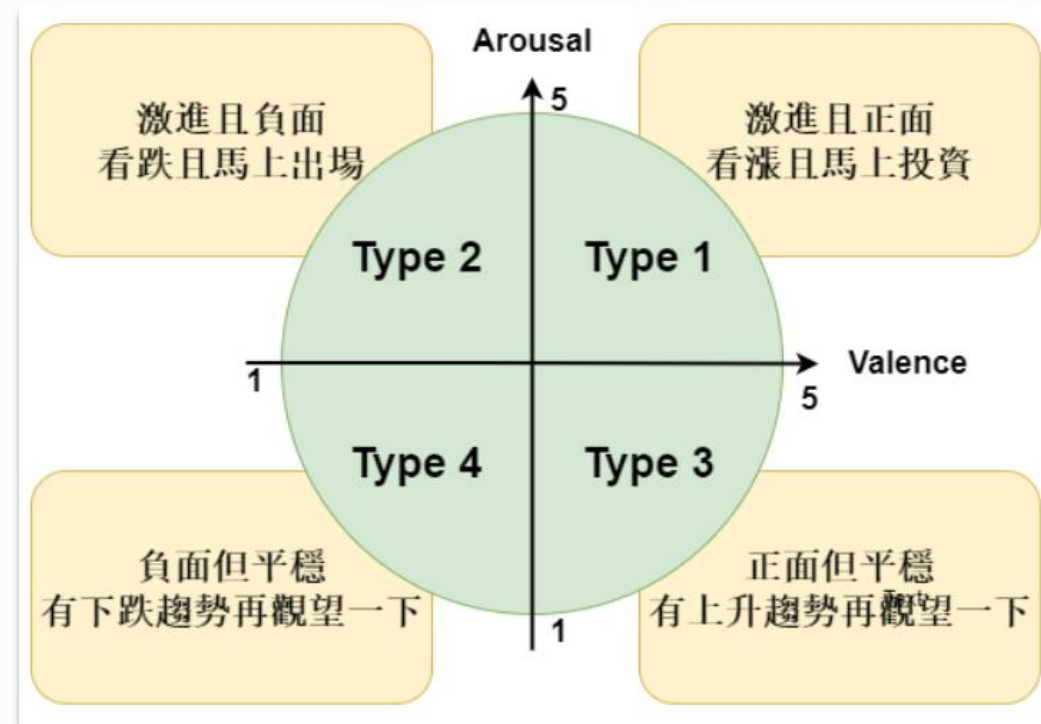
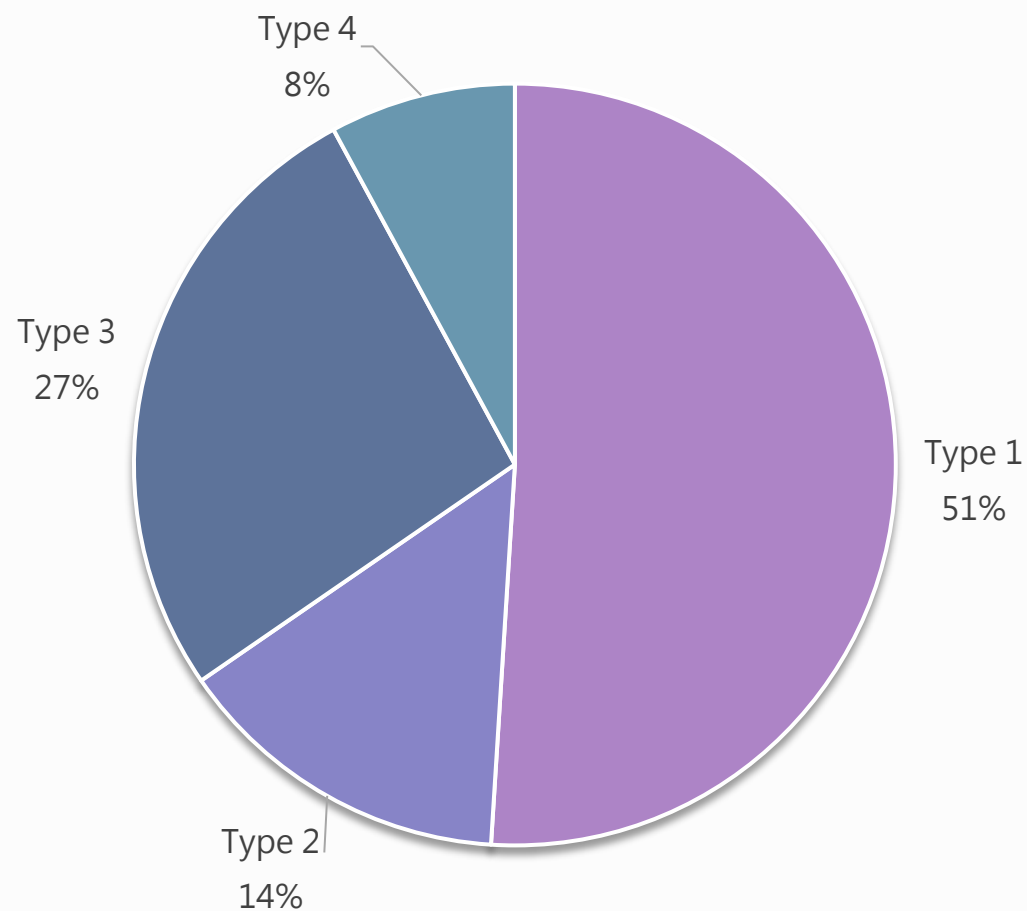
資料集 參數設計 分類模型說明 模型結果比較 模型效能分析



資料集 - 文章

	訓練	驗證	測試
資料筆數	2,628	498	462
平均句子數	5.28	5.20	5.20
平均詞數(每句)	7.19	7.22	7.25
平均詞數(每篇文章)	37.94	37.54	37.74
平均正面詞數	1.60	1.66	1.48
平均負面詞數	0.73	0.78	0.70

資料集 - 預測類別



參數設計

參數名稱	參數設定	說明
RNN type	{GRU, LSTM}	RNN 的類型
Learning rate	{0.0001, 0.001, 0.01}	學習率
Hidden size	{50, 100, 200}	隱藏層
Batch size	64	批次大小
Epoch	100	迭代次數
Dropout rate	0.0	隨機丟失比例

主要

分類模型說明

模型名稱	組成模型	說明
HAN	HAN	原始的 HAN 模型
SHAN	SAN + HAN	HAN 與 SAN 組合模型
GSHAN	SAN + HAN	SAN 的部分加入 GRU
DSHAN	SAN + HAN	SAN 和 HAN 之間額外使用 attention
SAN	SAN	SHAN 將 HAN 的部分移除
GSAN	SAN	GSHAN 將 HAN 的部分移除
DSAN	SAN	DSHAN 將 HAN 的部分移除

← 基準

主要

模型結果比較 – RNN type

Model	RNN type	Val loss	Val F1 score
HAN	GRU	0.85	0.66
	LSTM	0.94	0.60
SHAN	GRU	0.80	0.66
	LSTM	0.88	0.65
DSHAN	GRU	0.96	0.64
	LSTM	0.97	0.60
GSHAN	GRU	0.82	0.67
	LSTM	0.91	0.65



模型結果比較 – Learning rate

Model	Learning rate	Val loss	Val F1 score
HAN	0.0001	0.96	0.58
	0.001	0.85	0.66
	0.01	1.04	0.62
SHAN	0.0001	0.93	0.60
	0.001	0.80	0.66
	0.01	1.07	0.63
DSHAN	0.0001	1.03	0.58
	0.001	0.96	0.64
	0.01	0.97	0.60
GSHAN	0.0001	0.83	0.65
	0.001	0.82	0.67
	0.01	1.07	0.61



模型結果比較 – Hidden size

Model	Hidden size	Val loss	Val F1 score
HAN	50	0.83	0.65
	100	0.85	0.66
	200	0.94	0.61
SHAN	50	0.80	0.66
	100	0.86	0.64
	200	0.92	0.63
DSHAN	50	0.96	0.64
	100	0.99	0.58
	200	0.98	0.58
GSHAN	50	0.82	0.67
	100	0.91	0.65
	200	0.83	0.65




模型結果比較 – 不同情感詞典

Model	Sentiment dictionary	Val loss	Val F1 score
SHAN	Both	0.81	0.64
	CVAW	0.80	0.65
	NTUSD	0.80	0.66
DSHAN	Both	0.96	0.64
	CVAW	1.00	0.59
	NTUSD	0.96	0.63
GSHAN	Both	0.79	0.66
	CVAW	0.79	0.66
	NTUSD	0.82	0.67



模型結果比較 – 未使用 HAN 比較

HAN	Model	Val loss	Val F1 score
✓	SHAN	0.80	0.66
	SAN	1.00	0.59
✓	GSHAN	0.82	0.67 
	GSAN	0.98	0.60

測試集模型效能分析

Model	RNN type	Hidden size	Learning rate	Sentiment dictionary	Test F1 score	Rank
SHAN	GRU	50	0.001	NTUSD	0.70	1
GSHAN	GRU	50	0.001	NTUSD	0.66	2
HAN	GRU	100	0.001	None	0.65	3
DSHAN	GRU	50	0.001	Both	0.63	4
SAN	LSTM	100	0.0001	Both	0.59	5
GSAN	GRU	200	0.0001	Both	0.58	6
DSAN	LSTM	200	0.01	Both	0.58	7



結論與未來展望

結論 研究限制與未來展望



結論

- 本研究將 VA 模型導入到股市新聞的情緒分類上，並透過加入外部的情感詞典所建立的 SAN 模型提高模型表現。最終結果可知，本研究提出的三個加入 SAN 網路的模型，其中有兩個表現比HAN 好。但是當我們完全不使用HAN模型時，效果就大幅下降。總結來說，本研究提出的SAN模型可以輔助HAN模型的分類，但是獨立做分類任務時效果有限。
- 貢獻
 - 將 VA 模型的概念導入到股市新聞，透過深度學習可以有效捕捉新聞中的正負面情緒與情緒的強烈程度。可以作為一個股市新聞的情緒量化指標。
 - 加入情感詞典所建立的 SAN 可以有效的加強模型在情緒分類的任務上表現。

研究限制與未來展望

- 在情感的辭典上，可以在加強在股市新聞的用詞。
- 在詞向量的部分，可以嘗試使用 word2vec、Bert 方式做比較。
- 可以測試一下不同的分類模型與 SAN 結合的效果如何。
- 在加上 SAN 模型的訓練時間成本還需要在探討。

THE END

Thanks for listening