Hey Vergil at ROCLING-2025 Shared Task: Emotion-Space-Based System for Doctors' Self-Reflection Sentiment Analysis

Ting-Yi Lin *Cong-Ying Lin* *Jui-Feng Yeh*

Department of Computer Science and Information Engineering
National Chiayi University
Chiayi, Taiwan

thomas60611@gmail.com transformers0514@gmail.com ralph@mail.ncyu.edu.tw

摘要

本研究針對 **ROCLING 2025** 維度情感分析任務，提出 **EmoTracer**，一個基於情緒空間的醫師日誌情感分析系統。系統採用 XLNet、BERT 與 LSTM 模型，並以 **SLAKE 病症資料集**及中文資料集（如 **Chinese EmoBank**、**NRC-VAD**）訓練，以捕捉醫師在撰寫病症相關日誌時可能產生的情緒波動。EmoTracer 可將文本轉換為 **Valence** 與 **Arousal** 分數，實驗結果顯示準確率約 60%，皮爾森相關係數（PCC）達 0.9，均方誤差（MAE）約 0.3，可作為心理健康管理的參考工具。系統同時建立了簡易的 **前端 UI**，方便使用者輸入文本並查看分析結果，以完整呈現 **EmoTracer 系統**功能。

Abstract

In the ROCLING 2025 dimensional sentiment analysis task, we present EmoTracer. It is an emotion-space-based system for analyzing doctors' self-reflection texts. The system uses XLNet, BERT, and LSTM models. It is trained on the SLAKE medical dataset and Chinese datasets, such as Chinese EmoBank and NRC-VAD. This helps the system capture the possible emotional changes of doctors when they write patient-related reflections. EmoTracer converts texts into Valence and Arousal scores. The experiments show about 60% accuracy, a Pearson correlation coefficient (PCC) of 0.9, and a mean absolute error (MAE) of 0.3. These results can help support mental health management. The system also has a simple front-end UI. Users can enter texts and see the analysis results. This demonstrates the full functionality of the EmoTracer system.

關鍵字：情緒空間座標、文本情感分析、醫師自我反思文本

Keywords: Emotion Space Coordinates, Text Sentiment Analysis, Doctors' Self-Reflection Texts

1 介紹

近年來，醫師在臨床工作中面臨高度壓力，心理健康問題日益受到關注。過勞、患者照護壓力以及醫療決策責任都可能導致醫師產生焦慮、情緒波動，甚至心理危機。研究顯示，醫師是自殺的高風險族群，其標準化死亡比（SMR）為 1.44，女性醫師的自殺風險更高（SMR = 1.9），顯示這個族群的心理健康問題不容忽視。由於醫師往往忙於工作，缺乏充分的心理支持，早期察覺心理困擾變得困難。

本論文提出 **EmoTracer**，一個基於情緒空間的醫師自我反思日誌分析系統。醫師可透過**表達性書寫（Expressive Writing）**記錄日常臨床經驗與情緒感受，系統則利用 **自然語言處理（NLP）** 將日誌文本轉換為 **Valence** 與 **Arousal** 的二維情緒座標。EmoTracer 可幫助醫師自我識別情緒波動、調整心理狀態，並作為專業心理輔導的輔助工具。

透過對自我反思文本的持續追蹤與分析，醫師的情緒歷程可視化呈現，不僅增進個人自我認知，也能協助醫療機構掌握團隊壓力情況，為心理健康干預提供有效數據支援。

2 模型架構

本論文提出的模型採用了聚合模型 (Aggregation Model) 的方式，融合了 XLNet、

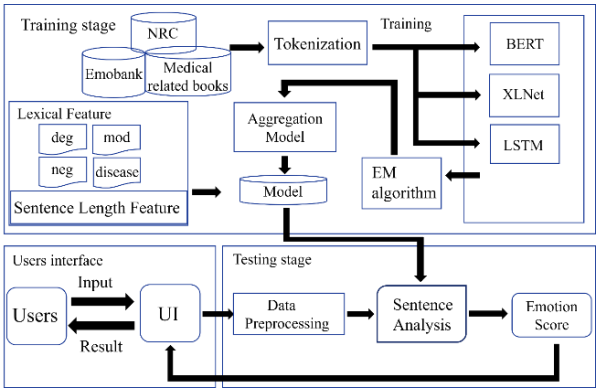
Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025), pages 407–412

BERT 以及 LSTM 三種模型，並使用 最大期望演算法 (EM Algorithm) 來動態分配各模型的權重，以彌補彼此的優缺點，並提高整體預測的準確性 。

在詞彙特徵標註方面，我們根據 Chinese EmoBank 和 SLAKE 醫療資料集，標註了病症詞、程度詞、語氣詞及否定詞等詞彙特徵，以擴充詞庫並提升準確度 。同時也針對雙重否定進行數值調整，並利用語句長短對情緒空間進行線性處理，以更精確地捕捉情緒的強度 。

我們使用獨立的測試集來驗證模型效能，並採用平均絕對誤差 (MAE)、皮爾遜相關係數 (PCC) 及正確率 (ACC) 作為評估指標 。最終， EmoTracer 日誌情感分析系統在正負性 (Valence) 預測上的正確率接近 60%，PCC 值超過 0.9，且 MAE 僅有 0.3 左右，顯示出優異的效能 。

最後結合 Python 視窗使用者介面，讓使用者能夠輸入文字或檔案進行分析，並透過二維情緒空間圖與長期情緒變化追蹤圖表，將量化的情緒指標視覺化，有助於識別學生在壓力下的情緒波動，並為預防心理危機提供依據 。



3 實驗方法

為了驗證系統效能，我們設計了兩個實驗，分別為模型準確率測試及語句特徵比較。

在模型準確率測試中，本論文使用準確率 (ACC)、皮爾遜相關係數 (PCC) 和均方誤差 (MAE) 三種指標來評估模型效果 。我們測試了多種模型，包括使用單一資料集 (Chinese EmoBank 或 NRC-VAD) 訓練的 XLNet 、LSTM 和 BERT 模型 。其中，未加入詞彙特徵的 XLNet 模型被設為基準線 (Baseline) 。最終，我們將所有模型合併成一個聚合模型 (Aggregation Model)，並測試其效能 。

在語句特徵比較實驗中，我們針對不同長度的語句 (詞數分別為 10、18、20、30) 進行了線性調整，並評估其對聚合模型效能的影響 。

微調方式

本論文的聚合模型根據最大期望演算法 (EM Algorithm) 對模型進行微調。由於同時使用 NRC-VAD 和 Chinese EmoBank 兩種資料集進行訓練，我們將其擴展為六個獨立的子模型 。最大期望演算法能夠動態地為這六個模型分配權重，其目的是整合各模型的優點，使其強弱項互相補足，並透過最小化整體預測誤差的目標函數來持續更新權重，最終目標在於提高整體分析的準確性與穩定性 。

為了驗證此優化策略的成效，並全面性地評估模型在情感強度預測這個迴歸問題 (regression problem) 上的性能表現，我們採用了以下三種互補的評估指標：

平均絕對誤差 (Mean Absolute Error, MAE) 是評估迴歸任務的核心指標之一，用於衡量模型預測值與實際標註值之間的平均誤差 。在本研究中，MAE 不僅是評估模型最終性能的關鍵，其概念也與我們模型微調過程中最小化的目標函數緊密相關。MAE 值越小，代表模型的預測在數值上越精準。

皮爾遜相關係數 (Pearson Correlation Coefficient, PCC) 則從另一個維度衡量模型的表現。相較於 MAE 關注預測的誤差大小，PCC 則是用於衡量模型預測結果與實際資料之間的線性相關性 。當 PCC 值越接近 1，表示模型的預測趨勢與真實值的變化趨勢高度一致，證明模型能準確捕捉情感分數的相對高低變化 。

最後，我們引入正確率 (Accuracy, ACC) 作為輔助指標，用以評估模型在判斷情感基本傾向上的表現。在計算上，我們以 Chinese EmoBank 資料集中的 Chinese valence-arousal sentences (CVAS) 子資料集所提供的標註分數作為評估的參考基準。由於情緒標註並不存在絕對的標準答案，因此 ACC 在此處的作用是量化模型預測結果與此參考基準的相符程度。提供了一個具實用價值的互補視角，用以驗證模型在捕捉情感方向的可靠性，從而與 MAE 和 PCC 共同構成了對模型更完整的性能驗證。

資料集

在訓練資料方面，我們採用了三個資料集(如表 1)，其中 Chinese EmoBank： 包含 11,043 筆數據，涵蓋單字、短語、單句和多句文本，並標註了情感的正負性與喚醒值 。而 NRC-VAD： 包含 119,791 筆中文資料，同樣採用情緒空間理論進行標註 。還有醫學生日誌相關的其他書籍，用於擴充領域詞彙 。

詞彙特徵標註的部分，我們根據 Chinese EmoBank 和 SLAKE 醫療資料集，標註了病症詞、程度詞、語氣詞和否定詞，以增強模型對醫療語境的識別能力 。

我們的測試資料集採用訓練資料集 Chinese EmoBank 中的 Chinese valence-arousal sentences (CVAS)，以該資料集的短句標註分數做為參考標準 。DSA-MST 比賽測試資料集： 一個獨立於訓練資料的測試集，用於評估模型的泛化能力 。

詞彙特徵處理

在情感分析系統中，詞彙特徵處理扮演著至關重要的角色。它不僅讓模型能夠識別語言中的情感，更能精確捕捉情感的強度、語氣與方向。我們將這個過程分為以下幾點：

1. 病症詞標註(2216 筆)：

本論文參考了醫療領域的 SLAKE 資料集，對日誌中常見的疾病詞彙（如「感冒」、「頭痛」、「失眠」等）進行了情感值標註。這樣做的目的是為了讓模型能夠理解，這些詞彙本身就帶有負面的情感和較高的喚醒值，從而更精準地識別醫學日誌中因身體狀況而產生的情緒。

2. 程度詞與語氣詞(共 65 筆)：

程度詞（如「非常」、「略微」）與語氣詞（如「也許」、「一定」）能夠為情感提供細緻的層次感。我們根據 Chinese EmoBank 的標準差和模型測試結果，為程度詞分配了不同的調整倍率，讓模型能夠區分「有點不開心」和「超級不開心」之間的強度差異。同樣地，語氣詞和標點符號的結合（如「?」和「!」）則能幫助模型識別情感中的不確定性、推測或強烈的情緒表達。

3. 否定詞與雙重否定(13 筆)：

否定詞（如「不」、「沒」）不僅能反轉情感的正負性，雙重否定（如「不得不」）更能強化語氣。我們針對否定詞進行了數值調整，並特別處理了雙重否定的情況，以確保模型能夠正確理解語義的轉變，例如將一個正向情感的表達轉為強烈的負面情緒。

透過這些細緻的詞彙特徵處理，本論文的聚合模型能夠超越傳統的情感分類，深入理解複雜的情緒變化，讓日誌情感分析系統的結果更加細膩且準確。

語句特徵比較

根據語言簡潔性對情緒空間的影響，中文語句在表達情感時能夠迅速、強烈地傳達情感，尤其是在簡短的語句中，情緒通常會顯得更加急迫、緊張或命令式。例如，簡單的祈使句如「去做功課」、「快走！」或「別說話」都能立即將情感強烈地傳達出來，且這些語句通常會省略主語，使得情緒表達更加直接。

在語句特徵比較的演算法中，我們通過對不同長度的語句進行分析，對語句的詞數進行篩選，並根據語句的詞數分別為 10、18、20、30 來計算斜率及截距。這樣的分析有助於更精確地捕捉情緒的強度，尤其是短語句所表達的強烈情感，並能在情緒空間中準確地映射出其對應的情感強度。

表 1：資料集數量

資料集種類	資料使用數量
NRC-VAD	119791
Chinese EmoBank	11043
其他書籍	92973
程度詞類	223
疾病詞 SLAKE	2216

4 實驗結果

綜合表 2 的所有實驗數據與分析，我們發現本論文提出的聚合模型(Aggregation Model)在情感分析的表現上取得了顯著的成功。相較於單一模型，本論文的聚合模型在所有測試中展現了更為優異且穩定的效能。

在單一模型中，使用 EmoBank 資料集訓練的 BERT 模型在正負性（Valence）的準確性和相關性上表現突出，正確率高達 58.32%，皮爾遜相關係數(PCC)達到 0.8701。然而，

BERT 在喚醒度（Arousal）的預測上誤差較大，而 XLNet 與 LSTM 則在這方面有更好的表現。

透過 最大期望演算法 (EM Algorithm) 的微調與權重分配，我們的聚合模型有效地融合了各個單一模型的優點，並顯著彌補了它們的不足。實驗結果顯示，聚合模型在正負性（Valence）與喚醒度（Arousal）的 正確率 (ACC)都超過 50%，分別達到了 59.02% 和 50.97%。

表 2 對 Aggregation Model 測試比較

模型	資料集	Valence Accuracy	Arousal Accuracy	PCC Valence	PCC Arousal	MAE Valence	MAE Arousal
Baseline XLNet	EmoBank	41.51%	47.06%	0.8112	0.2215	0.6541	1.0772
XLNet	EmoBank	54.33%	47.28%	0.7809	0.5508	0.5461	0.4295
XLNet	NRC	36.37%	48.14%	0.7899	0.0775	0.5962	0.4134
LSTM	EmoBank	55.65%	47.97%	0.5796	0.4885	0.6753	0.5593
LSTM	NRC	31.56%	48.87%	0.8185	0.7868	0.5868	0.4366
BERT	EmoBank	58.32%	50.56%	0.8701	0.5587	0.5172	0.8432
BERT	NRC	55.44%	44.57%	0.8112	0.1749	0.6283	1.1280
Aggregation Model	ALL	59.02%	50.97%	0.9120	0.6283	0.3146	0.3061
比賽結果	ALL			1.01	0.21	0.63	0.62

另外，該模型在正負性(Valence)的皮爾遜相關係數(PCC)為 0.9120，已接近上限值，同時其均方誤差（MAE）僅為 0.3146，是所有實驗組別中最低的。證明了聚合模型在預測的精準度上達到了最佳表現，能夠為日誌情感分析提供一個可靠且高效的解決方案。

表 3 的語句特徵比較實驗結果顯示，在語句長度接近 18 個詞時，聚合模型在正負性（Valence）與喚醒值（Arousal）上的 正確率（ACC）都有所提升。雖然這項處理使得皮爾遜相關係數（PCC）與均方誤差（MAE）略有下降，但 MAE 仍維持在 0.3 左右，顯示其誤差並未明顯增加。這證實本論文的語句長短特徵處理，確實能對整體模型的效能有所助益。

表 3 語句特徵比較之實驗結果

語句詞數	Valence Accuracy	Arousal Accuracy	PCC Valence	PCC Arousal	MAE Valence	MAE Arousal
未針對語句特徵做處理	59.02%	50.97%	0.9120	0.6283	0.3146	0.3061
10	59.02%	51.03%	0.9096	0.5538	0.3127	0.3911
18	59.13%	51.34%	0.9318	0.5679	0.3130	0.3957
20	58.81%	50.17%	0.9111	0.5860	0.3131	0.4026
30	58.03%	48.46%	0.9193	0.5066	0.3066	0.4488

這個結果與語言的簡潔性理論相符，亦即簡短的語句通常會更直接、更強烈地表達情感。透過分析與調整，我們的系統能更精準地捕捉日誌中因語句長短所帶來的細微情

緒變化，特別是短句所蘊含的強烈情感，進而提升情感分析的準確度。

5 結論

本論文透過自然語言處理技術，開發了一個能夠精準分析情緒日誌並長期追蹤情感變化的系統。我們深入研究了中文的語言特性，特別是針對否定詞、程度詞等詞彙特徵進行處理，有效地提升了模型的可靠性。

實驗結果顯示，本論文所提出的模型能夠將醫生的自我反思文本轉換為易於理解的情緒空間。在正負性（Valence）的預測上，正確率接近 60%，皮爾遜相關係數（PCC）超過 0.9，且均方誤差（MAE）維持在 0.3 左右，證明了系統在情感分析上的優越性能。

本系統不僅為醫生提供了一個能夠記錄與分析日誌內容的工具，也能透過持續追蹤的功能為其心理健康管理提供參考，幫助他們及時了解自身的情感波動，識別出潛在的心理疾病風險，提供早期預警的支持工具。

6 References

Eva S. Schernhammer and Graham A. Colditz. 2004. Suicide rates among physicians: a quantitative and gender-specific review of the literature. American Journal of Psychiatry, 161(12):2295–2302.

J. M. Smyth. 1998. Written emotional expression: effect sizes, outcome types, and moderating variables. Journal of consulting and clinical psychology, 66(1):174.

J. W. Pennebaker and C. K. Chung. 2011. Expressive Writing: Connections to Physical and Mental Health.

Lung-Hao Lee, Tzu-Mi Lin, Hsiu-Min Shih, Kuo-Kai Shyu, Anna S. Hsu, and Peih-Ying Lu. 2025. ROCLING-2025 Shared Task: Chinese Dimensional Sentiment Analysis for Medical Self-Reflection Texts. In Proceedings of the 37th Conference on Computational Linguistics and Speech Processing.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan R. Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In Advances in Neural Information Processing Systems, volume 32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

J. H. Wang, T. W. Liu, X. Luo, and L. Wang. 2018. An LSTM approach to short text sentiment classification with word embeddings. In Proceedings of the 30th Conference on Computational Linguistics and Speech Processing (ROCLING 2018), pages 214–223.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society: Series B (Methodological), 39(1):1–22.

Lung-Hao Lee, Jian-Hong Li and Liang-Chih Yu. 2022. Chinese EmoBank: Building Valence-Arousal Resources for Dimensional Sentiment Analysis. ACM Transactions on Asian and Low-Resource Language Information Processing, 21(4): article 65.

Huan-Ling Lin, Yu-Sheng Lu, Jheng-Wei Chen, et al. 2021. SLAKE: A Semantically-Labeled Knowledge-Enhanced Dataset for Medical VQA. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 9506–9513.

Pratibha Chauhan and Nitin Sharma. 2023. A systematic review on dimensional sentiment analysis. Multimedia Tools and Applications, 82(12):18011–18043.

Jing Zhao, Siyu Kang, Peijie Liu, Gerard de Melo, and Yaling Zhang. 2023. VADER-based iterative deep multi-task learning for valence-arousal prediction. In Proceedings of the AAAI Conference on Artificial Intelligence, 37(11):13264–13272.

Zhen Li, Jian-Hua Li, and Shi-Feng Wang. 2024. Context-Aware and Speaker-Sensitive Network for Dimensional Emotion Recognition in Conversations. IEEE Transactions on Affective Computing.

F. Miedema and S. Bhulai. 2018. Sentiment analysis with long short-term memory networks. Vrije Universiteit Amsterdam, pages 1-17.

B. Huang, Y. Ou, and K. M. Carley. 2018. Aspect level sentiment classification with attention-over-attention neural networks. In Proceedings

of the 11th International Conference on Social, Cultural, and Behavioral Modeling (SBP-BRiMS 2018), pages 197–206.

W. Wang, B. Bi, M. Yan, C. Wu, Z. Bao, J. Xia, and L. Si. 2019. Structbert: Incorporating language structures into pre-training for deep language understanding. arXiv preprint arXiv:1908.04577.

M. V. Koroteev. 2021. BERT: a review of applications in natural language processing and understanding. arXiv preprint arXiv:2103.11943.

A. F. Adoma, N. M. Henry, and W. Chen. 2020. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), pages 117–121.

H. F. Zhang, C. Zeng, and P. He. 2022. An Emotion Cause Detection Method Based on XLNet and Contrastive Learning. In Proceedings of the International Conference on Software Engineering and Knowledge Engineering (SEKE), pages 646–649.

N. Habbat, H. Anoun, and L. Hassouni. 2022. Combination of GRU and CNN deep learning models for sentiment analysis on French customer reviews using XLNet model. IEEE Engineering Management Review, 51(1):41–51.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, et al. 2016. Building Chinese Affective Resources in Valence-Arousal Dimensions. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics, pages 540–545.