Computational Linguistics and Chinese Language Processing
Vol. 26, No. 2, December 2021, pp. 1-16
© The Association for Computational Linguistics and Chinese Language Processing

以遷移學習改善深度神經網路模型於中文歌詞情緒辨識

Using Transfer Learning to Improve Deep Neural Networks for Lyrics Emotion Recognition in Chinese

廖家誼*、林亞宣、林冠成、張家瑋+

Jia-Yi Liao, Ya-Hsuan Lin, Kuan-Cheng Lin, and Jia-Wei Chang

摘要

情緒是音樂資訊檢索中的重要屬性，目前深度學習方法已被廣泛用於自動音樂情緒辨識。音樂情緒辨識主要以歌曲情緒為主，部分研究關注英文歌詞，罕見對於中文歌詞情緒辨識的研究。本研究提出運用遷移式學習改善深度神經網路模型—BERT 預訓練模型在中文歌詞的情緒分類任務上。實驗結果顯示，直接使用 BERT 對中文維度情緒語料庫建立中文情緒分類模型，對中文歌詞情緒分類僅有 50%的準確度，若使用 BERT 對中文維度情緒字典與片語建立情緒分類模型再遷移至中文維度情緒語料庫，能達到 71%的歌詞情緒分類準確度。

Abstract

Emotion is an important attribute in music information retrieval. Deep learning methods have been widely used in the automatic recognition of music emotion. Most of the studies focus on the audio data, the role of lyrics in music emotion classification remains under-appreciated. Due to the richness of English language resources, most previous studies were based on English lyrics but rarely in Chinese. This study proposes an approach without specific training for the Chinese lyrics

* 國立中興大學資訊管理學系
Department of Management Information Systems, National Chung Hsing University.
+ 國立臺中科技大學資訊工程系
Department of Computer Science and Information Engineering National Taichung University of Science and Technology
E-mail: jiaweichang.gary@gmail.com
The author for correspondence is Jia-Wei Chang.

emotional classification task: using transfer learning to improve deep neural networks, BERT pre-training model, for the emotion classification in Chinese lyrics. The experimental results show that directly using BERT to build an emotion classification model of CVAT only reach 50% of the classification accuracy. However, using BERT with transfer learning from CVAW, CVAP, to CVAT can achieve 71% classification accuracy.

關鍵詞：自然語言處理，音樂情緒辨識，遷移學習，中文歌詞

Keywords: Natural Language Processing, Music Emotion Recognition, Transfer learning, Chinese Lyrics.

1. 緒論 (Introduction)

音樂和人類情緒相互影響，在生活中扮演不可或缺的角色。音樂的搜尋通常以歌曲標題、詞曲作者、演唱者和演奏流派進行檢索，然而，情緒可以作為音樂的一個新且重要的搜尋屬性。隨著音樂串流平台使用者和歌曲庫的爆炸式增長，傳統的由專家進行情緒標註已不能滿足實際需求，推薦系統需要更快速的標註方法，自動情緒辨識因此成為重要的議題。音樂情緒辨識(Music Emotion Recognition) 用於觀察音樂與人類情感之相關性、對音樂抽取特徵並加以分析找出音樂特徵與人類對於音樂情緒感知的關聯。目前機器學習和深度學習方法已被廣泛用於辨識音樂的情緒。

支持向量機(Support Vector Machine, SVM)和支持向量回歸(Support Vector Regression, SVR)等機器學習方法(Han *et al*., 2009)。基於歌詞和音訊的歌曲情緒檢測方法來計算效價(Valence)和喚醒(Arousal)進行音樂之情緒分類(Jamdar *et al*., 2015)。用卷積神經網路預訓練模型對每 30 秒剪輯的印度古典音樂進行音樂情緒分類(Sarkar *et al*., 2021)。上述研究大多都集中利用聲學特徵進行音樂情緒辨識並無討論歌詞對於情緒的影響。歌詞被賦予情緒，在引發人類的情緒以及預測音樂情緒扮演著重要的角色(Hu & Downie, 2010)。雖然旋律和歌詞會同時對聽眾產生影響，但聽眾對於歌詞內容的偏好能進一步反映聽眾的特徵和傾向(Qiu *et al*., 2019)。Agrawal *et al*. (2021)提出歌詞可視為一連串彼此相關的句子，需捕捉上下文和長期依賴的關係，並在研究使用 Transformer-based 的模型進行歌詞情緒辨識，在多個英文歌詞情緒資料集上取得良好的成果，上述的英文歌詞資料集皆基於 Russell (1980)的 Valence-Arousal 心理學環繞情感模型進行音樂情緒的標註。

本研究提出一中文歌詞情緒分類方法。首先，運用基於 Transformer 語言預訓練模型對中文維度情緒字典(CVAW)與中文維度情緒片語(CVAP)進行建模，其次將模型遷移至中文維度情緒語料(CVAT)，最後將模型直接用於無標註的歌詞文本進行情緒的自動標註。

本研究其餘章節的組織如下：第二節回顧了心理學維度情緒模型、基於 Transformer 之模型和遷移學習的相關文獻。第三節的方法論說明了本研究所使用的兩個資料集、文本預處理並解釋本研究提出的架構。第四節為本研究模型訓練和歌詞驗證的結果。第五節對實驗結果進行相關討論。最後，在第六節總結本研究的成果和未來改進方向。

2. 文獻回顧 (Literature Review)

2.1 維度情緒模型 (Dimensional Models of Emotion)

情緒被視為一個連續體而非離散的形容詞，維度模型比起分類模型有著較低的歧義(Yang *et al*., 2008)。現有的研究大多採用 Russell (1980) 所提出的心理學環繞模型。Laurier *et al.* (2009) 的研究中表明，Russell 心理學情緒模型可以用於情緒分析或音樂情緒辨識任務。該維度模型的兩個維度的連續數值，分別為效價(Valence)和喚醒(Arousal)。效價(Valence)代表所有情緒體驗所固有的積極或消極，高效價(Valence)的歌曲聽起來更為積極、快樂，低效價(Valence)的歌曲聽起來較沮喪、憤怒。喚醒(Arousal)代表情緒的激動程度，歌曲的能量(energy)能對應於喚醒(Arousal)值，代表歌曲強度，能量高的歌曲通常越快速、響亮和強烈(Kim *et al*., 2011)。

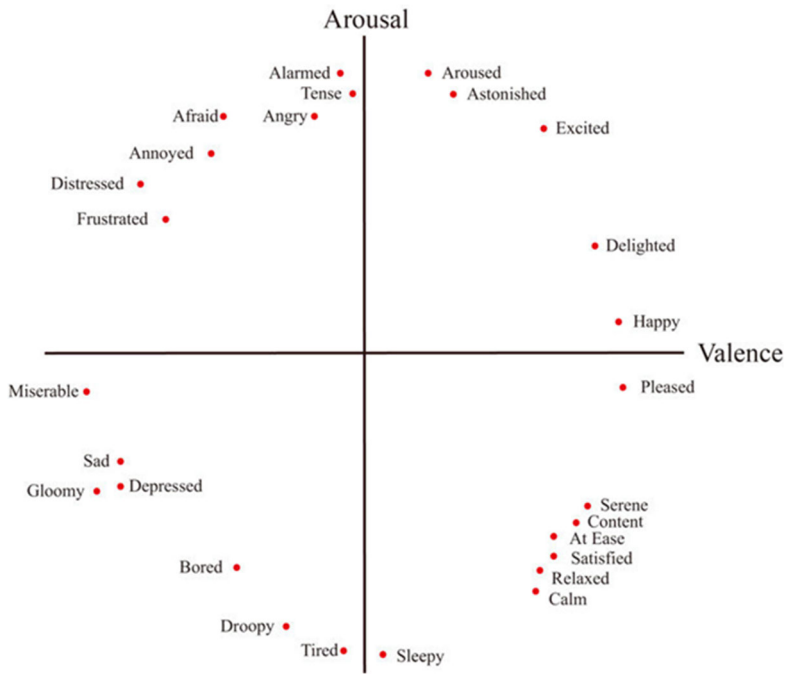


圖 1. Russell 提出之心理學維度情緒模型
[Figure 1. The circumplex model of affect (Russell, 1980)]

如圖 1 所示，情緒由效價(Valence)和喚醒(Arousal)兩個維度表示，情緒平面被分為四個象限，創建了四個情緒類別空間。在 Çano & Morisio (2017a)的研究中基於 Russell 維度情緒模型的四個象限將情緒分為四類別(Q1、Q2、Q3、Q4)，分別為快樂、憤怒、悲傷和輕鬆。因此，本研究在歌詞驗證的部分也依此方法將歌詞情緒分為四個象限類別。

2.2 基於Transformer之模型 (Transformer-based Model)

過去文本情緒分析是使用基於統計的詞袋模型和靜態特徵的詞向量模型將文本轉為向量特徵(Barry, 2017; Han *et al*., 2013)，但這些方法會遇到無法解讀多義詞的瓶頸。歌詞被視為是敘事而非彼此獨立的句子，需要捕捉上下文的依賴關係，在歌詞的音樂情緒分類任務上，若基於傳統詞典進行效果有限(Hu & Downie, 2010; Hu *et al*., 2009)，Abdillah *et al*. (2020)運用能捕捉時序關係的雙向長短期記憶(Long Short-Term Memory，LSTM)對 MoodyLyrics 資料集(Çano & Morisio, 2017b)進行歌詞的情緒分類，但遞歸架構難以具備平行運算的能力。Transformer (Vaswani *et al*., 2017)則改變過去序列網路的做法，自注意力機制藉由 Scaled dot-product Attention 讓資料得以平行運算， 考慮詞在不同空間映射的重要性，允許 BERT (Devlin *et al*., 2018)預訓練模型在多項任務中取得突破，包含實體辯識、序列或句子對分類、問答等 11 種任務，使得 Transformer-based 的模型架構在自然語言領域中成為主流。在歌詞情緒辨識的應用上，Agrawal *et al*. (2021)的研究便是使用基於 Transformer 的語言模型作為情緒分類的基礎架構，在多個英文歌詞情緒資料集上達到良好的成果，展示 Transformer-based 方法的高效能。

2.3 遷移學習 (Transfer Learning)

在某些領域中標籤的標記昂貴，若原始資料中含有標籤的數量太少，容易造成模型過度擬合。遷移學習常用的兩個技巧：特徵萃取和微調。遷移學習的有效性催生了多種應用，例如：學習情緒辯識(Hung *et al*., 2019)、時間序列任務(Fawaz *et al*., 2018)、3D 醫學影像分析(Chen *et al*., 2019)。在自然語言處理領域，也常運用遷移學習的技巧對於預訓練模型進行模型微調或特徵萃取，Transformer-based 的預訓練模型，證明微調在無註釋語料上預訓練大規模語言模型的有效性。Hung & Chang (2021)則提到多層遷移學習的有效性，表明了不管在電腦視覺任務或自然語言處理任務，經遷移學習的結果會優於未經過遷移的結果，因此，本篇研究提出的模型架構基於語言預訓練模型對文本進行遷移學習。

3. 方法論 (Methodology)

3.1 資料集 (Datasets)

- 中文維度情緒資料集(Yu *et al*., 2016; Yu *et al*., 2017)：資料如表 1 所示，包含中文情緒字典(Chinese Valence-Arousal Words, CVAW)、中文維度情緒片語(Chinese Valence-Arousal Phrases, CVAP)以及中文情緒語料庫(Chinese Valence-Arousal Text, CVAT)三個。CVAW 包含 5,512 個中文情緒詞；CVAP 中每個片語結合修飾符和來自 CVAW 中的詞，共 2,998 個中文情緒片語；CVAT 則從 720 篇來自 6 種不同類別的網路文章蒐集而來，共 2,009 個句子。三個資料集的每個詞或句子皆包含效價(Valence)和喚醒(Arousal)。效價(Valence)的範圍從 1 到 9 其分別代表極端負面和極端正面的情緒，喚醒(Arousal)的範圍從 1 到 9 其分別代表平靜和激動，效價(Valence)和喚醒(Arousal)若為 5 則代表沒有特定傾向的中性情緒。

- 歌詞資料集：為本研究自行收集並標籤的資料集。標籤包含象限一(Q1)、象限二(Q2)、象限三(Q3)及象限四(Q4)。Q1 代表正向激昂共 43 首，Q2 代表負向激昂共 45 首，Q3 代表負向平靜共 43 首，Q4 代表正向平靜共 39 首。V 和 A 分別代表效價(Valence)和喚醒(Arousal)，V 標記 1 代表正向情緒、0 代表負向情緒，A 標記為 1 代表激昂情緒、0 代表平靜情緒。

表 1. 中文維度情緒資料集
[Table 1. The datasets of Chinese valence-arousal]

名稱	總數	範例文字	Valence	Arousal
中文維度情緒字典 CVAW	5,512	不爽	2.8	7.2
中文維度情緒片語 CVAP	2,998	非常可愛	8	7.313
中文維度情緒語料庫 CVAT	2,009	這種記錄難免空洞，虛構也顯得薄弱。	3	3.5

3.2 提出之架構 (Proposed Architecture)

本研究提出的模型架構如圖 2，透過 BERT 預訓練模型建立 CVAT 中文維度情緒模型，將此模型直接用於歌詞情緒的標記，驗證在未學習過歌詞文本的情況下模型的成效。本章總共有三個小節，第一小節說明資料預處理，第二小節介紹模型實作的細節以及實驗的參數設定，第三小結討論將模型應用於歌詞文本情緒驗證的方法。

3.2.1 資料預處理 (Data Preproceessing)

CVAW、CVAP 和 CVAT 皆採用資料集內的文字、效價(Valence)平均和喚醒(Arousal)平均。由於 CVAW、CVAP 的文字較短且類似，因此將兩個資料集合併成一個資料集，稱 CVAW+CVAP，以 8 比 2 拆分為訓練集跟測試集。BERT 模型有別於傳統文本的方法，會將標點符號視為一個特徵值進行訓練，因此 CVAT 文字不進行刪除標點符號的預處理。歌詞的資料集共 170 首，由三位標註者將每首歌曲的針對效價(Valence)和喚醒(Arousal)分別標註為正或負，以中性情緒為原點，依照效價(Valence)和喚醒(Arousal)的正跟負分標記到四個象限。BERT 能夠訓練的最大文本長度為 512，考慮到 CVAP 和 CVAW 的文字都在 10 字以內，而 CVAT 的文本分佈大多集中在 100 字以內，為了避免產生過於稀疏向量，最大文本長度設定為 256 而非 512。輸入 BERT 模型前必須在每個序列開頭加上特殊字元符號[CLS]，此特殊字元代表整個輸入序列的向量表示，在序列尾巴則加上特殊字元符號[SEP]作為文本的結束，每個中文字會對應到 BERT 中文字典的一個索引值稱為 Token id，為了讓每一則輸入序列的長度保持一致，若文字長度不足則會在文字序列後端填充特殊字元[PAD]，最後轉為向量的序列和目標值轉為張量(Tensor)至 BERT 模型進行訓練。

3.2.2 實施細節 (Implementation Details)

本研究以知名的深度神經網路模型—BERT(Devlin *et al.*, 2018)為基礎架構，並進一步提出了多輸出(Multi-output)與單輸出(Single-output)兩種模型訓練架構。如圖 2 所示，多輸出(Multi-output)架構為一個 BERT 模型共享權重，一次輸出效價(Valence)和喚醒(Arousal)兩個連續值，單輸出(Single-output)為一個 BERT 模型輸出單一個值(例如 Valence)。由於是效價(Valence)、喚醒(Arousal)的數值預測，因此模型訓練時的損失函數選擇使用均方誤差(Mean square error, MSE)。兩種模型架構都實驗兩種方法：(a)使用從 CVAW+CVAP 遷移至 CVAT 資料集的遷移學習方法。(b)從零訓練 CVAT 未遷移的方法。最後進行兩種方法的比較。本研究基於微調方法進行實驗，微調方法的優點在於模型的許多參數不需要重新學習，即使只有少量訓練樣本也能達到良好的效果。在模型微調方面，在 BERT 預訓練模型加上一層 Dropout 和一層線性分類層，優化器為 Adam，學習速率本研究嘗試多種學習速率進行實驗，微調模型適合較小的學習速率避免預訓練的權重被修改破壞，也在實驗中發現，若學習速率不夠小會導致損失(loss)無法降低，最後選擇了 1e-05、1e-06 和 5e-05 三個超參數進行近一步實驗及比較，每次訓練最大 Epoch 設定為 100，加入 Early Stopping 的機制，將耐心(Patience)設至為 10。

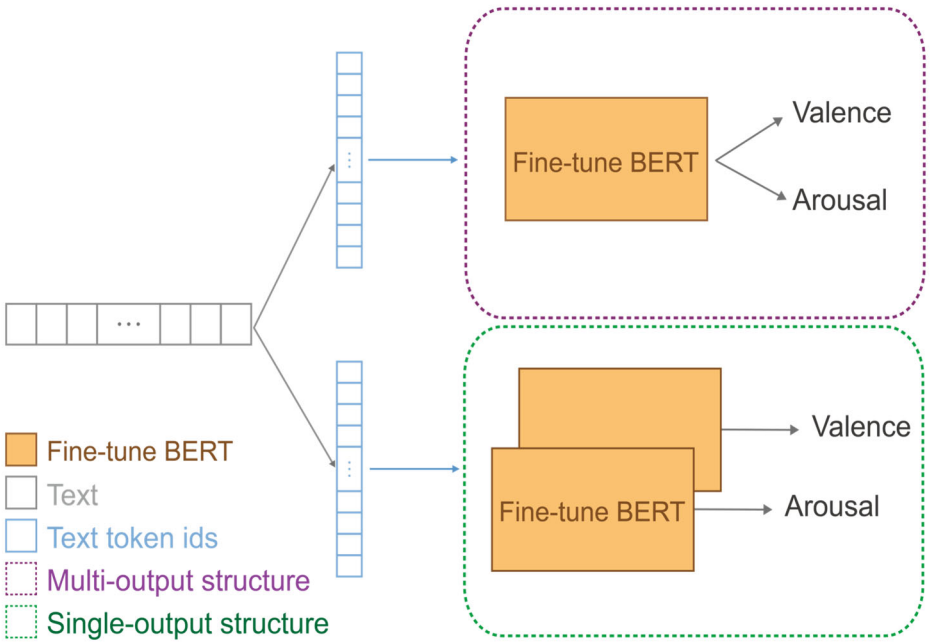


圖 2. 本研究提出之模型架構：包含兩種架構分別為單輸出與多輸出
[Figure 2. Training architecture of multi-output and single-output models]

3.2.3 歌詞情緒之分類 (Lyrics Emotion Classification)

此階段的目的在於驗證本研究提出的方法能在未學習過歌詞文本的情況下，能對於歌詞文本進行情緒的標註。首先，將歌詞文本進行與第一小節同樣的預處理後送入模型進行預測，輸出效價(Valence)和喚醒(Arousal)，其範圍為 0 到 9。本研究依照原資料集的敘述(Yu *et al*., 2016; Yu *et al*., 2017)，效價(Valence)和喚醒(Arousal)都以中性值 5 為閾值，因此，若預測出的效價(Valence)數值大於 5 表示模型預測該歌詞為正向情緒並標記為 1、效價(Valence)數值小於 5 則表示模型預測該歌詞為負向情緒並標記為 0，若預測喚醒(Arousal)數值大於 5 則表示模型預測該歌詞為激動情緒並標記為 1、喚醒(Arousal)值小於 5 表示模型預測該歌詞為平靜情緒並標記為 0。我們將效價(Valence)和喚醒(Arousal)標記之後的結果轉為四個象限 Q1、Q2、Q3 和 Q4 的情緒分類之結果，最後驗證其分類效果。

4. 實驗結果 (Experimental Result)

本章節將實驗結果分為兩個階段，第一階段是中文情緒的模型訓練結果，第二階段是驗證模型預測歌詞情緒的成效，每個段落包含在不同的模型架構和不同訓練方式的實驗結果。

4.1 中文情緒模型 (Chinese Emotion Model)

訓練模型的資料集切分皆以 8 比 2 進行，CVAP+CVAW 的訓練集和測試集分別為 6808 筆和 1702 筆。在多輸出(Multi-output)與單輸出(Single-output)架構的訓練結果，如表 2 所示，多輸出架構模型的均方誤差為 0.59126，單輸出模型的效價(Valence)和喚醒(Arousal)的均方誤差(MSE)分別為 0.3788 和 0.77339，兩個模型架構最佳的學習速率皆為 1e-05。

表 2. 在 CVAW + CVAP 資料集上之結果：包含多輸出(Multi-output) 與單輸出(Single-output) 兩種架構之結果
[Table 2. Results on multi-output and single-output models]

架構名稱	輸出	學習速率 (Lr)	損失 (loss)	Epoch
多輸出架構 Multi-output	-	1e-5	**0.59126**	14
		1e-6	0.65283	32
		5e-5	0.69301	17
單輸出架構 Single-output	Valence	1e-5	**0.3788**	24
		1e-6	0.39498	35
		5e-5	0.51918	4
	Arousal	1e-5	**0.77339**	12
		1e-6	0.92874	19
		5e-5	1.8867	12

如表 3 所示在多輸出架構底下，從零訓練 CVAT 資料集(Training From Scratch)和從 CVAP+CVAW 資料集模型遷移至 CVAT 資料集(Transfer Learning)的結果來看，兩者同樣都在學習速率皆為 1e-05 的訓練效果最佳，經過遷移的均方誤差為 0.65696 優於未經遷移的 0.72025。經過遷移的模型結果比未經遷移效果好，在不同學習速率下，經遷移的 CVAT 在不同的學習速率下都優於未經遷移的結果，有經過遷移學習的 CVAT 收斂速度也比未經遷移學習的快。表 4 為在單輸出架構底下，從零訓練 CVAT 資料集(Training From Scratch)和從 CVAP+CVAW 資料集模型遷移至 CVAT 資料集(Transfer Learning)的結果，單輸出(Single-output)架構是將效價(Valence)和喚醒(Arousal)作為獨立的兩個輸出，首先比較效價(Valence)輸出的結果，未經遷移(Training From Scratch)的均方誤差(MSE)為 0.50338，而經遷移學習(Transfer Learning)的均方誤差(MSE)為 0.46624，顯示經遷移學習的 CVAT 其結果優於未經遷移的結果。經遷移學習的最佳學習速率為 1e-06，未經遷移的最佳學習速率為 1e-5，就算同樣都在 1e-5 的學習速率底下，經遷移學習的均方誤差 0.47898 依然是優於未經遷移的均方誤差 0.50338。比較輸出為喚醒(Arousal)的結果，經遷移學習的 CVAT 其均方誤差為 0.84259 優於未經遷移的 0.87107，兩者同樣都在學習速率為 1e-05 的時候得到最佳結果。

表 3. 多輸出架構(Multi-output)經遷移學習與未經遷移學習在 CVAP 資料集之結果

[Table 3. CVAP results on the multi-output model with/without transfer learning]

方法	學習速率 (Lr)	損失 (loss)	Epoch
從 0 訓練 CVAT From Scratch	1e-5	**0.72025**	10
	1e-6	0.73979	58
	5e-5	0.80925	10
經遷移學習 Transfer Learning	1e-5	**0.65696**	3
	1e-6	0.67836	22
	5e-5	0.70594	2

表 4. 單輸出架構(Single-output)經遷移學習與未經遷移學習在 CVAP 資料集之結果
[Table 4. CVAP results on the single-output model with/without transfer learning]

方法	輸出	學習速率 (Lr)	損失 (loss)	Epoch
從 0 訓練 CVAT From Scratch	Valence	1e-5	**0.50338**	12
		1e-6	0.51199	44
		5e-5	0.55236	6
	Arousal	1e-5	**0.87107**	5
		1e-6	0.93317	28
		5e-5	0.9303	10
經遷移學習 Transfer Learning	Valence	1e-5	0.47898	4
		1e-6	**0.46624**	15
		5e-5	0.53422	5
	Arousal	1e-5	0.84259	1
		1e-6	0.88142	7
		5e-5	0.93479	11

4.2 中文歌詞情緒模型之驗證 (Verification of Chinese Lyrics Emotion Model)

此段落討論前述的中文情緒模型應用於中文歌詞分類之結果，第一小節描述中文歌詞的分類結果，第二小節更進一步說明 Valence-Arousal 平面的分類結果。中文歌詞的情緒分類是將模型輸出的效價(Valence)和喚醒(Arousal)基於中性值 5 作為閾值，轉換為座標平面上的四個象限類別(Q1、Q2、Q3、Q4)。歌詞情緒分類結果如表 8 所示，包含歌名、歌詞、模型預測的 Valence 數值和 Arousal 數值、預測的標籤和真實標籤。

4.2.1 中文歌詞情緒分類之結果 (Chinese Lyrics Emotion Classification Results)

經遷移學習 CVAT 模型與未經遷移學習 CVAT 模型的歌詞情緒分類的混淆矩陣的結果，如表 5 所示，在經遷移學習的 CVAT 分類，Q1 有 26%容易被錯分成 Q2，19%會被錯分成 Q4，僅有 2.3%會被分成 Q3，也就是在 Q1 的情緒類別中，效價(Valence)和喚醒(Arousal)都有被分類錯的可能，效價(Valence)和喚醒(Arousal)同時被錯分的機率僅 2.3%。Q4 有 56% 被錯分為 Q3，被分成 Q2 的可能為 5%，僅有 2.5%會被分成 Q1，也就是在 Q4 的情緒類別中，僅效價(Valence)容易被分類錯誤。Q2 的情緒幾乎都能夠準確辨識，僅有 2%會被錯分成 Q3，只有 2%會因為喚醒(Arousal)被錯誤分類。Q3 有 16%會被錯分成 Q2，其餘的都能被正確分類，也就是在 Q3 的情緒只有 16%會因為喚醒(Arousal)被錯誤分類。

未經遷移的 CVAT 分類結果中，Q1 有 14%被錯分成 Q2，25.6%被錯分成 Q3，35%被錯分成 Q4，Q2 只有 62%分類正確，其餘 37%皆被錯分成 Q3，Q3 有 14%被錯分成 Q2，其餘分類正確，Q4 有 7.7%被錯分成 Q2，69%容易被錯分成 Q3 ，只有 29%分類正確。

經遷移學習 CVAT 模型與未經遷移學習 CVAT 模型的歌詞情緒分類結果，如表 6 所示，經遷移學習的 CVAT 模型在歌詞情緒分類的準確度為 0.71，標籤 Q1 和 Q4 的 F1-score 較低，分別為 0.69 和 0.51，而 Q2 和 Q3 的 F1-score 較高，分別為 0.83 和 0.72。未經遷移學習的 CVAT 模型在歌詞情緒分類的準確度為 0.50，同樣是標籤 Q1 和 Q4 的 F1-score 較低，分別為 0.41 和 0.29，而 Q2 和 Q3 的 F1-score 較高，分別為 0.64 和 0.55。比較經遷移學習的模型與未經遷移學習的模型，經遷移學習的模型中每一個情緒標籤的分類結果都優於未經遷移學習的模型，可得知到在訓練階段 CVAT 模型學習效果較佳的模型，應用在歌詞的情緒分類也能得到較佳的結果，表示經遷移學習的模型在 CVAW+CVAP 資料集中所學習到的中文情緒特徵，有助於提升模型在歌詞文本的情緒辨識能力。

表 5. 歌詞情緒分類之混肴矩陣：經遷移學習與未經遷移學習
[Table 5. Results on the model with/without transfer learning by confusion matrix]

Prediction by CVAT: Transfer Learning					
True		Q1	Q2	Q3	Q4
	Q1	**23**	8	1	11
	Q2	0	**44**	1	0
	Q3	0	7	**36**	0
	Q4	1	2	**19**	17
Prediction by CVAT: Training from Scratch					
True		Q1	Q2	Q3	Q4
	Q1	**11**	6	11	15
	Q2	0	**28**	17	0
	Q3	0	6	**37**	0
	Q4	0	3	**27**	9

表 6. 歌詞分類結果之分數：經遷移學習與未經遷移學習
[Table 6. Results on the model with/without transfer learning]

CVAT : Transfer Learning			
Label	Precision	Recall	F1-score
Q1	0.96	0.53	0.69
Q2	0.72	0.98	0.83
Q3	0.64	0.84	0.72
Q4	0.61	0.44	0.51
Accuracy	0.71		
CVAT: Training From Scratch			
Label	Precision	Recall	F1-score
Q1	1.00	0.26	0.41
Q2	0.65	0.62	0.64
Q3	0.40	0.86	0.55
Q4	0.38	0.23	0.29
Accuracy	0.50		

4.2.2 Valence-Arousal 分類之結果 (Valence-Arousal Plane Classification Result)

此小節將說明效價(Valence)和喚醒(Arousal)的預測結果。經遷移學習的 CVAT 模型在 Valence-Arousal 平面之分類結果，如表 7 所示效價(Valence)和喚醒(Arousal)的分類準確率都為 0.76。在效價(Valence)的負向情緒中，Recall 為 1，表示 80 首負向情緒的歌詞都被正確分類，Precision 為 0.67，表示被預測為負向的歌詞總共有 120 首，有 80 首被正確分為負向情緒，但有 40 首標籤應為正向情緒的歌詞被錯誤分類為負向情緒。效價(Valence)的正向情緒中，Recall 為 0.56，表示有 50 首正向情緒的歌詞被正確分類，但有 40 首正向情緒的樣本被分類為負向情緒，Precision 為 1，表示被預測為正向情緒的歌詞總共有 50 首，而 50 首皆被正確分類。在喚醒(Arousal)的激動情緒中，Recall 為 0.71，表示總共有 105 首歌詞應被分類為激動情緒，有 75 首被正確分類、30 首應該被分類為激動情緒的歌詞被錯誤分類為平靜情緒，Precision 為 0.88，表示被預測為激動情緒的歌詞總共有 85 首，其中，有 75 首歌詞被正確分類為激動情緒，但有 10 首被錯誤分類為激動情緒。在喚醒(Arousal)的平靜情緒中，Recall 為 0.85，表示共有 65 首歌詞應被分類為平靜情緒，有 55 首被正確分類，有 10 首應被分成平靜情緒的歌詞被錯誤分類為激動情緒，Precision 為 0.65，表示被預測為平靜情緒的歌詞總共有 85 首，55 首歌詞被正確預測為平靜情緒，但有 30 首被錯誤分類為平靜情緒。

表 7. 經遷移學習 CVAT 模型之 Valence-Arousal 平面分類之結果
[Table 7. Classification results on the four categories of valence-arousal]

Valence			
Label	Precision	Recall	F1-score
1 (+)	1.00	0.56	0.71
0 (-)	0.67	1.00	0.80
Accuracy	0.76		
Arousal			
Label	Precision	Recall	F1-score
1 (+)	0.88	0.71	0.79
0 (-)	0.65	0.85	0.73
Accuracy	0.76		

5. 討論 (Discussion)

從單輸出模型架構的結果發現喚醒(Arousal)的特徵較難學習，該結果在多個研究中都有提到(Malheiro *et al*., 2016; Yu *et al*., 2016; Çano & Morisio, 2017b)，發現在無論中文或者英文的資料上，文字的喚醒(Arousal)維度較難以區分，推測激動程度或強度在文字上較難以顯示出來。不論在多輸出還是單輸出的模型架構底下，經過遷移的結果都優於未經遷移的結果且能提高模型的收斂速度，證明在 CVAW 和 CVAP 兩個資料集所學習到的特徵，有助於模型對 CVAT 中文維度情緒語料庫的學習。在驗證模型能否應用於歌詞文本的實驗結果中觀察到，CVAT 訓練結果較佳的遷移模型，應用於歌詞文本分類使其結果也會較佳，優於未遷移的 CVAT 模型，顯示經遷移學習學到的特徵是有助歌詞文本的情緒辨識成果，並且在未學習過歌詞文本的狀況下達到 71%的準確率。

另外，我們推測分類錯誤的原因是歌詞中的情緒在不同段落中可能帶有不同情緒，因此，較難以分類為單一情緒類別。以表 8 中，其真實標籤(Label)為正向激動(Q1)但模型預測(Predict)為負向激動(Q2)的「美好」這首歌曲進行每個段落的歌詞情緒分析，結果如表 9 所示，若只看段落一的句子中「你哭著說再見」、「揮手和你道別」會解讀出字句裡包含負面情緒，模型如期預測為負向情緒，相反的在段落六的句子中，「你是全世界的美好」、「你有最美麗的微笑」、「比你更重要就是和你 一起變老」，句子間流露出滿足、幸福等正向的情緒，模型如期預測為正向情緒。在段落三與段落五當中，兩段歌詞的標籤都是正向激昂的(Q1)，兩段歌詞差別只在於最後一句的不同，分別是「我不知道也永遠不要知道」和「我不知道也不用知道」，比較兩句話對於模型預測喚醒(Arousal)程度的影響，可以看出段落三的激動情緒為 5.40，略高於段落五的 5.371，我們推測第三段的歌詞，其關鍵字眼「永遠」提高整句話的激昂程度。在段落二與段落四當也能看到

類似的結果，段落二比起段落四多一句「對不對」，「對不對」表達強調或質問的語氣，因此加強句子的喚醒(Arousal)程度，從模型預測結果來看，段落二的喚醒(Arousal)也從4.88 提升至 4.94。

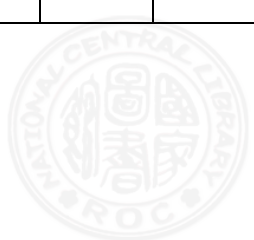
總結以上的結果說明本研究提出的模型能辨別出不同句子的正、負向情緒和字句間喚醒(Arousal)程度的差別，同時，我們觀察到在某些歌詞中其實隱含多種情緒，若歌詞只進行單一類別標註可能會對於情緒的標記不夠全面，因為當不同標註者對於歌詞關注的地方不同，標註的情緒也會有所不同。

表 8. 中文歌詞情緒分類之結果
[Table 8. Classification results on Chinese lyrics emotion]

歌曲	歌詞（未顯示完整歌詞）	V	A	Predict	Label
美好	我看見風吹過你的臉 有那年的感覺 那一年你哭著說再見 而我揮手和你道別	4.936	5.038	Q1	Q1
厭世吉娃娃	吉娃娃 吉娃娃 我是吉娃娃 我討厭爸爸 我討厭媽媽 我更討厭這個世界啊	2.654	6.474	Q2	Q2
你	風輕輕 我聽見你聲音 你對著我叮嚀 要注意自己的心情 風輕輕 我聽見你聲音 你拿著傘靠近	5.800	4.836	Q4	Q4
哼情歌	在無關景要的場合都會想起這首歌 是因為你曾經哼唱著 在平淡無奇的眼神都會想起你呢	4.360	4.809	Q3	Q3

表 9. 歌曲「美好」依照句子分割之 Valence-Arousal 預測結果
[Table 9. Prediction of valence-arousal by the example sentences from a song]

段落	歌詞	V(+/-)	A(+/1)
1	我看見風吹過你的臉 有那年的感覺 那一年你哭著說再見 而我揮手和你道別	4.509(-)	5.07(+)
2	時間並沒有帶走一切 反而給我們更多洗鍊 如果分離製造了更多的想念 那我們是否該更加感謝	5.156(+)	4.88(-)
3	你是全世界的美好 你有最美麗的微笑 你問我有什麼比你重要 我不知道也永遠不要知道	5.113(+)	5.40(+)
4	時間並沒有帶走一切 反而給我們更多洗鍊 如果分離製造了更多的想念 那我們是否該更加感謝 對不對	4.583(-)	4.94(-)
5	你是全世界的美好 你有最美麗的微笑 你問我有什麼比你重要 我不知道 也不用知道	5.766(+)	5.371(+)
6	你是全世界的美好 你有最美麗的微笑 你問我有什麼比你重要 總算知道 比你更重要 就是和你 一起變老	6.094(+)	5.226(+)



6. 結論 (Conclusion)

本研究提出以基於 Transformer 的語言預訓練模型對中文情緒資料集進行學習，將中文情緒資料庫的模型直接用於歌詞的效價(Valence)和喚醒(Arousal)進行標註。在實驗中比較了有遷移學習與未經遷移學習的模型，實驗結果證明在中文情緒字典與中文情緒片語學習到的特徵，有助於中文情緒文本的學習。同時，本研究將經遷移學習及未遷移的模型用於歌詞的情緒分類，發現經遷移學習的模型結果優於未經遷移的模型，證明在中文情緒資料集學習結果較佳的模型，用於歌詞情緒分類其結果也會較佳。在不同模型架構底下，發現喚醒(Arousal)的情緒較難以學習。最後，我們注意到歌詞中的情緒在不同段落會帶有不同情緒，不僅模型在人工標註時，若對於關注到不同語句則會對於歌詞情緒有不同的判斷，在未來研究方向可以將一首歌的歌詞視為多種情緒以多標籤任務方式進行。

致謝 (Acknowledgements)

特別感謝科技部計畫(108-2218-E-025-002-MY3)，「人工智慧音樂家-運用深度嵌入方法與生成對抗網路與和諧性為導向的詞曲生成器之設計」對於本研究的支持。

參考文獻 (References)

Abdillah, J., Asror, I., & Wibowo, Y. F. A. (2020). Emotion classification of song lyrics using bidirectional lstm method with glove word representation weighting. *Jurnal RESTI* (Rekayasa Sistem Dan Teknologi Informasi), *4*(4), 723-729. https://doi.org/10.29207/resti.v4i4.2156

Agrawal, Y., Shanker, R. G. R., & Alluri, V. (2021). Transformer-based approach towards music emotion recognition from lyrics. arXiv preprint arXiv:2101.02051.

Barry, J. (2017). Sentiment Analysis of Online Reviews Using Bag-of-Words and LSTM Approaches. In *AICS*, 272-274.

Chen, S., Ma, K., & Zheng, Y. (2019). Med3d: Transfer learning for 3d medical image analysis. arXiv preprint arXiv:1904.00625.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Fawaz, H. I., Forestier, G., Weber, J., Idoumghar, L., & Muller, P. A. (2018). Transfer learning for time series classification. arXiv e-prints, arXiv:1811 .01533.

Han, B. J., Rho, S., Dannenberg, R. B., & Hwang, E. (2009). SMERS: Music Emotion Recognition Using Support Vector Regression. In *Proceedings of 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, 651-656.

Han, Q., Guo, J., & Schuetze, H. (2013). Codex: Combining an svm classifier and character n-gram language models for sentiment analysis on twitter text. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, 520-524.

Hu, X., & Downie, J. S. (2010). When Lyrics Outperform Audio for Music Mood Classification: A Feature Analysis. In *Proceedings of 11th International Society for Music Information Retrieval Conference (ISMIR 2010)*, 619-624.

Hu, Y., Chen, X., & Yang, D. (2009). Lyric-based Song Emotion Detection with Affective Lexicon and Fuzzy Clustering Method. In *Proceedings of 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, 123-128.

Hung, J. C., Lin, K. C., & Lai, N. X. (2019). Recognizing learning emotion based on convolutional neural networks and transfer learning. *Applied Soft Computing*, 84, 105724. https://doi.org/10.1016/j.asoc.2019.105724

Hung, J. C., & Chang, J. W. (2021). Multi-level transfer learning for improving the performance of deep neural networks: Theory and practice from the tasks of facial emotion recognition and named entity recognition. *Applied Soft Computing*, 109, 107491. https://doi.org/10.1016/j.asoc.2021.107491

Jamdar, A., Abraham, J., Khanna, K., & Dubey, R. (2015). Emotion analysis of songs based on lyrical and audio features. arXiv preprint arXiv:1506.05012.

Kim, J., Lee, S., Kim, S., & Yoo, W. Y. (2011). Music mood classification model based on arousal-valence values. In *Proceedings of 13th International Conference on Advanced Communication Technology (ICACT 2011)*, 292-295.

Laurier, C., Sordo, M., Serra, J., & Herrera, P. (2009). Music Mood Representations from Social Tags. In *Proceedings of 10th International Society for Music Information Retrieval Conference (ISMIR 2009)*, 381-386.

Malheiro, R., Panda, R., Gomes, P., & Paiva, R. P. (2016). Emotionally-relevant features for classification and regression of music lyrics. *IEEE Transactions on Affective Computing*, *9*(2), 240-254. https://doi.org/10.1109/TAFFC.2016.2598569

Qiu, L., Chen, J., Ramsay, J., & Lu, J. (2019). Personality predicts words in favorite songs. *Journal of Research in Personality*, 78, 25-35. https://doi.org/10.1016/j.jrp.2018.11.004

Russell, J. A. (1980). A circumplex model of affect. *Journal of personality and social psychology*, *39*(6), 1161- 1178. https://doi.org/10.1037/h0077714

Sarkar, U., Nag, S., Basu, M., Banerjee, A., Sanyal, S., Sengupta, R., & Ghosh, D. (2021). Neural Network architectures to classify emotions in Indian Classical Music. arXiv preprint arXiv:2102.00616.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st International Conference on neural information processing systems (NIPS '17)*, 5998-6008.

Yang, Y. H., Lin, Y. C., Su, Y. F., & Chen, H. H. (2008). A regression approach to music emotion recognition. *IEEE Transactions on audio, speech, and language processing*, *16*(2), 448-457. https://doi.org/10.1109/TASL.2007.911513

Yu, L. C., Lee, L. H., Hao, S., Wang, J., He, Y., Hu, J., Lai, K. R., & Zhang, X. (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: Human Language Technologies, 540-545. https://doi.org/10.18653/v1/N16-1066

Yu, L. C., Lee, L. H., Wang, J., & Wong, K. F. (2017). IJCNLP-2017 Task 2: Dimensional Sentiment Analysis for Chinese Phrases. In *Proceedings of the IJCNLP 2017, Shared Tasks*, 9-16.

Çano, E., & Morisio, M. (2017a). Music mood dataset creation based on last. fm tags. In *Proceedings of Fourth International Conference on Artificial Intelligence and Applications (AIAP 2017)*, 15-26. https://doi.org/10.5121/csit.2017.70603

Çano, E., & Morisio, M. (2017b). Moodylyrics: A sentiment annotated lyrics dataset. In *Proceedings of the 2017 International Conference on Intelligent Systems, Metaheuristics & Swarm Intelligence (ISMSI '17)*, 118-124. http://dx.doi.org/10.1145/3059336.3059340