The 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)
Taoyuan, Taiwan, October 15-16, 2021. The Association for Computational Linguistics and Chinese Language Processing

運用遷移式學習改善 BERT 於中文歌詞情緒分類模型之研發
A Study on Using Transfer Learning to Improve BERT Model for Emotional Classification of Chinese Lyrics

廖家誼 **Jia-Yi Liao**[1]
plusoneee@smail.nchu.edu.tw

林亞宣 **Ya-Hsuan Lin** [1]

林冠成 **Kuan-Cheng Lin** [1]

張家瑋 **Jia-Wei Chang** [2]
jiaweichang.gary@gmail.com

[1] 國立中興大學資訊管理學系
Department of Management Information Systems
National Chung Hsing University

[2] 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

摘要

音樂庫的爆炸增長讓音樂資訊檢索和推薦成爲重要議題，以音樂情緒辨識爲基礎的推薦系統逐漸受到研究者的重視。音樂情緒辨識主要以歌曲情緒爲主，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。因此，本研究提出利用 BERT 預訓練模型和遷移學習來改善中文歌詞的情緒分類任務。實驗結果顯示，在未針對歌詞情緒分類任務訓練下:(a) 使用 BERT 針對 CVAT 建立之分類模型，只能達到 50% 的歌詞情緒分類準確度。(b) 使用 BERT 針對 CVAW+CVAP 建立分類模型再對 CVAT 資料集遷移學習後，能提升到 71% 的歌詞情緒分類準確度。

Abstract

The explosive growth of music libraries has made music information retrieval and recommendation a critical issue. Recommendation systems based on music emotion recognition are gradually gaining attention. Most of the studies focus on audio data rather than lyrics to build models of music emotion classification. In addition, because of the richness of English language resources, most of the existing studies are focused on English lyrics but rarely on Chinese. For this reason, We propose an approach that uses the BERT pre-training model and Transfer learning to improve the emotion classification task of Chinese lyrics. The following approaches were used without any specific training for the Chinese lyrics emotional classification task: (a) Using BERT, only can reach 50% of the classification accuracy. (b) Using BERT with transfer learning of CVAW, CVAP, and CVAT datasets can achieve 71% classification accuracy.

關鍵字：音樂情緒辨識、自然語言處理、中文歌詞

Keywords: Music Emotion Recognition, Natural Language Processing, Chinese Lyrics

1 緒論

音樂搜尋通常以歌曲標題、詞曲作者、演唱者和演奏流派做檢索。然而，情緒可以作爲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爲重要的議題。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，結合 ANEW 和 WordNet 來計算 Valence 和 Arousal 進行音樂情緒分類 (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., 2015)。

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 and Downie, 2010b)。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) 提出歌詞可視爲一連串彼此相關的句子，需捕捉上下文和長期依賴的關係，所以運用基於 Transformer 的模型進行歌詞情緒辨識，並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 的 Valence-Arousal 環繞模型進行音樂情緒的標註。截至本研究發表之前，尚未有中文歌詞文本包含情緒標註的大型資料集，因此，本研究提出一新的中文情緒辨識方法，運用基於 Transformer 語言預訓練模型對中文文字與片語進行建模，將模型遷移至中文文本資料集，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：第二節說明情緒模型、基於 Transformer 之模型和遷移學習的相關工作，第三節介紹本研究使用的資料集、文本預處理並解釋本研究提出的架構，第四節爲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，第六節總結本研究的成果。

2 相關研究

2.1 情緒維度模型

現有的研究大多採用Russell (1980) 提出的環繞模型。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。兩個維度的連續數值，分別爲 Valence 和 Arousal。Valence 代表所有情緒體驗所固有的積極或消極。Arousal 代表情緒的激動程度，歌曲的能量對應於 Arousal 值，代表歌曲強度 (Kim et al., 2011)。Çano and Morisio (2017a) 則基於 Russell 心理學情緒模型的四個象限將情緒分爲四類別，分別爲快樂、憤怒、悲傷和輕鬆 (Q1、Q2、Q3、Q4)，因此本研究在歌詞驗證也依此方法將歌詞情緒分爲四個象限類別。

2.2 基於 Transformer 之先進模型

歌詞被視爲是敘事而非彼此獨立的句子，需捕捉上下文的依賴關係，歌詞的音樂情緒分類任務若基於傳統詞典 (Barry 2017;Han et al. 2013) 進行效果有限 (Hu and Downie 2010a;Hu et al. 2009)。Abdillah et al. (2020) 運用捕捉時序關係的雙向長短期記憶 (Long Short-Term Memory，LSTM)，但遞歸架構難以具備平行運算的能力。Vaswani et al. (2017) 提出 Transformer 模型架構，該模型卓越的自注意力機制也使得目前自然語言處理領域中公認最先進的 BERT (Devlin et al., 2019) 亦以 Transformer 作爲模型設計的基礎。此外，Agrawal et al. (2021) 的研究使用基於 Transformer 作爲情緒分類的模型，在多個英文歌詞情緒資料集上達到傑出的成果，展現出 Transformer 的強大優勢。

2.3 遷移學習

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，容易過度擬合。遷移學習中有兩個常用的方法，特徵萃取和微調，特徵萃取技術是使用預先訓練好的模型作爲編碼器，爲目標任務提取有效的特徵。微調技術是將原有任務訓練所得到的模型結構和參數應用於目標任務的訓練，更新目標模型中的參數，達到提高訓練目標的學習能力。在自然語言處理領域，基於 Transformer 的預訓練模型 (Devlin et al., 2019) 已證明微調在大型無註釋語料庫上預訓練大規模語言模型的良好效能。Hung and Chang (2021) 則提到多層遷移學習在電腦視覺任務或自然語言處理任務的有效性，因此，本篇研究提出的模型架構便基於 Transformer 的語言預訓練模型對文本進行遷移學習。

3 方法論

3.1 資料集

- 中文情緒資料集 (Yu et al. 2016;Yu et al. 2017): 中文情緒字典 (CVAW)、中文情緒短語 (CVAP) 及中文情緒文本 (CVAT) 三個。CVAW 有 5,512 個中文情緒詞；CVAP 包含 2,998 個中文情緒片語；CVAT 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。每個詞或句子皆包含 Valence 和 Arousal 的數值，Valence 範圍從 1 到 9 分別代表極端負面和極端正面的情緒，Arousal 範圍從 1 到 9 分別代表平靜和激動，5 則代表沒有特定傾向的中性情緒。

- 歌詞資料集：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 共 43 首代表正向激昂，Q2 共 45 首代表負向激昂，Q3 共 43 首代表負向平靜，Q4 共 39 首代表正向平靜。V 和 A 分別代表 Valence 和 Arousal，V 標記 1 代表正向情緒、0 代表負向情緒，A 標記爲 1 代表激昂情緒、0 代表平靜情緒。

3.2 基於 BERT 的遷移式學習架構

本研究提出的模型架構如圖1，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，模型對於歌詞情緒分類的成效。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。

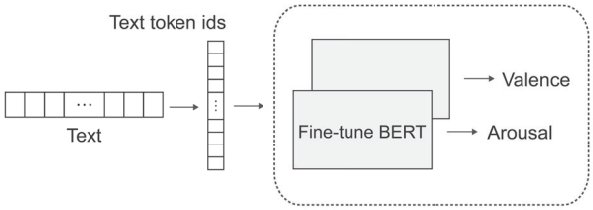

圖 1: 基於 BERT 的遷移式學習架構

3.2.1 資料預處理

CVAW、CVAP 和 CVAT 皆採用資料集內的文字、Valence 平均和 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 CVAW+CVAP 資料集，以 8 比 2 拆分爲訓練集跟測試集。BERT 模型有別於傳統文本的方法，會將標點符號視爲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 Valence 和 Arousal 分別標註爲正或負，以中性情緒爲原點，依照 Valence 和 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爲 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，爲避免產生過於稀疏向量，最大文本長度設定爲 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] 作爲文本的結束，每個中文字會對應到 BERT 中文字典的一個索引值稱爲 Token id，爲了讓每一則輸入序列的長度保持一致在文字序列後端填充特殊字元 [PAD]，最後，轉爲向量的序列和目標值轉爲 Tensor 至 BERT 模型進行訓練。

3.2.2 實施細節

本研究提出之模型架構如圖 1，模型輸出 Valence 和 Arousal 兩個數值，由於是數值的預測，因此損失函數選擇用均方誤差 (Mean square error，MSE)。實驗方法分別使用從 CVAW + CVAP 遷移至 CVAT 資料集的遷移學習方法比較從零直接訓練 CVAT 的未遷移的方法。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。在模型架構方面，在 BERT 欲訓練模型加上一層 Dropout 和一層線性分類層，優化器使用 Adam，學習速率在一開始嘗試多種學習速率，由於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，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，最大 Epoch 設定爲 100，並且加入 Early Stopping 的機制，將耐心 (Patience) 設至爲 10。

3.2.3 歌詞情緒之分類

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，對歌詞文本進行情緒的標註。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數值預測，模型輸出 Valence 值和 Arousal 值。依照原資料集的敘述，Valence 和 Arousal 都以中性值 5 爲閾值 (Yu et al. 2016;Yu et al. 2017)，因此，當模型輸出的 Valence 大於 5，表示模型預測該歌詞爲正向情緒並它標記爲 1、Valence 小於 5 則表示模型預測該歌詞爲負向情緒並標記爲 0，若 Arousal 值大於 5 表示模型預測該歌詞爲激動情緒並標記爲 1、Arousal 值小於 5 表示模型預測該歌詞爲平靜情緒並標記爲 0。我們將 Valence 和 Arousal 標記之後的結果轉爲 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

本章節將實驗結果分爲兩個階段，第一階段是模型訓練的結果，第二階段是驗證模型預測歌詞情緒的成效。

4.1 建立中文情緒模型

訓練模型的資料集切分皆以 8 比 2 進行，CVAP+CVAW 資料集的訓練集和測試集分別爲 6808 筆和 1702 筆。模型架構的訓練結果，如表1，模型的 Valence 和 Arousal 的均方誤差分別爲 0.3788 和 0.77339，且最佳的學習速率皆爲 1e-05。表2爲從零訓練 CVAT 資料集和從 CVAP+CVAW 資料集模型遷移至 CVAT 資料集的結果。本實驗模型架構是分別預測 Valence 和 Arousal 兩個數值，因此分別討論 Valence 和 Arousal 的結果。首先比較 Valence 輸出的結果，未經遷移的均方誤差爲 0.50338，經遷移學習的均方誤差爲 0.46624，

結果顯示經遷移學習的 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爲 1e-06，未經遷移的最佳學習速率爲 1e-5，就算同樣都在 1e-5 的學習速率下，經遷移學習的均方誤差 0.47898 還是優於未經遷移的均方誤差 0.50338。比較輸出爲 Arousal 的結果，經遷移的均方誤差爲 0.84259 優於未經遷移的 0.87107，從 Arousal 結果來看，經遷移學習的結果同樣優於未經遷移的結果。

Output	Learning Rate	Loss	Epoch
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

表 1: 在 CVAW+CVAP 資料集的訓練結果

Method	Output	Lr	Loss	Epoch
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
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

表 2: 經遷移學習與未經遷移學習之結果

4.2 驗證中文歌詞分類之結果

歌詞情緒分類是將模型輸出的 Valence 和 Arousal 基於中性值 5 轉換爲坐標平面上的四個象限類別。經遷移學習 CVAT 與未經遷移學習模型的歌詞情緒分類結果，如表 3，經遷移學習的 CVAT 模型在歌詞情緒分類的準確度爲 0.71，標籤 Q1 和 Q4 的 F1-score 較低，分別爲 0.69 和 0.51，而 Q2 和 Q3 的 F1-score 較高，分別爲 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爲 0.50，同樣是標籤 Q1 和 Q4 的 F1-score 較低，分別爲 0.41 和 0.29，而 Q2 和 Q3 的 F1-score 較高，分別爲 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

CVAT Transfer Learning			
Label	Precision	Recall	F1-score
Q1	0.96	0.53	**0.69**
Q2	0.72	0.98	**0.83**
Q3	0.63	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

表 3: 經遷移學習與未經遷移學習的歌詞分類結果之分數

都優於未經遷移學習的模型。由上述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能得到較佳的結果，證實經遷移學習的模型在 CVAW+CVAP 資料集中學習到的情緒特徵，有助於提升模型在歌詞文本的情緒辨識能力。

5 討論

從實驗結果可以看到 Arousal 的特徵較難學習其 loss 較高，在多個研究中都有提到中文或者英文的資料集上 Arousal 的維度難以區分，推測激動程度在文字上較難以顯示出來 (Malheiro et al. 2018; Yu et al. 2016;Çano and Morisio 2017b)。結果顯示經過遷移後的模型其結果都優於未經遷移的結果且提高了模型的收斂速度，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 中文情緒文本的學習。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。實驗結果表明了經遷移學習學到的情緒特徵是有助歌詞文本的情緒辨識成果。最後，在未學習過歌詞文本的狀況下，歌詞情緒分類結果達到 71% 的準確率。

6 結論

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 Valence 和 Arousal 進行標註。實驗比較了有遷移學習與未經遷移學習的模型，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，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。

致謝

特別感謝科技部給予研究經費的支持，本論文爲科技部計畫「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爲導向的詞曲生成器之設計」(108-2218-E-025-002-MY3) 之研究成果。

References

Jiddy Abdillah, I. Asror, and Y. Wibowo. 2020. Emotion classification of song lyrics using bidirectional lstm method with glove word representation weighting.

Yudhik Agrawal, Ramaguru Guru Ravi Shanker, and Vinoo Alluri. 2021. Transformer-based approach towards music emotion recognition from lyrics. arXiv:2101.02051.

James Barry. 2017. Sentiment analysis of online reviews using bag-of-words and lstm approaches. In *AICS*.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Byeong-Jun Han, Seungmin Rho, Roger B. Dannenberg, and Eenjun Hwang. 2009. Smers: Music emotion recognition using support vector regression. In *Proceedings of the 2017 International Conference on Intelligent Systems*, pages 651–656. ISMIR.

Qi Han, J. Guo, and Hinrich Schütze. 2013. Codex: Combining an svm classifier and character n-gram language models for sentiment analysis on twitter text. In *SemEval@NAACL-HLT*.

Xiao Hu and J. S. Downie. 2010a. When lyrics outperform audio for music mood classification: A feature analysis. In *ISMIR*.

Xiao Hu and J. Stephen Downie. 2010b. When lyrics outperform audio for music mood classification: A feature analysis. In *Proceedings of the 2017 International Conference on Intelligent Systems*, pages 619–624. ISMIR.

Yajie Hu, Xiaoou Chen, and Deshun Yang. 2009. Lyric-based song emotion detection with affective lexicon and fuzzy clustering method. In *ISMIR*.

Jason C. Hung and Jia-Wei Chang. 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. In *Applied Soft Computing*.

Adit Jamdar, Jessica Abraham, Karishma Khanna, and Rahul Dubey. 2015. Emotion analysis of songs based on lyrical and audio features. *Artificial Intelligence and Applications (IJAIA)*, arXiv:1506.05012.

Junghyun Kim, Seungjae Lee, Sungmin Kim, and Won young Yoo. 2011. Music mood classification model based on arousal-valence values. *13th International Conference on Advanced Communication Technology (ICACT2011)*, pages 292–295.

C. Laurier, M. Sordo, J. Serrà, and P. Herrera. 2009. Music mood representations from social tags. In *ISMIR*.

R. Malheiro, R. Panda, Paulo Gomes, and R. Paiva. 2018. Emotionally-relevant features for classification and regression of music lyrics. *IEEE Transactions on Affective Computing*, 9:240–254.

Lin Qiu, Jiayu Chen, Jonathan Ramsay, and Jiahui Lu. 2019. Personality predicts words in favorite songs. *Research in Personality*, 78:25–35.

J. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39:1161–1178.

Uddalok Sarkar, Sayan Nag, Medha Basu, Archi Banerjee, Shankha Sanyal, Ranjan Sengupta, and Dipak Ghosh. 2015. Neural network architectures to classify emotions in indian classical musics. arXiv:2102.00616.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

L. Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Y. He, Jun Hu, K. Lai, and Xue-Jie Zhang. 2016. Building chinese affective resources in valence-arousal dimensions. In *HLT-NAACL*.

L. Yu, Lung-Hao Lee, Jin Wang, and Kam-Fai Wong. 2017. Ijcnlp-2017 task 2: Dimensional sentiment analysis for chinese phrases. In *IJCNLP*.

Erion Çano and M. Morisio. 2017a. Music mood dataset creation based on last.fm tags.

Erion Çano and Maurizio Morisio. 2017b. Moodylyrics: A sentiment annotated lyrics dataset. In *Proceedings of the 2017 International Conference on Intelligent Systems,*, pages 118–124, Hong Kong. Metaheuristics Swarm Intelligence.