NCU-NLP at ROCLING-2021 Shared Task: Using MacBERT Transformers for Dimensional Sentiment Analysis

洪滿珍 Man-Chen Hung, 陳昭沂 Chao-Yi Chen
陳品蓉 Pin-Jung Chen, 李龍豪 Lung-Hao Lee
國立中央大學電機工程學系
Department of Electrical Engineering
National Central University
{109521068, 107501543, 107303034}@ncu.edu.tw, lhlee@ee.ncu.edu.tw

摘要

我們運用 MacBERT 模型在 CVAT 與 CVAS 資料集微調使其適用於 ROCLING 2021 的評測任務，並比較 MacBERT 與 BERT 和 RoBERTa 這兩個不同的模型，在 Valence 與 Arousal 維度上的效能差異。我們以平均絕對誤差 (MAE) 與關係係數 (r) 作為評分指標，在測試資料上能夠在 Valence 達到 MAE 與 r 分別為 0.611 與 0.904；而在 Arousal 達到 MAE 與 r 分別為 0.938 與 0.549 的效能。

Abstract

We use the MacBERT transformers and fine-tune them to ROCLING-2021 shared tasks using the CVAT and CVAS data. We compare the performance of MacBERT with the other two transformers BERT and RoBERTa in the valence and arousal dimensions, respectively. MAE and correlation coefficient (r) were used as evaluation metrics. On ROCLING-2021 test set, our used MacBERT model achieves 0.611 of MAE and 0.904 of r in the valence dimensions; and 0.938 of MAE and 0.549 of r in the arousal dimension.

關鍵字：情感運算、學習情緒、深度學習
Keywords: affective computing, learning emotions, deep learning

1 介紹

情感運算 (affective computing) 的目標是希望機器能夠讀懂人類的情感，進而做出相對應的動作，情感分析(sentiment analysis) 是自然語言處理研究中重要的研究領域，主要在於如何有效的提取文本中的情感資訊。根據情感的表達形式不同，可以區分為以下兩種方式：分類型情感分析或是維度型情感分析 (Calvo and Kim, 2013)。

傳統作法採用分類型 (categorical) 情感分析，將所有情感詞分成某幾個類別，例如：常見的正面、中立、負面三類情緒； 以及 Ekman (1992) 的六個基本情緒包含：憤怒 (anger)、高興 (happiness)、恐懼 (fear)、悲傷 (sadness)、厭惡 (disgust) 和驚喜 (surprise)，廣泛地應用各個領域 (Schouten and Frasincar, 2015; Pontiki et al., 2015)。

由於分類型情感分析無法精準的表達情緒詞中所帶有的情感，因此衍生出維度型 (dimensional) 情感分析，在多個情緒維度上，用連續性數值表示情感 (Russell, 1980)。圖 1 為常見的 Valance-Arousal 情感維度，採用二維平面表示，橫軸為 Valance 軸 (簡稱 V) 代表情感的正負面，數值範圍是 1 到 9 分，數值 1 表示最負面，數值 9 表示最正面，數值 5 則是中立沒有偏向。縱軸為 Arousal 軸 (簡稱 A) 表示情感激動程度， 1 為最平靜、9 為最激動，5 為中間值。任何字詞、片語、句子、篇章、段落都可以這個 VA 二維平面上表示，例如：「感激」這個詞在中文維度型情感字典 (CVAW) (Yu et al., 2016a) 的 VA 值分別為 6.8 和 7.2，位於 VA 平面屬於情緒正面且激動程度高的第一象限；「冷漠」一詞 (V: 4.0, A: 2.8) 則被歸到情緒負面且激動程度低的第三象限。任何情緒都可以在這個 VA 二維平面上用連續的

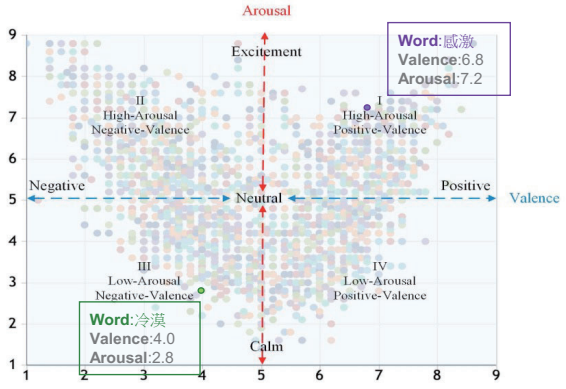

圖 1：VA 二維平面

數值表示，能更加細微的表達情感的差異，擴大應用領域。

IALP 2016 評測任務的目標是將此維度型分析應用在中文字詞 (Chinese words) (Yu et al; 2016b)；IJCNLP 2017 評測任務則是中文片語 (Chinese phrases)，近一步探討修飾語對於情感詞的情感狀態值變化 (Yu et al., 2017)。ROCLING 2021 今年的評測任務則專注於領域自調適(domain adaption)問題，目標是教育領域的學生學習心得情緒分析。在這個任務中，輸入一則學生的中文學習心得，情感分析系統需要分別對 Valence (V) 和 Arousal (A) 兩個維度，分析該輸入句蘊含的情感，各自輸出 1 到 9 的實數值，以下為編號 1 的中文心得，系統輸出值分別是 V 值 6.8 及 A 值 5.2。

- Input: 1, 今天教了許多以前沒有學過的東西，所以上起課來很新鮮。
- Output: 1, 6.8, 5.2

我們開發的 NCU-NLP 情感分析系統採用 MacBERT 作為主要的模型架構 (Cui et al., 2020)。我們用 CVAT 資料集 (Yu et al., 2016b) 以及自行建置的 CVAS 資料集，微調預訓練好的 MacBERT 模型，在測試資料上在 Valence 的 MAE 與 r 分別達到 0.611 與 0.904；在 Arousal 的 MAE 與 r 分別達到 0.938 與 0.549。

本文其他章節如下，第二章是維度型情感分析的相關研究，第三章敘述我們使用的模型架構，第四部份為實驗結果和效能比較，最後則是結論。

2 相關研究

維度型情感分析 VA 值預測模型，大致可分成三大類：辭典法 (lexicon-based)、回歸法 (regression-based)、以及神經網路法 (neural-network-based)。

辭典法將句子中出現的字詞，與情感字典進行比對，最後平均所有情感詞的 VA 值，將此平均結果當成文本的 VA 值預測結果 (Paltoglou et al., 2012)。

回歸法曾是最常被研究拿來預測 VA 值的方法，為了解決中文與英文這種跨語言的文字問題，Wei et al. (2011) 提出半自動標記中文詞語的方式，以英文為底，轉換成中文，最後才進行回歸模型的訓練，這種方式容易造成擬合度不足 (underfitting)。因此，Wanget al. (2016b) 提出區域性加權回歸方式 (a locally weighted method)，可以提升預測精準度。Wang et al. (2016a) 進一步提出 Community-based 加權圖模型 (weighted graph model)，能讓未見過的詞，有更相似的 Seed 字詞，對 VA 估計過程更有幫助，對英文和中文數據集都有更好的預測效能。此外，Amir et al. (2015) 同樣利用回歸法，並加上詞嵌入向量分析推特上的情緒成分。

近年來，深度學習方法廣泛用於維度型情感分析，有許多研究朝向使用詞嵌入向量和類神經網路的方式來預測 VA 情感值。在 IALP 2016 評測任務中，Du and Zhang (2016) 採用集成式單層增強神經網路 (an ensemble of several boosted one layer neural networks) 完成 VA 情感值的預測。在 IJCNLP 2017 的評測任務中，Wu et al. (2017) 中提出連結緊密的長短期記憶模型 (Long Short-Term Memory, LSTM) 預測中文詞語及片語。Yu et al. (2020) 採用兩層 NN 模型的疊接，第一層決定單詞的強度，第二層決定修飾詞語的位移權重，將有助於提升片語的精確度。Zhu et al. (2019) 使用基於注意力的對抗神經網路 (Adversarial Attention Network)，著重訓練為輸入詞加權的注意力層 (attention layer)，以達到確定某些詞語為特定情感做出的貢獻。之後更提出結合 CNN 和 LSTM 模型兩種類神經網路模型，將文本一部份當作 CNN 模型的輸入，以提取特徵值，而產生樹狀區域的 CNN-LSTM 模型，實驗結果

顯示比以前單純的類神經網路結構，效果更好 (Wang et al., 2020)。

學生的學習表現都會呈現在成績上，結構化的資料比如說是出席率、作業完成率及繳交率、課堂上的參與程度等等，都被老師來評斷成學生的學習狀況。學生學習心得這種非結構化的資料，因為不易結構化很容易被忽略掉，學生上課時的心得通常都富含情感詞，內容大多是對於課堂吸收程度的第一手情感資料，如果可以有效分析，將非常有助於課堂的調整，或是讓老師了解那些學生需要加強，對於教育領域很有幫助(Yu et al., 2018)。

3 NCU-NLP 模型架構

我們所使用的模型為 MacBERT (MLM as correction BERT) (Cui et al., 2020)，並對模型進行微調。MacBERT 是基於 BERT 改良的模型，該模型與 BERT 共享相同的預訓練任務，並對遮罩語言模型 (Masked Language Model, MLM) 任務進行修改：使用全詞遮罩及 n-gram 遮罩策略來選擇遮罩的詞候選，詞級別 1-gram 到 4-gram 的遮罩比例為 40%、30%、20%、10%。此外，MacBERT 不使用 [MASK] 符號來進行遮罩，因為在詞與符號的微調階段並沒有出現過 [MASK]，而改採用相似單詞進行遮罩，使用基於同義詞工具 (synonyms toolkit) 與 Word2vec (Mikolov et al., 2013)來計算相似度。若選擇一個 n-gram 來進行遮罩，則將會分別尋找相似的單詞，在找不到相似單詞時，則會降級使用隨機的單詞替換。MacBERT 使用 15% 的輸入文字來進行遮罩，80% 替換為相似的單詞，10% 替換為隨機單詞，剩下的 10% 為保留原本的輸入單詞。

4 實驗與評估

4.1 資料集

訓練資料來自中文維度型情感語料庫 (Chinese Valence-Arousal Text, CVAT) (Yu et al., 2016b) 及自行建置的 CVAS (Chinese Valence-Arousal Sentences)資料集。CVAT 資料集 (ver. 2.0) 內包含 2,969 個中文評論段落(Yu et al., 2016a)，CVAS 資料集包含 2582 個中文情感句子，皆已標記 VA 值。測試資料為主辦單位提供的學生中文學習心得，共有 1600 句。

4.2 實驗設定

我們比較以下三個模型的效能差異：BERT (Devlin et al., 2019)、RoBERTa (Liu et al.,2019) 與 MacBERT (Cui et al., 2020)。實驗採五折交互驗證，學習率 (learning rate) 設定為 5e-5，批次大小為 64，以及訓練次數 (epoch) 為 50 次。此外，我們比較只有 CVAT 資料集作為訓練資料，以及 CVAT 資料集加入 CVAS 資料集的效能差異。

4.3 評分指標

模型的表現程度，將情緒正負面 (V 值) 和激動程度 (A 值) 分開計算，以模型預測結果和標準答案間作比對，使用平均絕對誤差 (Mean absolute error, MAE) 及皮爾森相關係數 (Pearson correlation coefficient)來衡量。

- 平均絕對誤差 (Mean absolute error, MAE)

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|A_i - P_i| \qquad (1)$$

- 皮爾森相關係數 (Pearson correlation coefficient)

$$\mathrm{r} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{A_i - \bar{A}}{\sigma_A}\right)\left(\frac{P_i - \bar{P}}{\sigma_P}\right) \qquad (2)$$

4.4 結果

只有 CVAT 資料集的交互驗證結果如表 1，MacBERT 在 Valence 的 MAE 與 r 分數都較其他兩個模型來得更好；而在 Arousal 的 MAE 比其他兩者來的好，但 r 與 RoBERTa 有些微的落差。整體而言，MacBERT 的表現比 RoBERTa 與 BERT 佳。

表 1 與表 2 為 CVAT 與 CVAS 兩個資料集的交互驗證結果，MacBERT 與 RoBERTa 的 MAE 和 r 在 Valence 與 Arousal 都只有些微的差距，但都比 BERT 還要來得好。

綜合上述結果，我們決定採用MacBERT 作為系統架構，選擇兩種不同資料組合 (CVAT, CVAT+CVAS) 微調後的模型，作為 NCU-NLP 系統在測試集的效能。

CVAT		
Valence	MAE	r
BERT	0.475	0.854
RoBERTa	0.469	0.895
MacBERT	0.457	0.897
Arousal	MAE	r
BERT	0.668	0.62
RoBERTa	0.659	0.695
MacBERT	0.652	0.639

表 1、CVAT 資料集實驗結果

CVAT+CVAS		
Valence	MAE	r
BERT	0.531	0.854
RoBERTa	0.51	0.868
MacBERT	0.513	0.865
Arousal	MAE	r
BERT	0.763	0.582
RoBERTa	0.757	0.596
MacBERT	0.754	0.592

表 2、CVAT＋CVAS 資料集實驗結果

4.5 比較

表 3 為測試結果。在 Valence 上，CVAT+CVAS 資料集微調過的 MacBERT 模型的 MAE 分數，比只有 CVAT 資料集的少了 0.014，且 r 的分數也只高了 0.004。在 Arousal 上，只有 CVAT 資料集的 MAE 分數，比 CVAT+CVAS 資料集的分數少了 0.051，但 CVAT+CVAS 資料集的 r 分數，比只有 CVAT 資料集的分數還要好 0.033。

整體來說，以相關係數 r 來看，整體上 CVAT+CVAS 資料集的分數在 Valence 與 Arousal 上都比只有 CVAT 的來得高，而 MAE 在 Valence 上 CVAT+CVAS 資料集的分數較好，Arousal 則是只有 CVAT 資料集的分數較佳。

5 結論

在本次的評測任務中，經由 CVAT 資料集微調後的 MacBERT，在 Valence 的 MAE 與 r 分別為 0.625 與 0.9；在 Arousal 的 MAE 與 r 分別為 0.938 與 0.549。而經由 CVAT+CVAS 資料集微調後的 MacBERT ，在 Valence 的 MAE 與 r 分別為 0.611 與 0.904；在 Arousal 的 MAE 與 r 分別為 0.989 與 0.582，且在 r 指標都得到較好的成績。

ROCLING-2021 Test Set		
Valence	MAE	r
MacBERT - CVAT	0.625	0.9
MacBERT - CVAT +CVAS	0.611	0.904
Arousal	MAE	r
MacBERT - CVAT	0.938	0.549
MacBERT - CVAT +CVAS	0.989	0.582

表 3、ROCLING-2021 Test Set 實驗結果

致謝

This work was partially supported by the Ministry of Science and Technology, Taiwan under the grant MOST 108-2218-E-008-017-MY3.

參考資料

Silvio Amir, Ranmon F. Astudillo, Wang Ling, Bruno Martins, Mario Silva, and Isabel Trancoso. 2015. INESC-ID: A regression model for large scale twitter sentiment lexicon induction. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 613-618. https://doi.org/10.18653/v1/S15-2102

Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: dimensional and categorical models. *Computational Intelligence*, 29(3):527-543. https://doi.org/10.1111/j.1467-8640.2012.00456.x

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, Guoping Hu. 2020. Revisiting Pre-Trained Models for Chinese Natural Language Processing. *Association for Computational Linguistics,* pages 657–668.https://doi.org/10.18653/v1/2020.findings-emnlp.58.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Human Language Technologies, pages 4171-4186.* https://doi.org/10.18653/v1/N19-1423.

Steven Du and Xi Zhang. 2016. Aicyber's system for IALP 2016 shared task: Character-enhanced word vectors and Boosted Neural Networks. In *Proceedings of 2016 International Conference on Asian Language Processing,* pages 161-163. http://doi.org/10.1109/IALP.2016.7875958

Paul Ekman. 1992. An argument for basic emotions. *Cognition and Emotion*, 6(3-4):169-200. https://doi.org/10.1080/02699939208411068

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26,* pages 3111–3119. https://proceedings.neurips.cc/paper/2013/file/9aa42b31882ec039965f3c4923ce901b-Paper.pdf.

Georgios Paltoglou, Mathias Theunis, Arvid Kappas, and Mike Thelwall. 2012. Predicting emotional responses to long informal text. *IEEE Transactions on Affective Computing*, 4(1):106-115. http://doi.org/2010.1109/T-AFFC.2012.26

Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. SemEval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, pages 486-495. https://doi.org/10.18653/v1/S15-2082

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

Kim Schouten and Flavius Frasincar. 2015. Survey on aspect-level sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering*, 28(3):813-830. https://doi.org/10.1109/TKDE.2015.2485209

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016a. Community-based weighted graph model for valence-arousal prediction of affective words. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):1957-1968. https://doi.org/10.1109/TASLP.2016.2594287

Jin Wang,Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016b. Locally weighted linear regression for cross-lingual valence-arousal prediction of affective words. *Neurocomputing*, 194:271-278. https://doi.org/10.1016/j.neucom.2016.02.057

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2020. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing,* 28:581-591. http:/doi.org/10.1109/TASLP.2019.2959251

Wen-Li Wei, Chung-Hsien Wu, and Jen-Chun Lin. 2011. A regression approach to affective rating of Chinese words from ANEW. In *Proceedings of International Conference on Affective Computing and Intelligent Interaction,* pages 121-131. https://doi.org/10.1007/978-3-642-24571-8_13

Chuhan Wu, Fangzhao Wu, Yongfeng Huang, Sixing Wu, and Zhigang Yuan. 2017. THU_NGN at IJCNLP-2017 task 2: Dimensional sentiment analysis for Chinese phrases with deep LSTM. In *Proceedings of the 8th International Joint Conference on Natural Language Processing: Shared Tasks*, pages 47-52.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. https://arxiv.org/pdf/1907.11692.pdf

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, Xuejie Zhang 2016a. 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*, pages 540-545 https://doi.org/10.18653/v1/N16-1066

Liang-Chih Yu, Lung-Hao Lee, and Kam-Fai Wong. 2016b. Overview of the IALP 2016 shared task on dimensional sentiment analysis for Chinese words. In *International Conference on Asian Language Processing*, pages 156-160. http://doi.org/10.1109/IALP.2016.7875957

Liang-Chih Yu, Lung-Hao Lee, Jin Wang, and Kam-Fai Wong. 2017. IJCNLP-2017 Task 2: Dimensional Sentiment Analysis for Chinese Phrases. In *Proceedings of the 8th International Joint Conference on Natural Language Processing: Shared Tasks*, pages 9-16.

Liang-Chih Yu, Jin Wang, K. Robert Lai, and Xuejie Zhang. 2020. Pipelined neural networks for phrase-level sentiment intensity prediction. *IEEE Transactions on Affective Computing,* 11(3): 447-458. http://doi.org/10.1109/TAFFC.2018.2807819

Liang-Chih Yu, C.-W. Lee, H.I. Pan, C.Y. Chou, P.Y. Chao, Z.H. Chen, S.F. Tseng, C.L. Chan, and K.R. Lai. 2018. Improving early prediction of academic failure using sentiment analysis on self-evaluated comments. *Journal of Computer Assisted Learning*, 34(4):358-365. https://doi.org/10.1111/jcal.12247

Suyang Zhu, Shoushan Li, and Guodong Zhou. 2019. Adversarial attention modeling for multi-dimensional emotion regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 417-480. http://doi.org/10.18653/v1/P19-1045