ntust-nlp-1 at ROCLING-2021 Shared Task: Educational Texts Dimensional Sentiment Analysis using Pretrained Language Models

王繹崴
Yi-Wei Wang
國立台灣科技大學
Nation Taiwan University of Science and Technology

張維哲
Wei-Zhe Chang
國立台灣科技大學
Nation Taiwan University of Science and Technology

方柏翰
Bo-Han Fang
國立台灣科技大學
Nation Taiwan University of Science and Technology

陳奕嘉
Yi-Chia Chen
國立台灣科技大學
Nation Taiwan University of Science and Technology

黃偉愷
Wei-Kai Huang
國立台灣科技大學
Nation Taiwan University of Science and Technology

陳冠宇
Kuan-Yu Chen
國立台灣科技大學
Nation Taiwan University of Science and Technology

摘要

本研究為 Rocling 2021 共同任務：教育文本的維度式情感分析之成果報告。為了分析中文文本的情緒效價(Valence)與喚起程度(Arousal)，本研究基於當前流行的預訓練語言模型 BERT 與近期基於全詞遮蔽(Whole Word Masking)進行預訓練的 MacBERT，觀察模型在不同設定下的預測成果，並比較 BERT 與 MacBERT 在中文文本情緒預測效能的差異。我們發現，相較於 BERT，MacBERT 可以在驗證集上獲得些許的效能提升。因此，我們將數個使用不同訓練方法所得的預測模型進行預測結果平均，作為最終的輸出。

Abstract

This technical report aims at the ROCLING 2021 Shared Task: Dimensional Sentiment Analysis for Educational Texts. In order to predict the affective states of Chinese educational texts, we present a practical framework by employing pre-trained language models, such as BERT and MacBERT. Several valuable observations and analyses can be drawn from a series of experiments. From the results, we find that MacBERT-based methods can deliver better results than BERT-based methods on the verification set. Therefore, we average the prediction results of several models obtained using different settings as the final output.

關鍵字：情感分析、預訓練語言模型、BERT、MacBERT
Keywords: Sentiment Analysis, Pre-trained Language Model, BERT, MacBERT

1 緒論 (Introduction)

情緒分析已經是自然語言處理中備受矚目的任務之一，屬於文本分類的子任務，目標在於面對不同的文本時，能夠辨識出文本所欲表達的各類情緒量值，比如：正面、負面、情緒高漲、情緒低落等(Wei et al., 2011; Malandrakis et al., 2013; Wang et al., 2016; Du and Zhang, 2016; Wu et la., 2017; Yu et al., 2020, Kim et al., 2010; Paltoglou et al, 2013; Goel et la., 2017; Zhu et al., 2019; Wang et al., 2019; 2020)。情緒辨識可以廣泛地應用在我們的生活中，比如：分析網路上的社群評論、售後產品的相關回饋、客服機器人的應答等。

此次的共同任務：教育文本的維度式情感分析，其目標在於分析出中文教育文本中的喚起程度(Arousal)以及效價程度(Valence)(Russell, 1980)，其中喚起程度的高低意味著語者是興奮或是平靜，而效價程度則是代表

語者自身處於積極或是消極的態度(Patricia E. G. Bestelmeyer, 2017)。預訓練語言模型，能夠為文本萃取出含有豐富語意資訊的特徵向量，而這個特徵向量可被應用於其它下游任務中，完成各式不同的自然語言處理任務。因此在本次任務中，我們將會使用 BERT (Jac ob Devlin, 2018)與 MacBERT (Yiming Cui, 2020)作為文件的編碼器，透過編碼器產生具有語意資訊的特徵向量，接著將其輸入至下游任務的模型內，讓模型在預測喚起程度以及效價程度時，可以獲得更佳精確的結果。

2 研究背景 (Research Background)

在本章節中，我們將會介紹在自然語言處理任務中，處理文本資料時常會使用到的重要技術-詞嵌入。此外，由於本次任務為分數預測，屬於回歸任務的一種，因此在本章節中也會對回歸任務進行介紹。最後，在本章節的末段，我們將會介紹近年來在各項自然語言處理任務中大放異彩的預訓練語言模型-BERT 及其衍生模型 MacBERT。

2.1 詞嵌入 (Word Embedding)

詞嵌入的核心概念為「將一個單詞透過一個向量進行表示」。近年來較為熟悉的相關研究有 word2vec (Tomas Mikolov, 2013)、fast-text (Armand Joulin, 2016)和 Glove (Jeffrey Pennington, 2014) 。上述的這些研究都使用各自的方法將文字表達成向量，也成功地在很多自然語言處理任務上達到優秀的成果。不過這些詞嵌入的方法在「相同單詞但是不同語意」的時候，其表示向量卻是一樣的，為了解決此一問題，後續研究提出各式「動態」的詞嵌入表示法，比如：Cove (Bryan McCann, 2017)、ELMo (Matthew E. Peters, 2018)和 BERT (Jacob Devlin, 2018)。這種類型的詞嵌入方法會透過一個語言模型，將輸入文本根據其內容的語意，給予每一個詞一個基於上下文的詞嵌入表示向量(Enkhbold Bataa, 2019)。

2.2 回歸任務 (Regression)

回歸任務是讓機器根據訓練集的資料，學習如何為輸入的資料抽取特徵，並利用這些特徵資訊，轉換成正確的標記數值。本次的任務是分析輸入文本的喚起程度(Arousal)以及效價程度(Valence)，因此我們將這個任務視為一個回歸任務。

2.3 BERT

BERT 為 Bidirectional Encoder Representation from Transformer 的簡稱，為相當經典的預訓練語言模型，其架構為多層的 Bidirectional Transformer 層，而 BERT 在訓練上分為預訓練(Pre-training)與微調(fine-tuning)兩個步驟。在預訓練步驟裡，會使用大量的無標記文本來訓練 BERT 模型，而訓練方式則包括遮罩語言模型(Masked Language Model)以及下一句預測任務(Next Sentence Prediction)。在遮罩語言模型的任務中，會有一部分的字符(token)隨機的被遮罩或是替換成類似的字符，而模型必須去預測遮罩處的正確字符為何。下一句預測則是讓模型去判斷兩個連續的句子，後一句是否確實是接在前一句之後。在微調階段，模型將被訓練於解決目標任務。相較於預訓練，模型微調使用少量的標記資料，來對模型參數進行調整，使其得以符合下游任務的需求。

2.4 MacBERT

MacBERT(MLM as correlation BERT)是一個特別針對中文語言處理所設計的中文預訓練模型，跟 BERT 不同的地方在於：

- MacBERT 在填空部分使用全詞遮蔽(Whole Word Masking)，也就是在進行遮罩的時候，是以詞為遮罩單位而非單一個字符(token)為單位，避免一些連貫性很強的字符序列，就算被遮罩一部分，模型仍可輕易地預測出被遮罩的部分。

- 在遮罩方式上，追加使用 N 元遮罩(N-gram Masking)以及 Mac 遮罩(Mac-Masking)。N 元遮罩即將連續的 N 個字符一起遮罩；Mac 遮罩則是將所有被遮罩的字符都以向量上相近的字符作為替代，而非單純的$<mask>$符號，這是考慮到$<mask>$是不會出現在下游任務的。

- MacBERT 並非選擇使用預測下一句作為預訓練的任務，而是以語句順序判斷(Sentence Order Prediction)作為訓練目標。在語句順序判斷的任務中，模型必須辨識出兩句連續句子之間的先後關係。

以上三種與 BERT 不同的預訓練方式，使得 MacBERT 更能夠彌補預訓練階段與下游任務的差異性，也使得 MacBERT 在不同任務的中文資料集上都能夠得到比 BERT 還要優秀的成績。

3 方法 (Methods)

有鑑於近年來人工智慧、深度學習相關技術的蓬勃發展，尤其是 BERT 及其衍生模型在各式自然語言處理相關任務上大放異彩，刷新了各項成績。因此，在本研究中，我們將使用 BERT 及其衍生模型 MacBERT 進行後續實驗與討論，探究預訓練語言模型在中文教育文本的維度式情感分析任務上的成效。

3.1 模型架構 (Model Architectures)

圖 1 為本研究所使用之模型架構。在模型的輸入方面，我們在字符序列的最前面加上一個特別的字符$[CLS]$，在後續的分數預測時，我們則將這個特別字符的向量輸入至全連接層，分別得到輸入句子所對應之情緒效價與喚起程度的預測分數。

3.2 模型訓練 (Model Training)

由於本次任務屬於回歸任務，因此在誤差函式的設計方面，我們使用均方誤差(Mean Square Error, MSE)計算模型預測出的情緒效價分數V_i、喚起程度分數A_i，與正確答案$\widehat{V_i}$與$\widehat{A_i}$的誤差，並透過此誤差來優化模型的參數，完成模型的訓練：

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(V_i - \widehat{V_i})^2 + (A_i - \widehat{A_i})^2 \quad (1)$$

在模型訓練方面，我們使用多種方法來對模型進行訓練，方法包括將多個模型參數進行平均、多個模型預測結果平均、使用模型預測的結果當作虛擬標籤，並將虛擬標籤資料與原始訓練資料結合，進行二次訓練等方法。我們將於第四章節中詳細描述各個訓練方法的實作細節，並且比較各個方法所訓練出來的模型在驗證資料及測試資料上的效能表現。

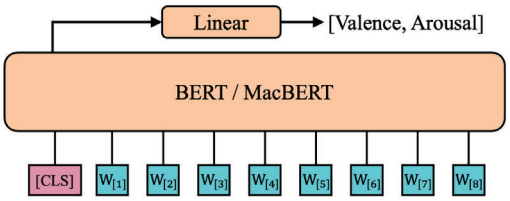

圖 1. 模型架構圖。

4 實驗 (Experiments)

在實驗的部分，我們使用了 5 種不同的設定方法，訓練出了 5 個子模型。而最終的輸出，則是這 5 個子模型的預測數值之平均。在本節中，我們將展示實作細節以及在測試集上模型集成（Ensemble）的結果。

4.1 訓練資料 (Training Data)

我們將 CVAW 4.0 的 5,512 個詞及 CVAP 2.0 的 2,998 個片語以及從 CVAT 2.0 的 2,969 個句子中抽取出 80%（2,375 筆）的句子合併作為訓練集；CVAT 2.0 剩餘的 20%（594 筆）個句子作為驗證集。經上述處理後，訓練集共有 10,885 筆文件，驗證集則有 594 筆資料。

4.2 子模型 (Sub-models)

各式子模型架構皆如表 1 所示，我們採用 BERT 或 MacBERT 作為基礎，藉由不同的訓練方法與設定，產生六個不同的子模型。

- **方法 1**：使用 BERT-base[1]作為基礎模型，採用 Adam 做為模型優化器，共迭代訓練 20 次，並使用 Noam(Ashish Vaswani, 2017)學習率調整器，再將 warmup_steps 設定為 25,000 來調整訓練時的學習率。最終，我們將訓練過程中，在驗證集上誤差最低的 5 個模型參數進行平均，作為最終的模型參數。

- **方法 2**：與方法 1 相同，只是基礎模型用 MacBERT-base[2]。

- **方法 3**：與方法 2 一樣使用 MacBERT-base 作為基礎模型，選擇 SGD 作為優化器，並迭代 5 次，而學習率固定為 1e-3。之後，我們額外加入 dianping[3] 資料集做

[1] https://huggingface.co/bert-base-chinese
[2] https://huggingface.co/hfl/chinese-macbert-base
[3] https://github.com/zhangxiangxiao/glyph

	Mean Absolute Error		Pearson Correlation Coefficient	
Sub-models	**Valence**	**Arousal**	**Valence**	**Arousal**
方法 1	0.463	0.614	0.890	0.674
方法 2	0.442	0.634	0.895	0.659
方法 3	0.514	0.679	0.880	0.624
方法 4	0.487	0.649	0.885	0.664
方法 5	0.469	0.623	0.888	0.667
方法 6	0.477	0.662	0.900	0.637

表 1：驗證集結果

	Mean Absolute Error		Pearson Correlation Coefficient	
	Valence	**Arousal**	**Valence**	**Arousal**
方法 1	0.586	0.885	0.901	0.585
集成模型	0.684	0.906	0.912	0.607
CYUT-run1	1.695	1.177	-0.017	0.040
CYUT-run2	1.685	1.252	0.007	-0.021
NCU-NLP-run1	0.625	0.938	0.900	0.549
NCU-NLP-run2	0.611	0.989	0.904	0.582
ntust-nlp-2-run1	0.654	0.880	0.905	0.581
ntust-nlp-2-run2	0.667	0.866	0.913	0.616
SCUDS-run1	0.953	1.054	0.694	0.375
SCUDS-run2	0.975	1.039	0.667	0.354
SoochowDS-run1	2.421	1.327	0.073	0.051
SoochowDS-run2	1.073	1.125	0.584	0.228

表 2：測試集結果

偽標籤（Pseudo Labeling），迭代 5 次，學習率固定為 1e-4。

- **方法 4：**與方法 2 一樣使用 MacBERT-base 作為基礎模型，迭代 20 次後，選出在驗證集上誤差最低的 4 個模型，我們將這四個模型的輸出結果取平均，作為最終的輸出。與方法 2 不同的是，方法 4 是對輸出結果做平均，而方法 2 是對模型參數做平均。

- **方法 5：**使用 BERT-base-uncased[4]、RoBERTa-wwm-ext[5]、MacBERT-base 作為基礎模型，三種基礎模型皆使用 Adam 優化器，學習率 2e-5，各自迭代 3 次後，選出在驗證集上誤差最低的模型參數。我們將 BERT、RoBERTa 與 MacBERT 輸出的結果取平均，作為最終的輸出。

- **方法 6：**使用 MacBERT-large[6]作為基礎模型，採用 SGD 為優化器迭代 12 次，學習率固定為 1e-4。我們將訓練集透過 word 軟體分別翻譯成英文、法文、德文、日文、俄語、義大利文後再翻譯回中文，因此相較於其他方法，方法 6 的訓練資料量擴增至原本的 7 倍。

4.3 實驗結果 (Experimental Results)

表 1 為各子模型在驗證集上的實驗結果。由於共同任務最終僅能繳交兩組系統，因此我們保留方法 1，作為一組系統；此外，我們將方法 2 至方法 6 的預測結果取平均，作為一個集成系統，當成第二組輸出。表 2 為方法 1 以及集成模型在測試集上的結果。

從表 2 中的數據可以發現，集成模型的均方誤差比方法 1 還要高，我們推測原因可能來自於：

[4] https://huggingface.co/bert-base-uncased
[5] https://huggingface.co/hfl/chinese-roberta-wwm-ext
[6] https://huggingface.co/hfl/chinese-macbert-large

- 方法 2 至方法 6 所訓練出來的子模型效能表現參差不一，觀察表 1，除了方法 2 之外，其餘方法的結果皆明顯比方法 1 差，因此即便進行預測結果整合，也無法彌補模型效能上的缺陷，導致預測結果不盡理想。

- 經觀察測試資料後，我們發現測試資料中的句子與訓練資料中的句子形式上有所差異。測試資料中的句子長度普遍較短，且內容相較於訓練資料差異較大。因此我們認為另一個照成集成模型效能較差的原因是集成模型的預測結果過於 overfitting 在訓練資料上，因此在測試資料上的預測表現不是很好。

5 結論 (Conclusions)

在 Rocling 2021 共同任務：教育文本的維度式情感分析的任務中，我們的方法 1 在情緒效價與喚起程度的均方誤差分別為 0.586 與 0.885，他們的皮爾森相關係數則分別為 0.901 與 0.585。與其他隊伍相較，我們成功取得本次共同任務裡最低的情緒效價均方誤差。因而證實我們所提出的方法能在教育文本的維度式情感分析的任務上擁有較好的效能表現。

Acknowledgment

This work was supported by the Ministry of Science and Technology of Taiwan under Grant MOST 110-2636-E-011-003 (Young Scholar Fellowship Program). We thank to National Center for High-performance Computing (NCHC) for providing computational and storage resources.

References

Patricia E. G. Bestelmeyer, Sonja A. Kotz, and Pascal Belin. 2017. *Effects of emotional valence and arousal on the voice perception network https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5597854/pdf/nsx059.pdf*

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. *Revisiting pre-trained models for chinese natural language processing*. In Findings of EMNLP. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *Bert: Pre-training of deep bidirectional transformers for language understanding*. In arXiv preprint arXiv: 1810.04805v2

Tomas Mikolov, Kai Chen, Greg Corrado and Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space.* In arXiv preprint arXiv: 1301.3781v3

Armand Joulinm Edouard Grave, Piotr Bojanowski and Tomas Mikolov. 2016. *Bag of Tricks for Efficient Text Classification.* In arXiv preprint arXiv:1607.01759v3

Jeffrey Pennington, Richard Socher and Christopher D. Manning. 2014. *GloVe: Global Vectors for Word Representation.* https://aclanthology.org/D14-1162.pdf

Bryan McCann, James Bradbury, Caiming Xiong and Richard Socher. 2017. *Learned in Translation: Contextualized Word Vectors* https://papers.nips.cc/paper/2017/file/20c86a628232a67e7bd46f76fba7ce12-Paper.pdf

Matthew E. Peters† , Mark Neumann† , Mohit Iyyer† , Matt Gardner†, Christopher Clark* , Kenton Lee* and Luke Zettlemoyer†*. 2018. *Deep contextualized word representations.* In arXiv preprint arXiv:1802.05365v2

Enkhbold Bataa and Joshua Wu. 2019. *An Investigation of Transfer Learning-Based Sentiment Analysis in Japanese* https://aclanthology.org/P19-1458.pdf

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser and Illia Polosukhin. 2017. *Attention Is All You Need.* In arXiv preprint arXiv:1706.03762v5

Rafael A. Calvo, and Sunghwan Mac Kim. 2013. Emotions in text: dimensional and categorical models. Computational Intelligence, 29(3):527-543. https://www.dhi.ac.uk/san/waysofbeing/data/health-jones-calvo-2013a.pdf

James A. Russell. 1980. A circumplex model of affect. Journal of Personality and Social Psychology, 39(6):1161.

Wen-Li Wei, Chung-Hsien Wu, and Jen-Chun Lin. 2011. A regression approach to affective rating of Chinese words from ANEW. In Proc. of ACII-11, pages 121-131.

N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan. 2013. Distributional semantic models for affective text analysis. IEEE Transactions on Audio, Speech, and Language Processing, 21(11): 2379-2392.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie Zhang. 2016a. Building Chinese affective resources in valence-arousal dimensions. In Proc. of NAACL/HLT-16, pages 540-545.

Steven Du and Xi Zhang. 2016. Aicyber's system for IALP 2016 shared task: Character-enhanced word vectors and Boosted Neural Networks, in Proc. of IALP-16, pages 161–163.

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 Proc. of IJCNLP-17, pages 42-52.

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.

Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A. Calvo. 2010. Evaluation of unsupervised emotion models to textual affect recognition. In Proc. of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pages 62-70.

Georgios Paltoglou, Mathias Theunis, Arvid Kappas, and Mike Thelwall. 2013. Predicting emotional responses to long informal text. IEEE Trans. Affective Computing, 4(1):106-115.

Pranav Goel, Devang Kulshreshtha, Prayas Jain and Kaushal Kumar Shukla. 2017. Prayas at EmoInt 2017: An Ensemble of Deep Neural Architectures for Emotion Intensity Prediction in Tweets, in Proc. of WASSA-17, pages 58–65.

Suyang Zhu, Shoushan Li and Guodong Zhou. 2019. Adversarial attention modeling for multi-dimensional emotion regression, in Proc. of ACL-19, pages 471–480.

Jin Wang, Liang-Chih Yu, K. Robert Lai and Xuejie Zhang. 2016. Community-based weighted graph model for valence-arousal prediction of affective words, IEEE/ACM Trans. Audio, Speech and Language Processing, 24(11):1957-1968.

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.