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

SoochowDS at ROCLING-2021 Shared Task: Text Sentiment Analysis Using BERT and LSTM

Ruei-Cyuan Su, Sing-Seong Chong, Tzu-En Su and Ming-Hsiang Su
Soochow University, Taiwan
70613rex, chongzhishan123, 70614roy, huntfox.su@gmail.com

摘要

在這次的挑戰賽中，本研究提出結合 BERT-based 詞向量模型和 LSTM 預測模型進行文本 Valence 和 Arousal 數值預測。其中 BERT-based 詞向量為 768 維，並依序將文句中每個詞向量依序輸入 LSTM 模型中進行預測。實驗結果得知我們所提出的 BERT 結合 LSTM 模型遠優於 Lasso Regression 回歸模型的結果。

Abstract

In this shared task, this paper proposes a method to combine the BERT-based word vector model and the LSTM prediction model to predict the Valence and Arousal values in the text. Among them, the BERT-based word vector is 768-dimensional, and each word vector in the sentence is sequentially fed to the LSTM model for prediction. The experimental results show that the performance of our proposed method is better than the results of the Lasso Regression model.

關鍵字：BERT、LSTM、Lasso Regression
Keywords: BERT、LSTM、Lasso Regression

1 Introduction

情緒分析是一個非常熱門的研究領域，學者們提出許多創新方法去分析和預測。公司可以根據資料(如產品留言)，進行顧客對產品的評價分析或尋找產品銷售問題，以便提高銷售量等。在文字情緒分析中，學者們採用兩大指標，別是 Valence 和 Arousal，進行文本情緒分析。其中 Valence 主要是區別情感正向與負向，而 Arousal 則是判斷情感是沉靜還是喚起。這兩大指標普遍用於檢測與識別文本情感訊息。如”最近上課遇到很多問題，情緒低落”對應的 Valence 和 Arousal 分別為 1.75 和 5.64。

會議紀錄對市場走勢起著重要作用，因為它們提供了對市場走勢的鳥瞰圖。 因此，人們越來越有興趣從大型金融文本中分析和提取各個方面的情緒以進行經濟預測。然而由於缺乏大型標記數據集，Aspect-based Sentiment Analysis (ABSA) 並未廣泛用於金融數據。於是 Wang [1] 提出一個模型來訓練 ABSA 的金融文件，並分析其對各種宏觀經濟指標的預測能力。Wang [1] 運用 FinBERT 技術合併文字來達到文件級別的分析。其實驗結果顯示 Federal Open Market Committee (FOMC) 的報告文件可以解釋63%市場的成長率，員工的情緒與通貨膨脹能解釋47%和19%相對應的經濟因數。

網路的普及創造了一個振興的數位媒體。隨著新聞點擊次數驅動的貨幣化，在網路新聞競爭激烈的氛圍中，記者們調整他們的報告以適應這樣的氛圍。由此產生的消極偏見是有害的，會導致焦慮和情緒障礙。Kumar 等人 [2] 在各種數據集上訓練 4 個管線化情感分析模型(Sequential、LSTM、BERT 和 SVM 模型)。 經過組合後，行動裝置 APP 只顯示會鼓舞人心的故事供用戶閱讀。結果顯示有 1,300 名用戶對該 APP 評價為 4.9 星，85% 的用戶回饋通過使用此 APP 改善了心理健康。

由於來自不同文化和教育背景的人對網路的使用呈指數增長，具仇恨攻擊的線上言論偵測已成為當今的一個關鍵問題。區分文本消息是否屬於仇恨言論和攻擊性語言是自動檢測文本內容的關鍵挑戰。Bencheng 等人 [3] 提出一種將推文自動分類為三類的方法：仇恨、攻擊性和兩者都不是。他們利用公共推文數據集，首先進行實驗構建 BERT-based embedding 結合 Bi-Directional Long Short-Term Memory (BI-LSTM) 模型，然後他們也嘗試使

用預訓練的 Glove-based embedding 結合相同的神經網絡架構。實驗考慮不同的神經網絡架構、學習率和歸一化方法，對他們所提之 BI-LSTM 模型進行超參數調整分析。在調整模型並使用最佳參數組合後，在測試數據上對其進行評估時達到了 92% 以上的準確率。在參考上述研究後，本研究提出結合 BERT-based 詞向量模型和 LSTM 模型進行訓練，以完成 Valence 和 Arousal 文本預測。

2 資料集說明

本研究第一個採用的是 CVAT 1.0 及 CVAT 2.0 中文維度情感語料庫 [4]，是一個情感語料庫。其中包含從網路中提取的 2,969 的條列句子(CVAT 2.0)及 2,009 的條列句子(CVAT 1.0)，並且分為六個不同類別：新聞文章、政治論壇、汽車論壇、酒店評論、書評和筆記本電腦評論。每個句子都用人工分類的方式標明了 Valence 和 Arousal 之維度的實值分數。這兩個維度的範圍從 1（高度消極或平靜）到 9（高度積極或興奮）。本研究第二個採用的是 CVAW 4.0 中文維度情感詞典，包含了 5,512 個單詞。每個單詞都使用人工分類的方式標明 Valence 和 Arousal 維度的實值分數。本研究第三個採用的是 CVAP 2.0 中文維度情感片語，包含 2,998 個多詞短語。每個短語由一個情感詞和一個或多個修飾詞組成，例如修飾詞的否定詞、情態詞和程度副詞。最後再使用測試資料來完成 Valence 和 Arousal 文本預測。

3 System framework

在系統架構說明中，本研究首先使用 Jiba 斷詞工具將 CVAT 1.0 及 CVAT 2.0 這兩個資料集進行斷詞處理。本研究使用了中文維度情感辭典 CVAW、中文維度型情感片語 CVAP 資料集、地名或人名等加入 Jiba 的字典裡，使斷詞更加正確。接著本研究以 word2vector, doc2vector 和 BERT 進行詞向量模型訓練。最後分別進行 Lasso Regression 和 LSTM 模型對 Valence 和 Arousal 進行訓練，並用以預測測試集得到 Valence 和 Arousal 的結果。

3.1 斷詞模型

目前中文有兩個斷詞模型可用，一個是 Jieba 斷詞工具，而另一個是中研院的 CKIP 斷詞工具。本研究分析使用 Jieba 斷詞工具，這個工具有 python 的介面，使用上非常容易，可輸入繁體字典。 因 Jiba 的本身斷詞效果有限，因此本研究加入中文維度情感辭典 CVAW、中文維度型情感片語 CVAP 資料集。實驗結果發現加入字典的地名、人名、專業名稱並無有效修正，因此本研究再整理出應修正的名稱加入字典，得到正確的斷詞結果。

3.2 詞向量模型

Word2Vector [5]是輕量級的神經網絡，其模型僅僅包括輸入層、隱藏層和輸出層，模型框架根據輸入輸出的不同，主要包括 CBOW 和 Skip-gram 模型。CBOW 的方式為知道詞w_t的上下文$w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$的情況預測當前詞$w_t$，而 Skip-gram 則是知道了詞$w_t$的情況，對詞的上下文$w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}$進行預測。

首先介紹 Simple CBOW Mode，在我們的設置中，詞彙量大小為V隱藏層大小為N。輸入x是一層 one-hot representation vector，這意味著對於給定的輸入上下文詞$\{x_1 \dots x_V\}$，裡面共V個單元，其中只有一個為 1，所有其他單元為 0。例如$x = [0, \dots, 1, \dots, 0]$。輸入層和輸出層之間的權重可以用一個V × N矩陣W表示。$W_{\mathrm{V\times N}} = \{w_{ki}\}$的每一行是輸入層關聯詞的$N$維向量表示$v_w$。給定一個上下文（一個詞），假設$x_k = 1, x_{k'} = 0, k \neq k'$，得

$$h = x^T W = W^T_{(k,\cdot)} \coloneqq v^T_{wI} \tag{1}$$

這只是將W的第k行複製到h。v_{wI}是輸入詞wI的向量表示。從隱藏層至輸出層，權重矩陣為

$$W'_{N\times V} = \{w'_{ij}\} \tag{2}$$

這是一個N × V的矩陣。使用這些權重，我們可以計算詞彙中每單詞的分數u_i，

$$u_j = {v'_{wj}}^T \cdot h \tag{3}$$

其中v'_{wj}是矩陣W'的第j列。然後我們可以使用 softmax 得到詞的後驗分佈，這是一個多項式分佈。

$$p = \left(w_j \middle| w_I\right) = y_j = \frac{\exp\left(u_j\right)}{\sum_{j'=1}^{V} \exp\left(u_{j'}\right)} \tag{4}$$

其中y_i是輸出層中第j個單元的輸出。對於損失函數，在訓練目標是在給定輸入上下文詞w_I的權重的情況下，最大化觀察實際輸出詞w_0將其在輸出層中的相應索引表示為j^*的條件概率

$$\max p(w_0|w_I) \tag{5}$$

$$= u_{j^*} - log \sum_{j'=1}^{V} \exp(u_{j'}) := -E$$

而$E = -logp(w_0|w_I)$為損失函數，其中j^*是輸出層中實際輸出詞的索引。

而在 Skip-gram Model 中，我們仍然使用 Simple CBOW Mode 對隱藏層輸出 h 相同的定義。在輸出層，我們不是輸出一個多項式分佈，而是輸出c個多項式分佈，

$$p(w_{c,j} = w_{O,c}|w_I) = \frac{\exp(u_{c,j})}{\sum_{j'=1}^{V} \exp(u_{j'})} \quad (6)$$

其中$w_{c,j}$是輸出層第 c 個面板上的第j個單詞；$w_{O,c}$是實際第C 個輸出上下文詞中的詞；w_I是唯一的輸入詞；$y_{c,j}$是第j個的輸出輸出層第C個面板上的單元；$u_{c,j}$是第j個單元在第c個單元上的淨輸入輸出層面板。其中因為輸出層面板共享相同的權重，因此$u_{c,j} = u_j = {v'_{wj}}^T \cdot h,\ c = 1,2,\dots,C$，其他參數如同 Simple CBOW Mode。損失函數與 Simple CBOW Mode 沒有太大差別，

$$E = -logp(w_{O,1}, w_{O,2}, \dots, w_{O,C}|w_I)$$
$$= -log \prod_{c=1}^{C} \frac{\exp(u_{c,j_c^*})}{\sum_{j'=1}^{V} \exp(u_{j'})} \quad (7)$$

最後詞向量模型為 BERT 模型(Bidirectional Encoder Representations from Transformer)，是 Google 以無監督的方式利用大量無標記文本的模型。訓練資料來源于 Wikipedia (2.5B 字)加上 Book coupus (800M 字)。批量大小為 1024 序列*128 長度或 256 序列*512 長度。BERT 分為兩種 BERT-Base (12-layer, 768-Hidden, 12-head)和 BERT-Large (24-layer, 1024hidden, 16-head)。BERT 無需標記好的資料或解釋即可進行分析。Transformer 是 BERT 的核心模組，而 Attention 是 transformer 的核心部分，主要是增強語義向量，在不同的字結合中，代表識別字所帶來的意思。

3.3 Lasso Regression

線性回歸(linear regression)，為用線性函數(hypothesis) $f(x) = wx + b$ 去擬合一組數據 $D = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$，找到一組$(w^*, b^*)$，使損失 $J = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$ (mse)最小。Lasso 的全稱 Least Absolute Shrinkage and Selection Operator，又譯最小絕對值收斂和選擇算子、套索算法，其 cost function 為

$$J = \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2 + \lambda\|w\|_1 \quad (8)$$

其中λ為乘子。目標為$min_{w,b}J$，因此也將它寫成

$$min_{w,b}\ \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2, \quad s.t.\ \|w\|_1 \le t \quad (9)$$

其中 t 可理解為正規化力度。以$x \in R^2$為例，對 w 的限制空間為正方形，因此$argmin_{w,b}\ \frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$的解容易切在 w 某一維為 0 的點可解決過度擬和問題以及來做 feature selection。

3.4 LSTM 模型

LSTM 是為了解決 RNN 的缺點，如不能準確處理長期序列、時間的資料。LSTM 是由四個結構所組成，輸入門 (Input Gate)，儲存細胞（Memory Cell），遺忘門 (Forget Gate)，輸出門 (Output Gate)。Input gate 主要負責控制這個值輸入，Memory Cell 儲存值，下階段在使用，Output Gate 輸出 output，Forget Gate 是否保留或刪除特徵(feature)。LSTM 操作思路就是把輸入到類神經網路層處理產生出結果，過程當中，記住某些特徵，然後會跟著這些經驗來判斷或學習。其中 (10)-(15)分別為 Input Gate, Forget Gate 和 Output Gate 計算公式。

$$f_t = \sigma(W_f \cdot [X_t, h_{t-1}] + b_f) \quad (10)$$
$$i_t = \sigma(W_i \cdot [X_t, h_{t-1}] + b_i) \quad (11)$$
$$c_t = tanh(W_c \cdot [X_t, h_{t-1}] + b_c) \quad (12)$$
$$C_t = f_t \times C_{t-1} + i_t \times c_t \quad (13)$$
$$o_t = \sigma(W_o \cdot [X_t, h_{t-1}] + b_o) \quad (14)$$
$$h_t = o_t \times tanh(C_t) \quad (15)$$

4 實驗結果

4.1 皮爾遜相關係數

皮爾遜相關係數 (Pearson product-moment correlation coefficient) [6]，又稱作 PPMCC 或 PCCs，常用 r 或 Pearson's r 表示。在統計學上，用於度量兩個變數x和y之間的相關程度（線性相依），其值介於-1 與 1 之間。於自然科學領域中，該係數廣用於度量兩個變數之間的線性相依程度。本使用的r為(x_i, y_i)樣本點的標準分數的均值估算：

$$r = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right) \quad (16)$$

其中 $\frac{x_i-\bar{x}}{\sigma_x}$ 為 x_i 樣本的標準分數，$\bar{x}$為x_i樣本的平均值，σ_x為x_i樣本的標準差。

4.2 平均絕對誤差

平均絕對誤差（Mean Absolute Error, MAE），對同一物理量，進行多次測量時，其各次測量值和其絕對誤差不會相同，因此把各次測量的絕對誤差取絕對值後再求平均值，稱為平均絕對誤差，由於離差被絕對值化，不會出現正負相抵消的情況，因能更佳地反映預測值誤差的實際情形。指各個變量值平均數的離差絕對值的算術平均數。f_i為預測值，y_i為真實值，$e_i = |f_i - y_i|$為絕對誤差，

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| = \frac{1}{n}\sum_{i=1}^{n}|e_i| \qquad (17)$$

從上可知， MAE 就是指你的預測值與真實值之間平均相差多大。

4.3 實驗參數設定

詞向量模型部分，使用的 Word2Vec 為 CBOW Multi-Word Context Model 的採用均值的方式。在表示當前詞與預測詞在一個句子的最大距離設為 10，對字典做截斷的詞頻次數少於 1 的單詞會被丟棄。訓練並行數為 4。使用 negative sampling 的技巧，採用 negative sampling 設置 5 個 noise words，初始化權重則使用 python 的 hash 函數。替代次數為 5，在分配 word index 的時候會先對單詞基於頻率降序排序，每一批的傳遞單詞數量為 10000，學習速率為 0.025。特徵向量的維度為 200。

使用的 Lasso Regression 的替代次數固定為 10000，調整調整正則化的強度的值(alpha)來尋找最佳模型，再使用預測出的數值和正確數值本身算出 MAE 和皮爾遜相關係數。Alpha(A)為 Arousal Lasso regression(簡 Arousal LR)模型參數的數值，Alpha (V) 為 Valence Lasso regression (簡 Valence LR) 模型參數的數值。

而 LSTM 模型中，我們設定一層 LSTM 和一層 Neural Network，作為預測模型，其中輸入為 BERT 所輸出之詞向量串接、Loss Function 是 Mean Square Error、Optimizer 是 Adam、訓練 epochs 是 200。

4.4 實驗模型選擇

分析數值後，我們對 Arousal LR 模型以及 Valence LR 模型，每個取 Alpha=0.00001 和 Alpha=0.0001 去預測正確結果。結果發現無論在 Arousal LR 模型或 Valence LR 模型，對於 Alpha=0.00001 所預測出的結果有些超過 Arousal 和 Valence 維度限制範圍 (1~9) 過多，因此最後使用 Alpha=0.0001 的數值作為 Arousal LR 模型和 Valence LR 模型的參數，並以之預測結果。其中 Table 1 和 Table 2 為 Lasso Regression 模型實驗結果，Table 3 為不同模型實驗結果，最後我們選擇 BERT+LSTM model 作為最終架構。

Table 1: Arousal Lasso Regression Evaluation

Alpha(α)	MAE	Pearson's *r*
0.00001	0.5338	-0.192
0.0001	0.8099	-0.052
0.001	0.8142	-0.046
0.01	0.8205	-0.033
0.1	0.8361	-

Table 2: Valence Lasso Regression Evaluation

Alpha(α)	MAE	Pearson's *r*
0.00001	0.6842	0.837
0.0001	1.1199	0.241
0.001	1.1346	0.211
0.01	1.1509	0.172
0.1	1.1702	0.081

Table 3: Model Evaluation

Model	MAE	Pearson's *r*
Arousal Lasso Regression	1.0107	-0.046
Valence Lasso Regression	1.3176	0.211
BERT+LSTM	**0.052**	**0.998**

5 Conclusion and Future Work

在這次的挑戰賽中，我們提出結合 BERT-based 詞向量模型和 LSTM 預測模型進行文本 Valence 和 Arousal 數值預測。實驗結果得知我們所提出的模型遠優於 Lasso Regression 回歸模型的結果。在未來的研究中，我們將持續修正模型中的參數設定和字典的擴增，以期能提升整體系統的效能。

References

[1] Sarah-Yifei Wang. 2021. *Aspect-based Sentiment Analysis in Document - FOMC Meeting Minutes on Economic Projection*, arXiv:2108.04080.

[2] Saurav Kumar, Rushil Jayant, and Nihaar Charagulla. 2021. *Sentiment Analysis on the News to Improve Mental Health*, arXiv:2108.07706.

[3] Bencheng Wei, Jason Li, Ajay Gupta, Hafiza Umair, Atsu Vovor, Natalie Durzynski. 2021. *Offensive Language and Hate Speech Detection with Deep Learning and Transfer Learning*. arXiv:2108.03305.

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

[5] Rong, Xin. 2014. *word2vec parameter learning explained*, arXiv:1411.2738.

[6] Lee Rodgers, Joseph, and W. Alan Nicewander. 1988. Thirteen ways to look at the correlation coefficient. *The American Statistician*, 42(1): 59-66. https://doi.org/10.1080/00031305.1988.10475524.