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

ROCLING-2021 Shared Task: Dimensional Sentiment Analysis for Educational Texts

Liang-Chih Yu[1], Jin Wang[2], Bo Peng[3], Chu-Ren Huang[3]
[1]Department of Information Management, Yuan Ze University
[2] School of Information Science and Engineering, Yunnan University
[3]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
Contact: lcyu@saturn.yzu.edu.tw, wangjin@ynu.edu.cn,
peng-bo.peng@polyu.edu.hk, churen.huang@polyu.edu.hk

Abstract

This paper presents the ROCLING 2021 shared task on dimensional sentiment analysis for educational texts which seeks to identify a real-value sentiment score of self-evaluation comments written by Chinese students in the both valence and arousal dimensions. Valence represents the degree of pleasant and unpleasant (or positive and negative) feelings, and arousal represents the degree of excitement and calm. Of the 7 teams registered for this shared task for two-dimensional sentiment analysis, 6 submitted results. We expected that this evaluation campaign could produce more advanced dimensional sentiment analysis techniques for the educational domain. All data sets with gold standards and scoring script are made publicly available to researchers.

1 Introduction

The goal of sentiment analysis is to automatically identify affective information within texts. There are two major models to represent affective states: categorical and dimensional approaches (Calvo and Kim, 2013). The categorical approach represents affective states as several discrete classes (e.g., positive, negative, neutral), while the dimensional approach represents affective states as continuous numerical values on multiple dimensions, such as valence-arousal (VA) space (Russell, 1980), as shown in Fig. 1. The valence represents the degree of pleasant and unpleasant (or positive and negative) feelings, and the arousal represents the degree of excitement and calm. Based on this two-dimensional representation, any affective state can be represented as a point in the VA coordinate plane by determining the degrees of valence and arousal of given words (Wei et al., 2011; Malandrakis et al., 2013; Wang et al., 2016a; Du and Zhang, 2016; Wu et la., 2017) or texts (Paltoglou et al, 2013; Goel et la., 2017; Zhu et al., 2019; Wang et al., 2019; 2020; Cheng et al, 2021; Wu et al., 2021, Xie et al., 2021). In 2016, we hosted a first dimensional sentiment analysis task for Chinese words (Yu et al., 2016b). In 2017, we extended this task to include both word- and phrase-level dimensional sentiment analysis (Yu et al., 2017). This year, we explore the sentence-level dimensional sentiment analysis task on educational texts (students' self-evaluated comments).

Structured data such as attendance, homework completion and in-class participation have been extensively studied to predict students' learning performance. Unstructured data, such as self-evaluation comments written by students, is also a useful data resource because it contains rich emotional information that can help illuminate the emotional states of students (Yu et al., 2018). Dimensional sentiment analysis is an effective technique to recognize the valence-arousal ratings from texts, indicating the degree from most negative to most positive for valence, and from most calm to most excited for arousal. This shared task provides an evaluation platform for the development and implementation of advanced techniques for dimensional sentiment analysis in the educational domain.

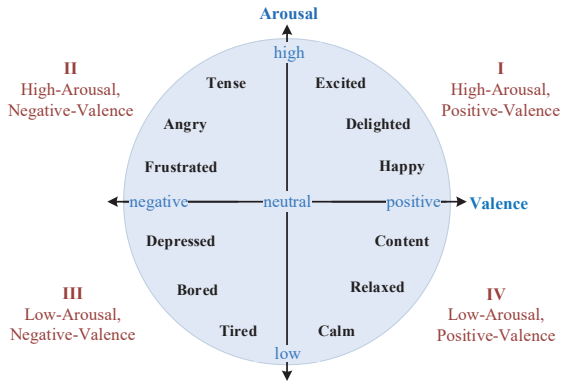

Figure 1: Two-dimensional valence-arousal space.

2 Task Description

In this task, participants are asked to provide a real-valued score from 1 to 9 for both valence and arousal dimensions for each self-evaluation comment. The input format is "sentence_id, sentence", and the output format is "sentence_id, valence_rating, arousal_rating". Below are the input/output formats of the example sentences.

Example 1:

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

Example 2:

Input: 2, 覺得課程進度有點快，內容難以消化

Output: 2, 3.0, 4.0

3 Datasets

Training set: There are three datasets annotated with valence-arousal ratings for training: 1) Chinese Valence-Arousal Words (CVAW)[1] (Yu et al., 2016a), which contains 5,512 single words; 2) Chinese Valence-Arousal Words (CVAP)[2] (Yu et al., 2017), which contains 2,998 multi-word phrases; 3) Chinese Valence-Arousal Words (CVAT)[3] (Yu et al., 2016a), which contains 2,969 sentences.

Test set: A total of 1,600 sentences were collected from the self-evaluated comments written by university students. Each sentence was then annotated with valence-arousal ratings by seven annotators and the average ratings were taken as ground truth. Once the rating process was finished, a corpus clean-up procedure was performed to remove outlier ratings that did not fall within the mean plus/minus 1.5 standard deviations. They were then excluded from the calculation of the average ratings for each sentence.

The policy of this shared task was implemented as is an open test. That is, in addition to the above official datasets, participating teams were allowed to use other publicly available data for system development, but such sources should be specified in the final technical report.

[1] http://nlp.innobic.yzu.edu.tw/resources/cvaw.html
[2] http://nlp.innobic.yzu.edu.tw/resources/cvap.html
[3] http://nlp.innobic.yzu.edu.tw/resources/cvat.html

4 Evaluation Metrics

Prediction performance is evaluated by examining the difference between machine-predicted ratings and human-annotated ratings, in which valence and arousal are treated independently. The evaluation metrics include Mean Absolute Error (MAE) and Pearson Correction Coefficient (*r*), as shown in the following equations.

- **Mean absolute error (MAE)**

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

- **Pearson correlation coefficient (*r*)**

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

where A_i is the actual value, P_i is the predicted value, n is the number of test samples, $\bar{A}$ and $\bar{P}$ respectively denote the arithmetic mean of A and P, and σ is the standard deviation. The MAE measures the error rate and the PCC measures the linear correlation between the actual values and the predicted values. A lower MAE and a higher PCC indicate more accurate prediction performance.

5 Evaluation Results

5.1 Participants

Table 1 summarizes the submission statistics for 7 participating teams (CYUT, NCU-NLP, DeepNLP, NTUST-NLP-1, NTUST-NLP-2, SCUDS and SoochowDS). In the testing phase, each team was allowed to submit at most two runs. Six teams submitted two runs, yielding a total of 12 runs for comparison.

Team	Affiliation	#Run
CYUT	Chaoyang University of Technology	2
NCU-NLP	National Central University	2
DeepNLP	Nanjing University	0
NTUST-NLP-1	National Taiwan University of Science and Technology	2
NTUST-NLP-2	National Taiwan University of Science and Technology	2
SCUDS	Soochow University	2
SoochowDS	Soochow University	2

Table 1: Submission statistics for all participating teams.

Team	Valence MAE	Valence *r*	Arousal MAE	Arousal *r*
Baseline	1.143	0.457	0.954	0.278
CYUT-run1	1.695	-0.017	1.177	0.040
CYUT-run2	1.685	0.007	1.252	-0.021
NCU-NLP-run1	0.625	0.900	0.938	0.549
NCU-NLP-run2	0.611	0.904	0.989	0.582
ntust-nlp-1-run1	0.684	0.912	0.906	0.607
ntust-nlp-1-run2	**0.586**	0.901	0.885	0.585
ntust-nlp-2-run1	0.654	0.905	0.880	0.581
ntust-nlp-2-run2	0.667	**0.913**	**0.866**	**0.616**
SCUDS-run1	0.953	0.694	1.054	0.375
SCUDS-run2	0.975	0.667	1.039	0.354
SoochowDS-run1	2.421	0.073	1.327	0.051
SoochowDS-run2	1.073	0.584	1.125	0.228
Late-CYUT-run1	0.674	0.870	0.901	0.531
Late-CYUT-run2	0.600	0.900	0.877	0.565

Table 2: Comparative results of valence-arousal prediction on the test set.

5.2 Baseline

We implemented a baseline using a lexicon-based method to calculate the VA ratings of texts by averaging the VA ratings of affective words match between the texts and CVAW 4.0 (Yu et al., 2016a). For the test instances that contain no affective words in the lexicon, their VA ratings will be assigned with 5.

5.3 Results

Tables 2 shows the results of valence-arousal prediction on the test set. Most of the results outperformed the baseline. The three best performing systems are summarized as follows.

- Valence MAE: NTUST-NLP-1-run2, NCU-NLP-run2 and NCU-NLP-run1
- Valence *r*: NTUST-NLP-2-run2, NTUST-NLP-1-run1 and NTUST-NLP-2-run1
- Arousal MAE: NTUST-NLP-2-run2, NTUST-NLP-2-run1 and NTUST-NLP-1-run2
- Arousal *r*: NTUST-NLP-2-run2, NTUST-NLP-1-run1 and NTUST-NLP-1-run2

There is a late submission for the CYUT team because the order of their scores in the initial submission is not consistent with that of the test set, thus yielding a negative correlation. The results of the late submission show the actual performance of their proposed method.

6 Conclusions

This study describes an overview of the ROCLING 2021 shared task on dimensional sentiment analysis for educational texts, including task design, data preparation, performance metrics and evaluation results. We hope the data sets collected and annotated for this shared task can facilitate and expedite future development in this research area. Therefore, all data sets with gold standard and scoring script are publicly available[4].

Acknowledgments

This work was supported by the Ministry of Science and Technology, Taiwan, ROC, under Grant No. MOST 110-2628-E-155-002.

References

Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: dimensional and categorical models. *Computational Intelligence*, 29(3):527-543.

Yu-Ya Cheng, Yan-Ming Chen, Wen-Chao Yeh, and Yung-Chun Chang.2021. Valence and Arousal-Infused Bi-Directional LSTM for Sentiment Analysis of Government Social Media Management. *Applied Sciences*, 11(2):880.

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*.

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. WASSA-17*, page 58–65.

Nikos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan, 2011. Kernel models for affective lexicon creation. In *Proc. of INTERSPEECH-11*, pages 2977-2980.

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.

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

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 Trans. Audio, Speech and Language Processing*, 24(11):1957-1968.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016b. Dimensional sentiment analysis using a regional CNN-LSTM model. In *Proc. of ACL-16*, pages 225-230.

Jin Wang, Liang-Chih Yu, K. Robert Lai and Xuejie Zhang. 2019. Investigating Dynamic Routing in Tree-Structured LSTM for Sentiment Analysis. In *Proc. of EMNLP/IJCNLP-19*, pages 3423-3428.

Jin Wang, Liang-Chih Yu, K. Robert Lai and Xuejie Zhang. 2020. Tree-Structured Regional CNN-LSTM Model for Dimensional Sentiment Analysis. *IEEE/ACM Trans. Audio, Speech and Language Processing*, 28:581-591.

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.

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, Shared Tasks*, pages 47–52.

Jheng-Long Wu, Min-Tzu Huang, Chi-Sheng Yang, and Kai-Hsuan Liu. 2021. Sentiment analysis of stock markets using a novel dimensional valence–arousal approach. *Soft Computing*, 25:4433–4450.

Housheng Xie, Wei Lin, Shuying Lin, Jin Wang, and Liang-Chih Yu. 2021. A multi-dimensional relation model for dimensional sentiment analysis. *Information Sciences*, 579:832-844.

Liang-Chih Yu et al. 2018. Improving Early Prediction of Academic Failure Using Sentiment Analysis on Self-evaluated Comments. *Journal of Computer Assisted Learning*, 34(4):358-365.

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.

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 *Proc. of IALP-16*, pages 156-160.

Liang-Chih Yu, Lung-Hao Lee, Jin Wang and Kam-Fai Wong. 2017. IJCNLP-2017 Task 2: Dimensional Sentiment Analysis for Chinese Phrases. In *Proc. of IJCNLP-17, Shared Tasks*, pages 9-16.

Suyang Zhu, Shoushan Li and Guodong Zhou. 2019. Adversarial Attention Modeling for Multi-dimensional Emotion Regression. In *Proc. of ACL-19*, pages 471–480.

[4] https://rocling2021.github.io/