IJCNLP-2017 Task 2: Dimensional Sentiment Analysis for Chinese Phrases

Liang-Chih Yu¹,², Lung-Hao Lee³, Jin Wang⁴, Kam-Fai Wong⁵
¹Department of Information Management, Yuan Ze University, Taiwan
²Innovative Center for Big Data and Digital Convergence, Yuan Ze University, Taiwan
³Graduate Institute of Library and Information Studies, National Taiwan Normal University
⁴School of Information Science and Engineering, Yunnan University, Yunnan, China
⁵The Chinese University of Hong Kong, Hong Kong, China
Contact: lcyu@saturn.yzu.edu.tw, lhlee@ntnu.edu.tw,
        wangjin@ynu.edu.cn, kfwong@se.cuhk.edu.hk

Abstract

This paper presents the IJCNLP 2017 shared task on Dimensional Sentiment Analysis for Chinese Phrases (DSAP) which seeks to identify a real-value sentiment score of Chinese single words and multi-word phrases 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 19 teams registered for this shared task for two-dimensional sentiment analysis, 13 submitted results. We expected that this evaluation campaign could produce more advanced dimensional sentiment analysis techniques, especially for Chinese affective computing. All data sets with gold standards and scoring script are made publicly available to researchers.

1 Introduction

Sentiment analysis has emerged as a leading technique to automatically identify affective information within texts. In sentiment analysis, affective states are generally represented using either categorical or 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., 2011; Wang et al., 2016a) or texts (Kim et al., 2010; Paltoglou et al, 2013; Wang et al., 2016b). Dimensional sentiment analysis has emerged as a compelling topic for research with applications including antisocial behavior detection (Munezero et al., 2011), mood analysis (De Choudhury et al., 2012) and product review ranking (Ren and Nickerson, 2014).

The IJCNLP 2017 features a shared task for dimensional sentiment analysis for Chinese words, providing an evaluation platform for the development and implementation of advanced techniques for affective computing. Sentiment lexicons with valence-arousal ratings are useful resources for the development of dimensional sentiment applications. Due to the limited availability of such VA lexicons, especially for Chinese, the objective of the task is to automatically acquire the valence-arousal ratings of Chinese affective words and phrases.

The rest of this paper is organized as follows. Section II describes the task in detail. Section III introduces the constructed datasets. Section IV proposes evaluation metrics. Section V reports the results of the participants’ approaches. Conclusions are finally drawn in Section VI.
2 Task Description

This task seeks to evaluate the capability of systems for predicting dimensional sentiments of Chinese words and phrases. For a given word or phrase, participants were asked to provide a real-valued score from 1 to 9 for both the valence and arousal dimensions, respectively indicating the degree from most negative to most positive for valence, and from most calm to most excited for arousal. The input format is “term_id, term”, and the output format is “term_id, valence_rating, arousal_rating”. Below are the input/output formats of the example words “好” (good), “非常好” (very good), “滿意” (satisfy) and “不滿意” (not satisfy).

Example 1:
Input: 1, 好
Output: 1, 6.8, 5.2

Example 2:
Input: 2, 非常好
Output: 2, 8.500, 6.625

Example 3:
Input: 3, 滿意
Output: 3, 7.2, 5.6

Example 4:
Input: 4, 不滿意
Output: 4, 2.813, 5.688

3 Datasets

Training set: For single words, the training set was taken from the Chinese Valence-Arousal Words (CVAW)¹ (Yu et al., 2016a) version two, which contains 2,802 affective words annotated

Test set: For single words, we selected 750 words that were not included in the CVAW 2.0 from NTUSD (Ku and Chen, 2007) using the same method presented in our previous task on Dimensional Sentiment Analysis for Chinese Words (Yu et al, 2016b).

Each single word in both training and test sets was annotated with valence-arousal ratings by five annotators and the average ratings were taken as ground truth. Each multi-word phrase was rated by at least 10 different annotators. 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 phrase.

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.

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)

¹ http://nlp.innobic.yzu.edu.tw/resources/cvaw.html
and Pearson Correction Coefficient (PCC), as shown in the following equations.

- **Mean absolute error (MAE)**
  \[
  MAE = \frac{1}{n} \sum_{i=1}^{n} | A_i - P_i | 
  \]  
  (1)

- **Pearson correlation coefficient (PCC)**
  \[
  PCC = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{A_i - \overline{A}}{\sigma_A} \right) \left( \frac{P_i - \overline{P}}{\sigma_P} \right) 
  \]  
  (2)

where \(A_i\) is the actual value, \(P_i\) is the predicted value, \(n\) is the number of test samples, \(\overline{A}\) and \(\overline{P}\) respectively denote the arithmetic mean of \(A\) and \(P\), and \(\sigma\) 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 19 participating teams including 7 from universities and research institutes in China (CASIA, G-719, LDCCNLP, SAM, THU_NGN, TeeMo and XMUT), 6 from Taiwan (CIAL, CKIP, NCTU-NTUT, NCYU, NLPSA and NTOU), 2 private films (AL_I_NLP and Mainiway AI), 2 teams from India (DeepCybErNet and Dlg), one from Europe (DCU) and one team from USA (UIUC). Thirteen of the 19 registered teams submitted their testing results. In the testing phase, each team was allowed to submit at most two runs. Three teams submitted only one run, while the other 10 teams submitted two runs for a total of 23 runs.

| Team            | Affiliation                                                                 | #Run |
|-----------------|-----------------------------------------------------------------------------|------|
| AL_I_NLP        | Alibaba                                                                     | 2    |
| CASIA           | Institute of Automation, Chinese Academy of Sciences                        | 1    |
| CIAL            | Academia Sinica & Taipei Medical University                                 | 2    |
| CKIP            | Institute of Information Science, Academia Sinica                           | 2    |
| DeepCybErNet    | Amrita University, India                                                    | 0    |
| Dlg             | IIT Hyderabad                                                               | 0    |
| DCU             | ADAPT Centre, Dublin City University, Ireland                              | 0    |
| G-719           | Yunnan University                                                           | 0    |
| LDCCNLP         | Fuzhou University                                                           | 2    |
| Mainiway AI     | Shanghai Mainiway Corp.                                                     | 2    |
| NCTU-NTUT       | National Chiao Tung University & National Taipei University of Technology   | 2    |
| NCYU            | National Chiayi University                                                  | 2    |
| NLPSA           | Institute of Information Science, Academia Sinica                           | 2    |
| NTOU            | National Taiwan Ocean University                                            | 2    |
| SAM             | Soochow University                                                          | 1    |
| THU_NGN         | Department of Electronic Engineering, Tsinghua University                   | 2    |
| TeeMo           | Southeast University                                                        | 0    |
| UIUC            | University of Illinois at Urbana Champaign                                  | 0    |
| XMUT            | Xiamen University of Technology                                             | 1    |

Table 1: Submission statistics for all participating teams.
5.2 Baseline

We implemented a baseline by training a linear regression model using word vectors as the only features. For single words, the regression was implemented by directly training word vectors to determine VA scores.

Given a word $w_i$, the baseline regression model is defined as

$$\begin{align*}
    Val_{w_i} &= W_{w}^{val} \cdot \text{vec}(w_i) + b_{w}^{val} \\
    Aro_{w_i} &= W_{w}^{aro} \cdot \text{vec}(w_i) + b_{w}^{aro}
\end{align*}$$

(3)

where $Val_{w_i}$ and $Aro_{w_i}$ respectively denote the valence and arousal ratings of $w_i$. $W$ and $b$ respectively denote the weights and bias. For phrases, we first calculate the mean vector of the constituent words in the phrase, considering each modifier word can also obtain its word vector. Give a phrase $p_j$, its representation can be obtained by,

$$\begin{align*}
    \text{vec}(p_j) &= \text{mean}[\text{vec}(w_1), \text{vec}(w_2), ..., \text{vec}(w_n)] \quad (4)
\end{align*}$$

where $w_i \in p_j$ is the word in phrase $p_j$. The regression was then trained using $\text{vec}(p_j)$ as a feature, defined as

$$\begin{align*}
    Val_{p_j} &= W_{p}^{val} \cdot \text{vec}(p_j) + b_{p}^{val} \\
    Aro_{p_j} &= W_{p}^{aro} \cdot \text{vec}(p_j) + b_{p}^{aro}
\end{align*}$$

(5)

| Word-Level  | Valence MAE | Valence PCC | Arousal MAE | Arousal PCC |
|-------------|-------------|-------------|-------------|-------------|
| Baseline    | 0.984       | 0.643       | 1.031       | 0.456       |
| AL_1_NLP-Run1 | 0.547       | 0.891       | 0.853       | 0.677       |
| AL_1_NLP-Run2 | 0.545       | 0.892       | 0.857       | 0.678       |
| CASIA-Run1  | 0.725       | 0.803       | 1.069       | 0.428       |
| CIAL-Run1   | 0.644       | 0.853       | 1.039       | 0.423       |
| CIAL-Run2   | 0.644       | 0.85       | 1.036       | 0.426       |
| CKIP-Run1   | 0.602       | 0.858       | 0.949       | 0.576       |
| CKIP-Run2   | 0.665       | 0.855       | 1.133       | 0.569       |
| LDCCNL-Run1 | 0.811       | 0.769       | 0.996       | 0.479       |
| LDCCNL-Run2 | 1.219       | 0.521       | 1.235       | 0.346       |
| MainiwayAI-Run1 | 0.715       | 0.796       | 1.032       | 0.509       |
| MainiwayAI-Run2 | 0.706       | 0.800       | 0.985       | 0.552       |
| NCTU-NTUT-Run1 | 0.632       | 0.846       | 0.952       | 0.543       |
| NCTU-NTUT-Run2 | 0.639       | 0.842       | 0.94       | 0.566       |
| NCYU-Run1   | 0.922       | 0.645       | 1.155       | 0.428       |
| NCYU-Run2   | 1.235       | 0.663       | 1.177       | 0.402       |
| NLPSA-Run1  | 1.108       | 0.561       | 1.207       | 0.351       |
| NLPSA-Run2  | 1.000       | 0.604       | 1.207       | 0.351       |
| NTOU-Run1   | 0.913       | 0.700       | 1.133       | 0.163       |
| NTOU-Run2   | 1.061       | 0.544       | 1.114       | 0.35       |
| SAM-Run1    | 1.098       | 0.639       | 1.027       | 0.378       |
| THU_NGN-Run1 | 0.610       | 0.857       | 0.940       | 0.623       |
| THU_NGN-Run2 | 0.509       | 0.908       | 0.864       | 0.686       |
| XMUT-Run1   | 0.946       | 0.701       | 1.036       | 0.451       |

Table 2: Comparative results of valence-arousal prediction for single words.
The word vectors were trained on the Chinese Wiki Corpus\textsuperscript{2} using the CBOW model of word2vec\textsuperscript{3} (Mikolov et al., 2013a; 2013b) (dimensionality=300 and window size=5).

5.3 Results

Tables 2 shows the results of valence-arousal prediction for single words. The three best performing systems are summarized as follows.

- Valence MAE: THU_NGN, AL_I_NLP and CKIP.
- Valence PCC: THU_NGN, AL_I_NLP and CKIP.

| Phrase-Level      | Valence MAE | Valence PCC | Arousal MAE | Arousal PCC |
|-------------------|-------------|-------------|-------------|-------------|
| Baseline          | 1.051       | 0.610       | 0.607       | 0.730       |
| AL_I_NLP-Run1     | 0.531       | 0.900       | 0.465       | 0.855       |
| AL_I_NLP-Run2     | 0.526       | 0.901       | 0.465       | 0.854       |
| CASIA-Run1        | 1.008       | 0.598       | 0.816       | 0.683       |
| CIAL-Run1         | 0.723       | 0.835       | 0.914       | 0.756       |
| CIAL-Run2         | 1.152       | 0.647       | 1.596       | 0.286       |
| CKIP-Run1         | 0.492       | 0.921       | \textbf{0.382} | 0.908       |
| CKIP-Run2         | 0.444       | 0.935       | 0.395       | 0.904       |
| LDCCNLP-Run1      | 0.822       | 0.762       | 0.489       | 0.828       |
| LDCCNLP-Run2      | 0.916       | 0.632       | 0.605       | 0.742       |
| MainiwayAI-Run1   | 0.612       | 0.861       | 0.554       | 0.793       |
| MainiwayAI-Run2   | 0.577       | 0.874       | 0.524       | 0.813       |
| NCTU-NTUT-Run1    | 0.454       | 0.928       | 0.488       | 0.847       |
| NCTU-NTUT-Run2    | 0.453       | 0.931       | 0.517       | 0.832       |
| NCYU-Run1         | 1.035       | 0.725       | 0.735       | 0.670       |
| NCYU-Run2         | 1.175       | 0.670       | 0.801       | 0.666       |
| NLPSA-Run1        | 0.709       | 0.818       | 0.632       | 0.732       |
| NLPSA-Run2        | 0.689       | 0.829       | 0.633       | 0.727       |
| NTUO-Run1         | 0.472       | 0.910       | 0.420       | 0.882       |
| NTUO-Run2         | 0.453       | 0.929       | 0.441       | 0.870       |
| SAM-Run1          | 0.960       | 0.669       | 0.722       | 0.704       |
| THU_NGN-Run1      | 0.349       | 0.960       | 0.389       | 0.909       |
| THU_NGN-Run2      | \textbf{0.345} | \textbf{0.961} | 0.385       | \textbf{0.911} |
| XMUT-Run1         | 1.723       | 0.064       | 1.163       | 0.084       |

Table 3: Comparative results of valence-arousal prediction for multi-word phrases.

- Arousal MAE: AL_I_NLP, THU_NGN and NCTU-NTUT.
- Arousal PCC: THU_NGN, AL_I_NLP and CKIP.

Tables 3 shows the results of valence-arousal prediction for multi-word phrases. The three best performing systems are summarized as follows.

- Valence MAE: THU_NGN, CKIP and NCTU-NTUT.
- Valence PCC: THU_NGN, CKIP and NCTU-NTUT.
- Arousal MAE: CKIP, THU_NGN and NTOU.
- Arousal PCC: THU_NGN, CKIP and NTOU.

\textsuperscript{2}https://dumps.wikimedia.org/
\textsuperscript{3}http://code.google.com/p/word2vec/
Table 4 shows the overall results for both single words and multi-word phrases. We rank the MAE and PCC independently and calculate the mean rank (average of MAE rank and PCC rank) for ordering system performance. The three best performing systems are THU_NGN, AL_I_NLP and CKIP.

Table 5 summarizes the approaches for each participating system. CASIA, SAM and XMUT did not submit reports on their developed methods. Nearly all teams used word embeddings. The most commonly used word embeddings were word2vec (Mikolov et al., 2013a; 2013b) and GloVe (Pennington et al., 2014). Others included FastText (Bojanowski et al., 2017), character-enhanced word embedding (Chen et al., 2015) and Cw2vec (Cao et al., 2017). For machine learning algorithms, six teams used deep neural networks such as feed-forward neural network (CKIP), boosted neural network (BNN) (AL_I_NLP), convolutional neural network (CNN) (NLPSA), long short-term memory (LSTM) (NCTU-NTUT and THU_NGN) and ensembles (Mainiway AI and THU_NGN). Three teams used regression-based methods such as support vector regression (CIAL, CKIP, LDCCNLP) and linear regression (CIAL). Other methods included a lexicon-based

| All-Level  | V-MAE | V-MAE Rank | V- PCC | V- PCC Rank | A-MAE | A-MAE Rank | A-PCC | A-PCC Rank | Mean Rank |
|------------|-------|------------|--------|-------------|-------|------------|-------|------------|-----------|
| THU_NGN-Run2 | 0.427 | 1 | 0.9345 | 1 | 0.6245 | 1 | 0.7985 | 1 | 1 |
| THU_NGN-Run1 | 0.4795 | 2 | 0.9085 | 2 | 0.6645 | 4 | 0.766 | 3 | 2.75 |
| AL_I_NLP-Run2 | 0.5355 | 3 | 0.8965 | 3 | 0.661 | 3 | 0.766 | 2 | 2.75 |
| AL_I_NLP-Run1 | 0.539 | 4 | 0.8955 | 4 | 0.659 | 2 | 0.761 | 4 | 3.5 |
| CKIP-Run1 | 0.547 | 7 | 0.8895 | 6 | 0.6655 | 5 | 0.742 | 5 | 5.75 |
| NCTU-NTUT-Run1 | 0.543 | 5 | 0.887 | 7 | 0.72 | 6 | 0.695 | 8 | 6.5 |
| NCTU-NTUT-Run2 | 0.546 | 6 | 0.8865 | 8 | 0.7285 | 7 | 0.699 | 7 | 7 |
| CKIP-Run2 | 0.5545 | 8 | 0.895 | 5 | 0.764 | 10 | 0.7365 | 6 | 7.25 |
| MainiwayAI-Run2 | 0.6415 | 9 | 0.837 | 10 | 0.7545 | 9 | 0.6825 | 9 | 9.25 |
| MainiwayAI-Run1 | 0.6635 | 10 | 0.8285 | 11 | 0.793 | 13 | 0.651 | 11 | 11.25 |
| LDCCNLP-Run1 | 0.8165 | 14 | 0.7655 | 13 | 0.7425 | 8 | 0.635 | 10 | 11.25 |
| NTOU-Run2 | 0.757 | 13 | 0.7365 | 15 | 0.7775 | 12 | 0.61 | 12 | 13 |
| CIAL-Run1 | 0.6835 | 11 | 0.844 | 9 | 0.9765 | 21 | 0.5895 | 14 | 13.75 |
| NTOU-Run1 | 0.6925 | 12 | 0.805 | 12 | 0.7765 | 11 | 0.5225 | 22 | 14.25 |
| CASIA-Run1 | 0.8665 | 16 | 0.7005 | 17 | 0.9425 | 19 | 0.5555 | 15 | 16.75 |
| NLPSA-Run2 | 0.8445 | 15 | 0.7165 | 16 | 0.92 | 17 | 0.539 | 20 | 17 |
| Baseline | 1.0175 | 20 | 0.6265 | 22 | 0.819 | 14 | 0.593 | 13 | 17.25 |
| NLPSA-Run1 | 0.9085 | 18 | 0.6895 | 18 | 0.9195 | 16 | 0.5415 | 18 | 17.5 |
| NCYU-Run1 | 0.9785 | 19 | 0.685 | 19 | 0.945 | 20 | 0.549 | 16 | 18.5 |
| SAM-Run1 | 1.029 | 21 | 0.654 | 21 | 0.8745 | 15 | 0.541 | 19 | 19 |
| CIAL-Run2 | 0.898 | 17 | 0.7485 | 14 | 1.316 | 24 | 0.356 | 23 | 19.5 |
| LDCCNLP-Run2 | 1.0675 | 22 | 0.5765 | 23 | 0.92 | 18 | 0.544 | 17 | 20 |
| NCYU-Run2 | 1.205 | 23 | 0.6665 | 20 | 0.989 | 22 | 0.534 | 21 | 21.5 |
| XMUT-Run1 | 1.3345 | 24 | 0.3825 | 24 | 1.0995 | 23 | 0.2675 | 24 | 23.75 |

Table 4: Comparative results of valence-arousal prediction for both words and phrases.

4 https://github.com/facebookresearch/fastText
Conclusions
This study describes an overview of the IJCNLP 2017 shared task on dimensional sentiment analysis for Chinese phrases, including task design, data preparation, performance metrics, and evaluation results. Regardless of actual performance, all submissions contribute to the common effort to develop dimensional approaches for affective computing, and the individual report in the proceedings provide useful insights into Chinese sentiment analysis.

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.

Acknowledgments
This work was supported by the Ministry of Science and Technology, Taiwan, ROC, under Grant No. MOST 105-2221-E-155-059-MY2 and MOST 105-2218-E-006-028, and the National Natural Science Foundation of China (NSFC) under Grant No.61702443 and No.61762091.

References
Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5:135–146.

5 http://nlp.innobic.yzu.edu.tw/tasksdsa_p/
Rafael A. Calvo and Sunghwan Mac Kim. 2013. Emotions in text: dimensional and categorical models. *Computational Intelligence*, 29(3):527-543.

Shaosheng Cao, Wei Lu, Jun Zhou, and Xiaolong Li. 2017. Investigating chinese word embeddings based on stroke information.

Xinxiong Chen, Lei Xu, Zhiyuan Liu, Maosong Sun, and Huan-Bo Luan. 2015. Joint learning of character and word embeddings. In *Proc. of the 25th International Joint Conference on Artificial Intelligence (IJCAI-15)*, pages 1236–1242.

Munmun De Choudhury, Scott Counts, and Michael Gamon. 2012. Not all moods are created equal! Exploring human emotional states in social media. In *Proc. of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM-12)*, pages 66-73.

Shu-Ling Huang, Yueh-Yin Shih, and Keh-Jiann Chen. 2008. Knowledge representation for comparative constructions in extended-HowNet. *Language and Linguistics*, 9(2):395-413, 2008.

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.

Lun-Wei Ku and Hsin-Hsi Chen. 2007. Mining Opinions from the Web: Beyond Relevance Retrieval. Journal of American Society for Information Science and Technology, *Special Issue on Mining Web Resources for Enhancing Information Retrieval*, 58(12):1838-1850.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Annual Conference on Advances in Neural Information Processing Systems (NIPS-13)*, pages 1-9.

Tomas Mikolov, Greg Corrado, Kai Chen, and Jeffrey Dean. 2013b. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR-2013)*, pages 1-12.

Myriam Munezero, Tuomo Kakkonen, and Calkin S. Montero. 2011. Towards automatic detection of antisocial behavior from texts. In *Proc. of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP) at IJCNLP-II*, pages 20-27.

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.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-14)*, pages 1532-1543.

Jie Ren and Jeffrey V. Nickerson. 2014. Online review systems: How emotional language drives sales. In *Proc. of the 20th Americas Conference on Information Systems (AMCIS-14)*.

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 the 4th International Conference on Affective Computing and Intelligent Interaction (ACII-11)*, pages 121-131.

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 the 20th International Conference on Asian Language Processing (IALP-16)*, pages 156-160.

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 the 54th Annual Meeting of the Association for Computational Linguistics (ACL-16)*, pages 225-230.