Chinese EmoBank: Building Valence-Arousal Resources for Dimensional Sentiment Analysis

LUNG-HAO LEE and JIAN-HONG LI, National Central University
LIANG-CHIH YU, Yuan Ze University

An increasing amount of research has recently focused on dimensional sentiment analysis that represents affective states as continuous numerical values on multiple dimensions, such as **valence-arousal (VA)** space. Compared to the categorical approach that represents affective states as distinct classes (e.g., positive and negative), the dimensional approach can provide more fine-grained (real-valued) sentiment analysis. However, dimensional sentiment resources with valence-arousal ratings are very rare, especially for the Chinese language. Therefore, this study aims to: (1) Build a Chinese valence-arousal resource called Chinese EmoBank, the first Chinese dimensional sentiment resource featuring various levels of text granularity including 5,512 single words, 2,998 multi-word phrases, 2,582 single sentences, and 2,969 multi-sentence texts. The valence-arousal ratings are annotated by crowdsourcing based on the **Self-Assessment Manikin (SAM)** rating scale. A corpus cleanup procedure is then performed to improve annotation quality by removing outlier ratings and improper texts. (2) Evaluate the proposed resource using different categories of classifiers such as lexicon-based, regression-based, and neural-network-based methods, and comparing their performance to a similar evaluation of an English dimensional sentiment resource.

CCS Concepts: • **Computing methodologies** → **Language resources**;

Additional Key Words and Phrases: Valence-arousal prediction, dimensional sentiment analysis, affective computing

ACM Reference format:
Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese EmoBank: Building Valence-Arousal Resources for Dimensional Sentiment Analysis. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 21, 4, Article 65 (January 2022), 18 pages.
https://doi.org/10.1145/3489141

1 INTRODUCTION

Sentiment analysis has emerged as a leading technique for the automatic identification of affective information texts [Pang and Lee 2008; Calvo and D'Mello 2010; Liu 2012; Feldman 2013]. In sentiment analysis, representation of affective states is an essential issue and can be generally divided into categorical and dimensional approaches [Calvo and Kim 2013].

This work is an extended version of our conference paper [Yu et al. 2016] published at NAACL-HLT 2016. This work is partially supported by the Ministry of Science and Technology, Taiwan under grants MOST 108-2218-E-008-017-MY3, MOST 107-2628-E-155-002-MY3 and MOST 110-2628-E-155-002.
Authors' addresses: L.-H. Lee and J.-H. Li, National Central University, No. 300, Zhongda Rd., Zhongli District, Taoyuan City 32001, Taiwan (R.O.C); emails: lhlee@ee.ncu.edu.tw, 105501004@cc.ncu.edu.tw; L.-C. Yu (corresponding author), Yuan Ze University, No. 135, Yuan-Tung Rd., Zhongli District, Taoyuan City 32003, Taiwan (R.O.C); email: lcyu@saturn.yzu.edu.tw.

2375-4699/2022/01-ART65
https://doi.org/10.1145/3489141

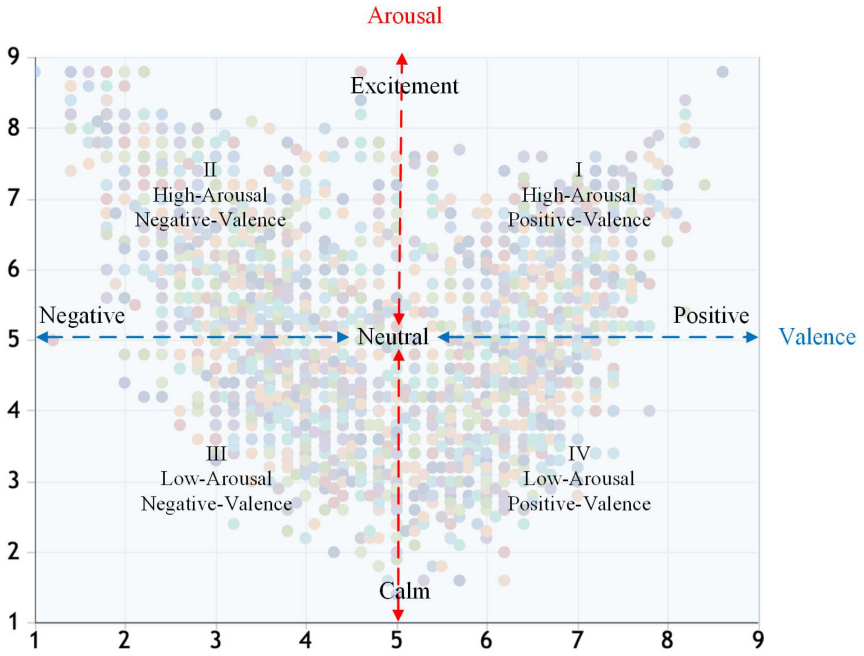

Fig. 1. Two-dimensional valence-arousal space.

The categorical approach represents affective states as several discrete classes such as positive, neutral, and negative, Ekman's six basic emotions (i.e., anger, happiness, fear, sadness, disgust and surprise) [Ekman 1992], and Plutchik's [1991] eight emotions (Ekman's six plus trust and anticipation). The dimensional approach represents affective states as continuous numerical values in multiple dimensions, such as **valence-arousal (VA)** space [Russell 1980], as shown in Figure 1. The valence represents the degree of pleasant and unpleasant (i.e., positive and negative) feelings, while the arousal represents the degree of excitement and calm. Based on this representation, any sentiment expressions can be represented as a point in the VA coordinate plane by recognizing their valence-arousal ratings. Any affective state can be represented as a point in the VA coordinate plane. Applications can benefit from such representation to provide more fine-grained (real-valued) sentiment analysis. For instance, mood analysis systems can identify high risk Twitter users with different mental illnesses because analysis of Twitter posts suggests that depressive users express lower valence and arousal than those with **post-traumatic stress disorder (PTSD)**, and both are lower than control (normal) subjects [Preoţiuc-Pietro et al. 2015]. Product review systems can prioritize high-arousal positive (or high-arousal negative) reviews because recent marketing research suggests that these reviews are usually of interest and could drive purchasing behavior [Ren and Nickerson 2014].

Affective lexicons and corpora with VA ratings are useful resources for the development of sentiment applications. For English, researchers have developed several dimensional lexicons such as the **Affective Norms for English Words (ANEW)** [Bradley and Lang 1999], Extended ANEW [Warriner et al. 2013], and NRC-VAD [Mohammad 2018b], and corpora such as **Affective Norms for English Text (ANET)** [Bradley and Lang 2007], Facebook posts [Preoţiuc-Pietro et al. 2016], and EmoBank [Buechel and Hahn 2017]. For Chinese, dimensional sentiment resources are very rare, including only one small lexicon of 162 words [Wei et al. 2011] and no corpora.

Therefore, this study focuses on building a Chinese valence-arousal resource named Chinese EmoBank, the first Chinese dimensional sentiment resource featuring various levels of text granularity including words, phrases, sentences and multi-sentence texts. Chinese EmoBank consists of two lexicons called **Chinese valence-arousal words (CVAW)** and **Chinese valence-arousal phrases (CVAP)** and two corpora called **Chinese valence-arousal sentences (CVAS)** and **Chinese valence-arousal texts (CVAT)**. The CVAW contains 5,512 single words collected from two polarity-based sentiment lexicons, the **Chinese LIWC (C-LIWC)** [Huang 2012] and NTUSD [Ku and Chen 2007]. The CVAP contains 2,998 multi-word phrases where each phrase is composed of an affective word in the CVAW and a set of modifiers (e.g., negator, degree adverb, and modal) that modify the affective word. The CVAS contains 2,582 single sentences selected from the Twitter microblogging and social networking service. The CVAT contains 2,969 multi-sentence texts extracted from web forums, reviews, and news articles. The annotation of VA ratings is accomplished by crowdsourcing based on the **Self-Assessment Manikin (SAM)** rating scale [Bradley and Lang 1994]. A corpus cleanup procedure is also used to improve annotation quality by removing outlier ratings and improper texts. To further demonstrate the feasibility of the constructed resource, we evaluate it using different categories of classifiers such as lexicon-based, regression-based, and neural-network-based methods, and compare their performance to a similar evaluation of an English dimensional sentiment resource.

The rest of this paper is organized as follows. Section 2 introduces existing lexicons, corpora, and prediction methods for dimensional sentiment analysis. Section 3 describes the process of building Chinese EmoBank. Section 4 presents the analysis results and feasibility evaluation. Conclusions are finally drawn in Section 5.

2 RELATED WORK

This section presents existing single-dimension and multi-dimensions sentiment lexicons and corpora, followed by a description of automatic methods for dimensional score prediction at the word-, phrase- and sentence-levels.

2.1 Dimensional Sentiment Resources

Table 1 presents the language resources for dimensional sentiment analysis. A number of one-dimensional sentiment lexicons provide sentiment intensity or strength of words, including SentiWordNet [Baccianella et al. 2010], SentiFul [Neviarouskaya et al. 2011], SO-CAL [Taboada et al. 2011], AFINN [Nielsen 2011], SentiStrength [Thelwall et al. 2012], and VADER [Hutto and Gilbert 2014]. Specifically, NRC-EIL provides sentiment intensity for eight emotions [Mohammad 2018a]. The SemEval and WASSA shared tasks also released several datasets for single words, multi-word phrases [Rosenthal et al. 2015; Kiritchenko et al. 2016], and sentences [Cortis et al. 2017; Mohammad and Bravo-Marquez 2017; Mohammad et al. 2018]. Stanford Sentiment Treebank [Socher et al. 2013] provided fully labeled parse trees containing sentiment scores at both the phrase- and sentence-levels.

Among multi-dimensional resources, ANEW is the first three-dimensional lexicon providing real-valued scores for the valence, arousal, and dominance dimensions [Bradley and Lang 1999]. ANEW has been extended from 1,034 words to 13,915 words [Warriner et al. 2013]. NRC-VAD provides 20,007 English words with valence, arousal, and dominance ratings [Mohammad 2018b]. In addition to lexicon resources, several multi-dimensional corpora have been proposed. ANET is the first three-dimensional corpus providing valence, arousal, and dominance ratings [Bradley and Lang 2007]. A corpus of 2,895 Facebook posts [Preoţiuc-Pietro et al. 2016] was annotated to provide two-dimensional valence and arousal ratings. EmoBank [Buechel and Hahn 2017] provides

Table 1. Language Resources for Dimensional Sentiment Analysis

Lexicon	Granularity	Size	Scale	Dimension
SentiWordNet [Baccianella et al. 2010]	Word	147,306	Continuous [0, 1]	Valence
SentiFul [Neviarouskaya et al. 2011]	Word	12,900	Continuous [0, 1]	Valence
SO-CAL [Taboada et al. 2011]	Word	5,042	Multi-point [−5, 5]	Valence
AFINN [Nielsen 2011]	Word	2,477	Multi-point [−5, 5]	Valence
SentiStrength [Thelwall et al. 2012]	Word	2,609	Multi-point [−4, 4]	Valence
VADER [Hutto and Gilbert 2014]	Word	7,520	Continuous [−4, 4]	Valence
NRC-EIL [Mohammad 2018a]	Word	9,921	Continuous [0, 1]	Valence for Eight emotions
SemEval 2015 Task 10 [Rosenthal et al. 2015]	Word/Phrase	1,515 (subtask E)	Continuous [0, 1]	Valence
SemEval 2016 Task 7 [Kiritchenko et al. 2016]	Word/Phrase	3,207 (subtask 1)	Continuous [−1, 1]	Valence
SST [Socher et al. 2013b]	Sentence	11,855	Continuous [0, 1]	Valence
SemEval-2017 Task 5 [Cortis et al. 2017]	Tweets (subtask 1) Headlines (subtask 2)	2,510 (subtask 1) 1,647 (subtask 2)	Continuous [−1, 1]	Valence
WASSA-2017 [Mohammad and Bravo-Marquez 2017]	Tweets	7,097	Continuous [0, 1]	Valence for four emotions
SemEval-2018 Task 1 [Mohammad et al. 2018]	Tweets	12,634 (EI-reg) 2,567 (V-reg)	Continuous [0, 1]	Valence for four emotions
ANEW [Bradley and Lang 1999]	Word	1,034	Continuous [1,9]	Valence, Arousal, Dominance
Extended ANEW [Warriner et al. 2013]	Word	13,915	Continuous [1,9]	Valence, Arousal, Dominance
NRC-VAD [Mohammad 2018b]	Word	20,007	Continuous [0, 1]	Valence, Arousal, Dominance
ANET [Bradley and Lang 2007]	Text	120	Continuous [1,9]	Valence, Arousal, Dominance
Facebook posts [Preoţiuc-Pietro et al. 2016]	Sentence	2,895	Continuous [1,9]	Valence, Arousal
EmoBank [Buechel and Hahn 2017]	Sentence	10,062	Continuous [1,9]	Valence, Arousal, Dominance

10,062 sentences with valence, arousal, and dominance ratings. All of the above multi-dimensional lexicons and corpora are scored from 1 to 9.

2.2 Dimension Score Prediction

The above dimensional sentiment resources have been used for dimension score prediction at the word-, phrase-, and sentence-levels. These approaches can be categorized as lexicon-based [Paltoglou et al. 2013], regression-based [Wei et al. 2011, Malandrakis et al. 2013; Paltoglou and Thelwall 2013; Amir et al. 2015; Wang et al. 2016a; 2016b], and neural-network-based models [Du and Zhang 2016; Vilares et al. 2016; Wu et al. 2017; Goel et al. 2017; Zhu et al. 2019; Yu et al. 2020; Wang et al. 2020; Huang et al. 2020].

Lexicon-based methods typically determine the sentiment score of a text by averaging the sentiment scores of the words in the text [Paltoglou et al. 2013]. Regression-based methods have been intensively studied for dimension score prediction. Wei et al. [2011] proposed a cross-lingual approach that trained a linear regression model using the dimension scores of a set of English seed words (source) and their translated Chinese seed words (target). Wang et al. [2016a] further extended their work using a locally weighted linear regression model. Malandrakis et al. [2013] built a linear regression model using n-grams with sentiment scores as features. Both Paltoglou and Thelwall [2013] and Amir et al. [2015] used **support vector regression (SVR)**. Wang et al. [2016b] developed a community-based weighted graph model that performed the regression task on a graph using a social network method to predict the dimension scores of words.

Recently, deep neural network models with word embeddings [Mikolov et al. 2013a; 2013b; Pennington et al. 2014; Bojanowski et al. 2017] or sentiment embeddings [Tang et al. 2016; Yu et al. 2018] have been widely applied to dimensional score prediction. Du and Zhang [2016] used a boosted neural network trained on character-enhanced word embeddings to predict the dimension scores of words. Vilares et al. [2016] used a CNN trained on Twitter word embeddings to determine the sentiment of tweets from highly negative to highly positive using a five-point scale. Wu et al. [2017] introduced a densely connected deep LSTM model to concatenate features at different levels to predict the dimension scores of both words and phrases. Goel et al. [2017] presented an ensemble of different neural networks to determine the intensity level for different emotion categories such as anger, fear, joy, and sadness. Zhu et al. [2019] presented an adversarial attention network to predict the dimension scores of short texts. Yu et al. [2020] proposed a pipelined neural network model to sequentially learn word intensity and modifier weights for phrase-level sentiment intensity prediction. Wang et al. [2020] developed a regional CNN-LSTM model that integrates both local (regional) information within sentences and long-distance dependency across sentences to predict the dimension scores of long texts. Huang et al. [2020] incorporated a context-dependent sentiment lexicon into a 3-channel CNN to predict the strength of both words and texts.

3 THE CHINESE EMOBANK CONSTRUCTION

This section describes the process of building the Chinese EmoBank including the CVAW, CVAP, CVAS, and CVAT.

3.1 Data Collection

The words in CVAW are collected from two polarity-based sentiment lexicons, C-LIWC [Huang 2012] and NTUSD [Ku and Chen 2007]. These affective words are then combined with a set of modifiers such as negators (e.g., *not*), degree adverbs (e.g., *very*), and modals (e.g., *should*) to form multi-word phrases, i.e., the CVAP. The frequency of each phrase is then retrieved from a large web-based corpus. Only phrases with a frequency greater than or equal to 3 are retained as candidates. To prevent several modifiers from dominating the whole resource, each modifier (or modifier

combination) can have at most 50 phrases. In addition, the phrases are selected to maximize the balance between positive and negative words, which helps prevent positive or negative words from dominating the whole resource. The balance is accomplished by, for example, randomly selecting 25 positive words and 25 negative words to constitute the 50 phrases for each modifier. Finally, a total of 2,998 multi-word phrases are selected into the CVAP. To build the CVAS, we first collect Chinese tweets from the social networking service Twitter, selecting tweets that contain the greatest number of affective words found in the CVAW. The tweets are then split into sentences using existing punctuation for a total of 2,582 single sentences excluding emoticons, URLs, and abusive language. For the CVAT, we collect web texts from six different categories: news articles, political discussion forums, car discussion forums, hotel reviews, book reviews, and laptop reviews. Texts containing incomplete semantics and abusive language are excluded. Finally, a total of 2,969 multi-sentence texts containing the greatest number of affective words found in the CVAW were selected for annotation.

3.2 Annotation Details

The annotation of VA ratings is accomplished by crowdsourcing. For the CVAW, each word is randomly assigned to five annotators for rating, and each instance of CVAP, CVAS, and CVAT is randomly assigned to 10 annotators. The number of annotators used for the CVAW is lower than that of CVAP, CVAS, and CVAT because the word-level rating is relatively easier to determine than at the phrase-, sentence-, and multi-sentence levels.

The annotation platform is implemented with the SAM rating scale [Bradley and Lang 1994] using the Google app engine. Figure 2 shows an example of the annotation screen. The top part of Figure. 2 presents an example sentence in the CVAT, followed by the picture-oriented SAM rating scale. Both the valence and arousal dimensions use a nine-degree scale. Value 1 on the valence and arousal dimensions respectively denotes extremely high-negative and low-arousal sentiment, while 9 denotes extremely high-positive and high-arousal sentiment, and 5 denotes a neutral and medium-arousal sentiment. The picture-oriented protocol can help annotators determine the VA ratings more precisely. Volunteer annotators use the annotation screen to provide the VA ratings for each instance in the CVAW, CVAP, CVAS, and CVAT.

3.3 Corpus Cleanup

Once the annotation process is finished, a cleanup procedure is performed to remove outlier ratings and improper instances (e.g., those containing abusive or vulgar language). Outliers are identified if they do not fall into the interval of the mean plus/minus 1.5 **standard deviation (SD)**. They are then excluded from the calculation of the average VA ratings for each instance in the constructed Chinese EmoBank. Table 2 shows the annotation results of the example sentence presented in Figure 2. For the valence dimension, the rating 8 provided by annotator A10 is marked as an outlier because it exceeds the mean plus 1.5 standard deviation. Similarly, the rating 2 in the arousal dimension provided by annotator A1 is also marked as an outlier. After excluding outlier ratings, the (mean, SD) of the example sentence for the valence dimension is (5.778, 0.416) and that for the arousal dimension is (5.333, 0.667). To use the Chinese EmoBank to train a prediction model, the standard deviation is a useful metric to exclude instances with inconsistent annotations. For example, a previous study suggested excluding instances with a standard deviation higher than 2 [Paltoglou et al. 2013].

4 RESULTS AND EVALUATION

This section presents the results of the annotation statistics, a visualization of the corpus, and an evaluation of dimension score prediction.

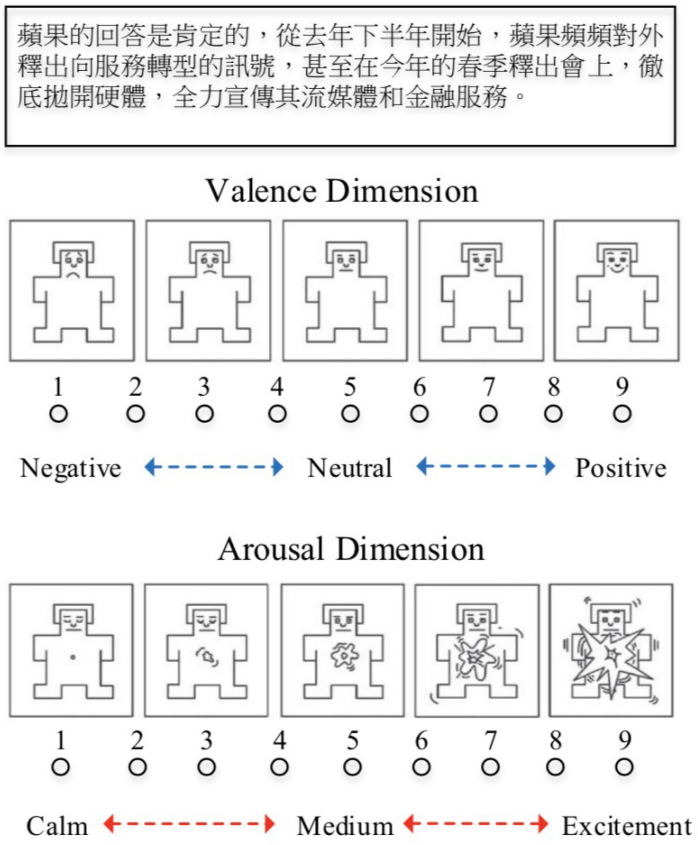

Fig. 2. Annotation screen with the modified 9-point SAM rating scale.

4.1 Results of the Chinese EmoBank

Table 3 shows the mean and standard deviation of the VA ratings for the Chinese EmoBank. The results show that the standard deviation of the arousal dimension is greater than that of valence, indicating that the arousal ratings are more difficult to be determined by the annotators.

Figure 3 shows the scatter plot of the CVAW. It presents a smile curve, indicating that both high-positive and high-negative words usually have a high arousal value. Table 4 lists several example words for the four quadrants of the VA plane.

For the CVAP, a total of 52 modifiers (including 4 negators, 42 degree adverbs, and 6 modals - see Table 5) are combined with the affective words in the CVAW to form the multi-word phrases. Table 6 shows the distribution of different phrase patterns in the CVAP. Table 7 lists an example phrase for each pattern. Figure 4 shows the scatter plot of the CVAP.

Table 8 shows the distribution of the text categories along with their sentence and word counts in CVAS and CVAT. The CVAS is collected from Twitter alone and the CVAT is collected from web texts based on six different categories where News is the major class (50.83%). Figures 5 and 6 respectively show the scatter plot of single sentences and multi-sentence texts in the CVAS and CVAT. Both scatter plots present a smile curve, which is consistent with that of the CVAW and CVAP. Tables 9 and 10 respectively list several example sentences for the CVAS and CVAT.

Table 2. Example of Corpus Cleanup

Sentence	蘋果的回答是肯定的，從去年下半年開始，蘋果頻頻對外釋出向服務轉型的訊號，甚至在今年的春季釋出會上，徹底拋開硬體，全力宣傳其流媒體和金融服務。 (Apple Inc. gave the affirmative response. Starting from the second half of last year, Apple Inc. had frequently shown signs of its service transformations. Even at the release meeting this spring, it completely put aside hardware and fully promoted its streaming media and financial services.)											
Annotator	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10	Mean	SD
Valence	5	5	6	6	6	6	6	6	6	8*	5.778	0.416
Arousal	2*	6	4	5	5	5	5	6	6	6	5.333	0.667

*denotes an outlier.

Table 3. Annotation Statistics of the Chinese EmoBank

	Number of Instances	Valence		Arousal	
		Mean	SD	Mean	SD
CVAW	5,512	4.540	0.717	5.023	1.276
CVAP	2,998	4.594	0.451	5.617	0.561
CVAS	2,582	4.637	0.410	4.967	1.035
CVAT	2,969	4.803	0.664	4.845	1.084

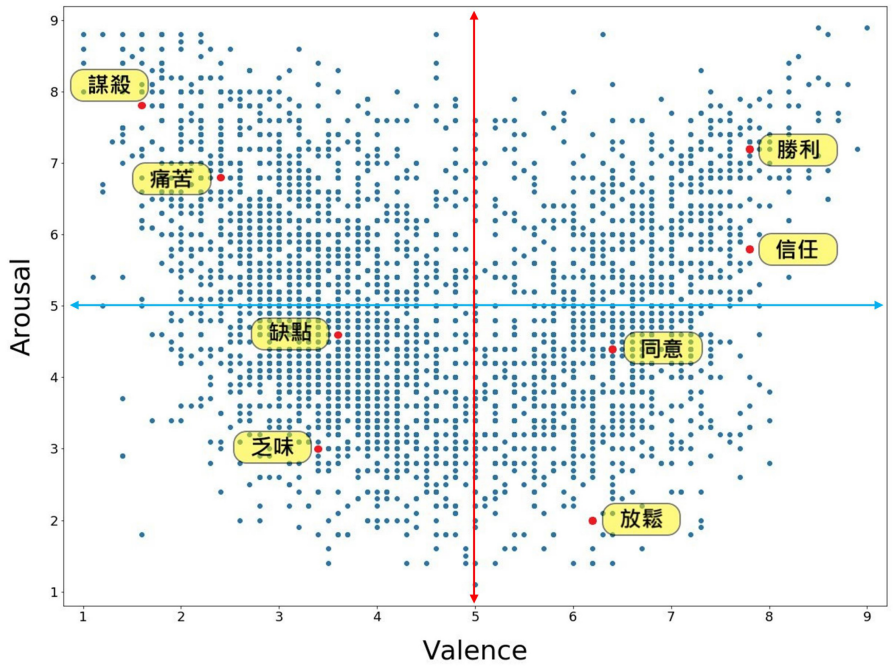

Fig. 3. Scatter plot of the CVAW lexicon.

Table 4. Annotated Examples of the CVAW

Word	Valence		Arousal	
	Mean	SD	Mean	SD
勝利 (victory)	7.8	0.748	7.2	1.166
信任 (trust)	7.8	0.748	5.8	1.470
痛苦 (pain)	2.4	0.490	6.8	0.748
謀殺 (murder)	1.6	0.490	7.8	0.748
乏味 (tedious)	3.4	0.800	3.0	1.414
缺點 (fault)	3.6	0.490	4.6	1.356
同意 (agree)	6.4	0.490	4.4	1.200
放鬆 (relaxed)	6.2	0.748	2.0	0.894

Table 5. Modifier Set Used in the CVAP

Modifier Type	Modifier
Negator	不, 不能, 沒, 沒有
Degree Adverb	有點, 稍, 稍許, 稍稍, 稍微, 略, 略微, 還, 蠻, 愈, 越, 越加, 越發, 好, 老, 怪, 尤其, 較, 較為, 比較, 完全, 更, 更加, 更為, 非常, 挺, 很, 太, 相當, 十分, 格外, 特別, 異常, 最, 最為, 無比, 超, 超級, 極其, 極度, 極為, 萬分
Modal	也許, 可能, 本來, 應該, 本該, 本能

Table 6. Distribution of the Phrase Pattern in the CVAP

Phrase Type	Pattern	Number (ratio%)
2-word phrases	Negator + Word	181 (6.04%)
	Degree Adverb + Word	1,160 (38.69%)
	Modal + Word	143 (4.77%)
3-word phrases	Negator + Degree Adverb + Word	373 (12.44%)
	Degree Adverb + Negator + Word	646 (21.55%)
	Modal + Negator + Word	151 (5.05%)
	Modal + Degree Adverb + Word	323 (10.77%)
	Degree Adverb + Modal + Word	21 (0.70%)
Total	All	2,998 (100%)

4.2 Valence-Arousal Rating Prediction

To demonstrate the application of the constructed affective resources, this section evaluates the performance of the lexicon-based, regression-based, and neural-network-based methods for the valence-arousal rating prediction of the affective corpora.

Table 7. Annotated Examples of the CVAP

Phrase	Pattern	Valence		Arousal	
		Mean	SD	Mean	SD
不喜歡 (do not like)	Negator + Word	3.033	0.481	5.788	0.605
有點冷 (a little cold)	Degree Adverb + Word	3.333	0.471	4.375	0.484
應該開心 (should be happy)	Modal + Word	5.986	0.242	5.350	0.456
沒有太難過 (not so sad)	Negator + Degree Adverb + Word	4.478	0.413	4.675	0.538
完全不同意 (totally not degree)	Degree Adverb + Negator + Word	2.600	0.490	6.833	0.687
應該不嚴重 (should not be serious)	Modal + Negator + Word	5.750	0.433	4.611	0.809
可能更喜歡 (maybe more like)	Modal + Degree Adverb + Word	7.075	0.268	6.444	0.497

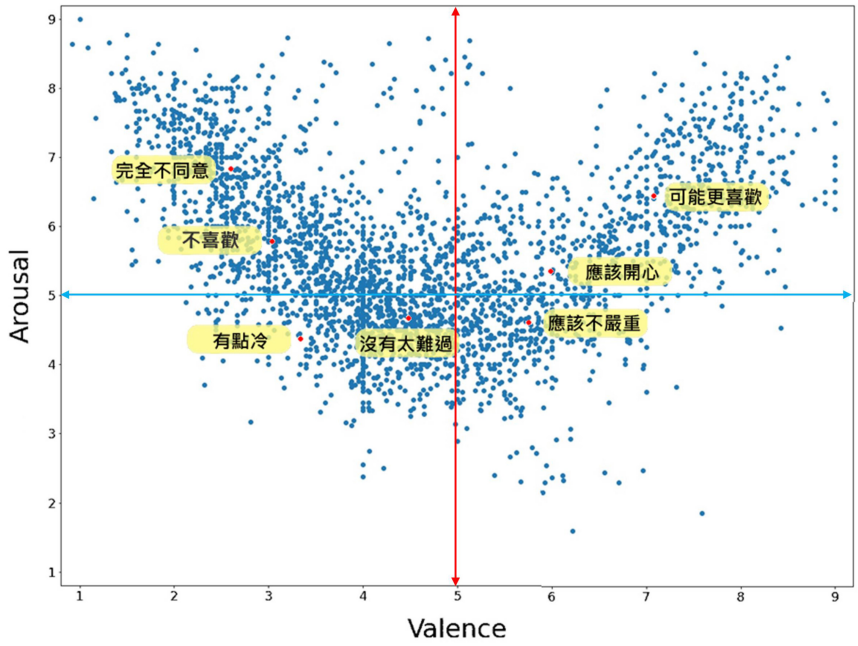

Fig. 4. Scatter plot of the CVAP.

This experiment used three affective corpora. (i) **EmoBank** [Buechel and Hahn 2017] contains 10,062 multi-sentence texts. Each text was rated with individual dimensions (valence/arousal/dominance) in the range of (1, 5). (ii) **Chinese valence-arousal single-sentences (CVAS)** and (iii) **Chinese valence-arousal multi-sentence texts (CVAT)** are our constructed

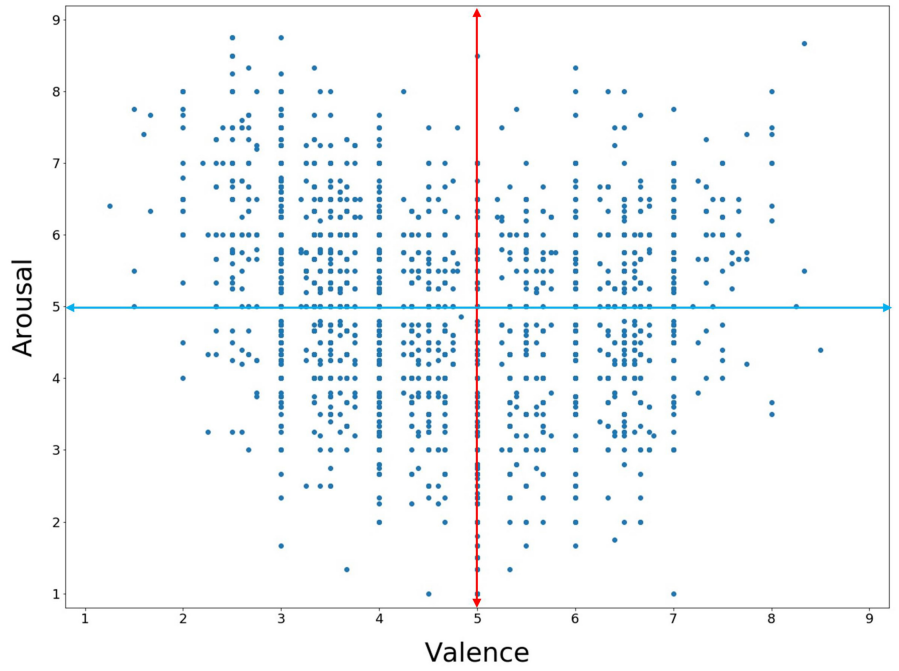

Fig. 5. Scatter plot of the CVAS corpus.

Table 8. Distribution of Text Categories in the CVAS and CVAT

	Category	Num. of Texts (ratio%)	Num. of Sentences	Num. of Words	Avg. Words
CVAS	Twitter	2,582 (100%)	2,582	18,383	7.12
CVAT	Book Review	287 (9.67%)	1,007	6,958	6.91
	Car Forum	253 (8.52%)	859	11,124	12.95
	Hotel Review	299 (10.07%)	1,001	6,101	6.10
	Laptop Review	182 (6.13%)	738	4,538	6.15
	Politics Forum	439 (14.78%)	1,717	13,420	7.82
	News Article	1,509 (50.83%)	6,771	50,096	7.40
	Total	2,969 (100%)	12,093	92,237	7.63

affective corpora with VA ratings. The former is collected from Twitter. The latter consists of texts collected from six categories, including book reviews, car reviews, hotel reviews, laptop reviews, political commentary, and news.

The following methods were compared to demonstrate their performance. (i) **Lexicon-based method** [Paltoglou et al. 2013]: For the EmoBank, the Extended ANEW [Warriner et al. 2013] was used to predict the valence (or arousal) ratings of a given sentence by averaging the valence (or arousal) ratings of the words matched in the Extended ANEW. For both CVAS or CVAT, we used the CVAW and CVAP. (ii) **Regression-based method:** including the **linear regression (LR)** [Wei et al. 2011] and SVR [Paltoglou and Thelwall 2013; Amir et al. 2015]. (iii) **Neural-Network-based method:** including CNN, RNN, LSTM, attention LSTM [Yang et al. 2016], BERT [Devlin et al. 2018], and XLNet [Yang et al. 2019].

Experimental settings are described as follows. We used a 5-fold cross validation technique to evaluate the effectiveness of the above methods. In addition, the suggested default parameters

Table 9. Annotated Examples of the CVAS

Sentence	Valence		Arousal	
	Mean	SD	Mean	SD
他真的是萬能的 (He is really multitalented.)	8.333	0.471	5.500	1.118
無一例外都離婚了 (Everyone of them is divorced.)	2.667	0.943	6.333	0.471
這趟旅程真的不怎麼令人盡興 (This trip is really not very exciting.)	3.000	0.816	4.667	0.943
坦誠地與美方交換意見 (A sincere exchange of opinions with the U.S.)	6.400	0.489	3.750	0.829

Table 10. Annotated Examples of the CVAT

Text	Valence		Arousal	
	Mean	SD	Mean	SD
CPU 顯卡也完全夠用，接口也非常齊全，總體來說很滿意！ (The CPU graphic card is also fully functional, and the interface is also very complete. Generally speaking, I am very satisfied!)	7.143	0.350	5.000	0.816
房間裏黴味，煙味撲鼻，沒有窗戶通風，骯髒的地毯上的斑斑點點的污蹟，令人觸目驚心。 (The room smelled musty and smoky, with no windows for ventilation, and the stained carpet was shockingly dirty.)	1.889	0.737	6.875	0.927
很多車主抱怨新車怠速抖動嚴重----冷車時更嚴重。 (Many car owners complained that the new car has a serious idle jitter, and this is more pronounced when starting the car in cold weather.)	3.250	1.090	5.667	1.054
飛安帶來更多保障，也提供旅客更安心的服務品質。 (Improved flight safety provides passengers with greater confidence in service quality.)	7.000	0.535	4.222	1.133

shown in Table 11 were selected without further fine-tuning. The word vectors for English and Chinese were trained using the BERT technique [Devlin et al. 2018]. Pre-trained models with whole word masking were downloaded from official BERT GitHub website. For English, we used BERT-Large, Cased (24-layer 1024-hidden, 16-heads, 340 parameters). The dimensionality is 1024. For Chinese, we used BERT-Base, Chinese Simplified, and Traditional (12-layer, 768-hidden, 12-heads, 110 parameters). The dimensionality is 768. The pre-trained XLNet[1] and BERT[2] models are publicly available online for evaluation.

Performance was evaluated using the **mean absolute error (MAE)** and Pearson correction coefficient (r), defined as the follows

[1]XLNet-mid, Chinese: https://github.com/ymcui/Chinese-XLNet, XLNet-Base, Cased: https://github.com/zihangdai/xlnet.
[2]BERT-wwm, Chinese: https://github.com/ymcui/Chinese-BERT-wwm, BERT-Base, Cased: https://github.com/google-research/bert.

Table 11. Hyper-parameters Used in the Classifiers

Methods	CNN	RNN, LSTM	Attention	XLNet	BERT
Filter Number	60	–	60	–	–
Filters Length	3	–	3	–	–
Pool Length	2	–	2	–	–
Hidden State Dim.	–	120	120	120	120
Layer Number	–	–	–	24 (Cht.) 12 (Eng.)	12
Hidden Number	–	–	–	768	768
Head Number	–	–	–	12	12
Optimizer	Adam				
Batch Size	32				
(Recurrent) Dropout	0.25				
Epoch	20				

- Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|a_i - p_i| \tag{1}$$

- Person Correlation Coefficient (r):

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{a_i - \mu_A}{\sigma_A}\right)\left(\frac{p_i - \mu_P}{\sigma_A}\right) \tag{2}$$

where $a_i \in A$ and $p_i \in P$ respectively denote the i-th actual value and predicted value, n is the number of test samples, μ_A and σ_A respectively represent the mean value and the standard deviation of A, while μ_P and σ_P respectively represent the mean value and the standard deviation of P. The MAE measures the error rate, and the r measures the linear correlation between the actual values and the predicted values. A lower MAE and a higher r indicate more accurate prediction performance.

Tables 12, 13, and 14 respectively show the prediction results for CVAP, CVAS and CVAT. All three datasets produce nearly consistent findings. The lexicon-based method provides baseline results. For both regression-based methods, the SVR approach outperformed the LR in both the valence or arousal dimensions. The BERT model outperformed the other neural-network-based methods in all dimensions. For CVAP, the Attention model underperformed the LSTM, possibly because phrases are usually very short with one or two modifiers to a word, thus raising challenges for the attention mechanism. Comparing the results achieved by our constructed CVAS and CVAT, all models on the CVAS clearly underperformed the corresponding model results on the CVAT. Based on our observations, the CVAS data containing single sentences from Twitter has more difficulty predicting valence-arousal ratings than multi-sentence texts that provide more information in CVAT. Table 15 also shows the EmoBank results for reference. The consistent conclusions confirm the reliability of our constructed Chinese EmoBank corpus.

5 CONCLUSIONS AND FUTURE WORK

This study constructs a language resource, the Chinese Emobank, annotated with valence-arousal ratings for dimensional sentiment analysis. The Chinese EmoBank presents a Chinese affective lexicon with 5,512 single words (CVAW) and 3,000 multi-word phrases (CVAP), and a

Table 12. Comparative Results of Different Methods in CVAP

CVAP (Chinese)			
Valence		MAE	*r*
Lexicon	CVAW	1.468	0.467
Regression	LR	1.541	0.286
	SVR	1.355	0.290
Neural Networks	CNN	0.735	0.821
	RNN	0.630	0.861
	LSTM	0.613	0.864
	Attention	0.633	0.853
	XLNet	0.472	0.932
	BERT	0.339	0.965
Arousal		MAE	*r*
Lexicon	CVAW	0.794	0.660
Regression	LR	1.071	0.297
	SVR	0.862	0.413
Neural Networks	CNN	0.559	0.804
	RNN	0.509	0.837
	LSTM	0.492	0.846
	Attention	0.522	0.824
	XLNet	0.484	0.855
	BERT	0.410	0.902

Table 13. Comparative Results of Different Methods in CVAS

CVAS (Chinese)			
Valence		MAE	*r*
Lexicon	CVAW/CVAP	0.940	0.593
Regression	LR	1.079	0.493
	SVR	0.886	0.612
Neural Networks	CNN	0.920	0.564
	RNN	0.909	0.573
	LSTM	0.871	0.602
	Attention	0.857	0.621
	XLNet	0.680	0.766
	BERT	0.656	0.788
Arousal		MAE	*r*
Lexicon	CVAW/CVAP	1.214	0.266
Regression	LR	1.183	0.264
	SVR	0.954	0.414
Neural Networks	CNN	0.975	0.364
	RNN	0.964	0.373
	LSTM	0.946	0.429
	Attention	0.943	0.432
	XLNet	0.953	0.428
	BERT	0.926	0.460

Table 14. Comparative Results of Different Methods in CVAT

CVAT (Chinese)			
Valence		MAE	r
Lexicon	CVAW/CVAP	0.928	0.621
Regression	LR	0.791	0.701
	SVR	0.710	0.760
Neural Networks	CNN	0.814	0.665
	RNN	0.716	0.740
	LSTM	0.657	0.777
	Attention	0.621	0.802
	XLNet	0.531	0.855
	BERT	0.509	0.871
Arousal		MAE	r
Lexicon	CVAW/CVAP	0.871	0.279
Regression	LR	0.833	0.423
	SVR	0.716	0.530
Neural Networks	CNN	0.799	0.396
	RNN	0.738	0.493
	LSTM	0.715	0.534
	Attention	0.696	0.553
	XLNet	0.701	0.559
	BERT	0.690	0.581

Table 15. Comparative Results of Different Methods in EmoBank

EmoBank (English)			
Valence		MAE	r
Lexicon	Extended ANEW	0.783	0.406
Regression	LR	0.395	0.675
	SVR	0.369	0.689
Neural Networks	CNN	0.436	0.606
	RNN	0.389	0.655
	LSTM	0.357	0.724
	Attention	0.347	0.743
	XLNet	0.352	0.696
	BERT	0.324	0.766
Arousal		MAE	r
Lexicon	Extended ANEW	1.095	0.175
Regression	LR	0.362	0.488
	SVR	0.342	0.520
Neural Networks	CNN	0.390	0.434
	RNN	0.352	0.494
	LSTM	0.347	0.526
	Attention	0.340	0.546
	XLNet	0.355	0.452
	BERT	0.334	0.569

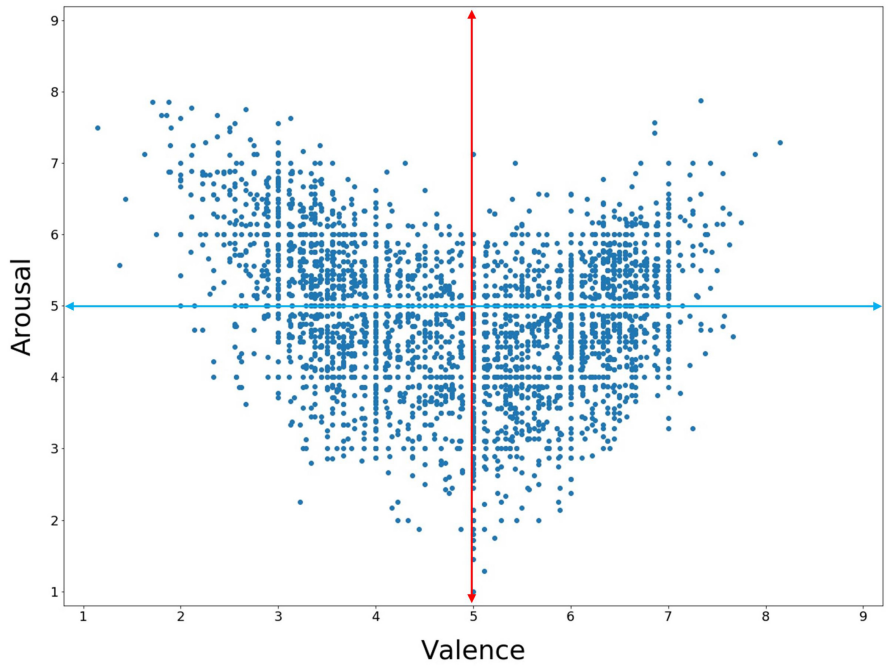

Fig. 6. Scatter plot of the CVAT corpus.

Chinese affective corpus of 2,582 single sentences (CVAS) and 2,969 multi-sentence texts (CVAT) with six different categories, all annotated with valence-arousal values. A cleanup procedure removed outlier ratings and improper texts to improve annotation quality. Experimental results provide a feasibility evaluation and baseline performance for VA prediction using the constructed resources.

Future work will focus on developing advanced VA prediction methods and building useful dimensional sentiment applications based on the constructed resources. For example, Figures 3–6 show that the valence and arousal dimensions may correlate with each other. It is worth investigating how relations between dimensions can be integrated into the prediction model to enhance performance. Finally, we will release the entire Chinese EmoBank with fully annotated valence-arousal ratings to facilitate future development in related research areas.

REFERENCES

S. Amir, R. F. Astudillo, W. Ling, B. Martins, M. Silva, and I. Trancoso. 2015. INESC-ID: A regression model for large scale Twitter sentiment lexicon induction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15).*

S. Baccianella, A. Esuli, and F. Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10).* 2200–2204.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics* 5 (2017), 135–146.

M. M. Bradley and P. J. Lang. 1994. Measuring emotion: The Self-Assessment Manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry* 25, 1 (1994), 49–59.

M. M. Bradley and P. J. Lang. 1999. Affective norms for English words (ANEW): Instruction manual and affective ratings. *Technical Report C-1,* University of Florida, Gainesville, FL.

M. M. Bradley and P. J. Lang. 2007. Affective norms for English text (ANET): Affective ratings of text and instruction manual. *Technical Report D-1,* University of Florida, Gainesville, FL.

S. Buechel and U. Hahn. 2017. EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL'17)*. 578–585.

R. A. Calvo and S. D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing* 1, 1, (2010), 18–37.

R. A. Calvo and S. M. Kim. 2013. Emotions in text: Dimensional and categorical models. *Computational Intelligence* 29, 3 (2013), 527–543.

K. Cortis, A. Freitas, T. Daudert, M. Huerlimann, M. Zarrouk, S. Handschuh, and B. Davis. 2017. SemEval-2017 Task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17)*. 519–535.

J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

S. Du and X. Zhang. 2016. Aicyber's system for IALP 2016 shared task: Character-enhanced word vectors and boosted neural networks. In *Proceedings of the 2016 International Conference on Asian Language Processing (IALP'16)*. 161–163.

P. Ekman. 1992. An argument for basic emotions. *Cognition and Emotion* 6 (1992), 169–200.

R. Feldman. 2013. Techniques and applications for sentiment analysis. *Communications of the ACM* 56, 4 (2013), 82–89.

P. Goel, D. Kulshreshtha, P. Jain, and K. K. Shukla. 2017. Prayas at EmoInt 2017: An ensemble of deep neural architectures for emotion intensity prediction in Tweets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA'17)*. 58–65.

C. L. Huang, C. K. Chung, N. Hui, Y. C. Lin, Y. T. Seih, B. C. P. Lam, and J. W. Pennebaker. 2012. Development of the Chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology* 54, 2 (2012), 185–201.

M. Huang, H. Xie, Y. Rao, J. Feng, and F. L. Wang. 2020. Sentiment strength detection with a context-dependent lexicon-based convolutional neural network. *Information Sciences*. 520 (2020), 389–399.

C. J. Hutto and E. Gilbert. 2014. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of 8th International AAAI Conference on Weblogs and Social Media*. 216–225.

S. Kiritchenko, S. M. Mohammad, and M. Salameh. 2016. SemEval-2016 Task 7: Determining sentiment intensity of English and Arabic phrases. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 42–51.

L. W. Ku and H. H. Chen. 2007. Mining opinions from the web: Beyond relevance retrieval. *Journal of the American Society for Information Science and Technology* 58, 2 (2007), 1838–1850.

B. Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool, Chicago, IL.

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 (2013), 2379–2392.

T. Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems (NIPS'13)*.

T. Mikolov, G. Corrado, K. Chen, and J. Dean. 2013b. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR'13)*.

S. M. Mohammad. 2018a. Word affect intensities. In *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference (LREC'18)*. 174–183.

S. M. Mohammad. 2018b. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL'18)*, 174–184.

S. M. Mohammad and F. Bravo-Marquez. 2017. WASSA-2017 shared task on emotion intensity. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA'17)*, 34–49.

S. M. Mohammad, F. Bravo-Marquez, M. Salameh, and S. Kiritchenko. 2018. Semeval-2018 Task 1: Affect in tweets, In *Proceedings of the 12th International Workshop on Semantic Evaluation (SemEval'18)*. 1–17.

A. Neviarouskaya, H. Prendinger, and M. Ishizuka. 2011. SentiFul: A lexicon for sentiment analysis. *IEEE Transactions on Affective Computing* 2, 1 (2011), 22–36.

F. Å. Nielsen. 2011. A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on Making Sense of Microposts: Big Things Come in Small Packages*.

G. Paltoglou and M. Thelwall. 2013. Seeing stars of valence and arousal in blog posts. *IEEE Transactions on Affective Computing* 4, 1 (2013), 116–123.

G. Paltoglou, M. Theunis, A. Kappas, and M. Thelwall. 2013. Predicting emotional responses to long informal text. *IEEE Transactions on Affective Computing* 4, 1 (2013), 106–115.

B. Pang and L. Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2, 1–2 (2008), 1–135.

J. Pennington, R. Socher, and C. D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'14)*. 1532–1543.

R. Plutchik. 1991. *The Emotions*, Lanham, MD, USA: Univ. Press Amer.

D. Preoţiuc-Pietro, J. Eichstaedt, G. Park, M. Sap, L. Smith, V. Tobolsky, H. A. Schwartz, and L. Ungar. 2015. The role of personality, age, and gender in tweeting about mental illness. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. 21–30.

D. Preoţiuc-Pietro, H. A. Schwartz, G. Park, J. Eichstaedt, M. Kern, L. Ungar, and E. Shulman. 2016. Modelling valence and arousal in Facebook posts. In *Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA'16)*. 9–15.

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

S. Rosenthal, P. Nakov, S. Kiritchenko, S. M. Mohammad, A. Ritter, and V. Stoyanov. 2015. SemEval-2015 Task 10: Sentiment analysis in Twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. 451–463.

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

R. Socher, A. Perelygin, J. Y. Wu, J. Chuang, C. D. Manning, A. Ng, and C. Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP'13)*. 1631–1642.

D. Tang, F. Wei, B. Qin, N. Yang, T. Liu, and M. Zhou. 2016. Sentiment embeddings with applications to sentiment analysis. *IEEE Transactions on Knowledge and Data Engineering* 28, 2 (2016), 496–509.

M. Taboada, J. Brooke, M. Tofiloski, K. Voll, and M. Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational Linguistics* 37, 2 (2011), 267–307.

M. Thelwall, K. Buckley, and G. Paltoglou. 2012. Sentiment strength detection for the social web. *Journal of the Association for Information Science and Technology* 63, 1 (2012), 163–173.

D. Vilares, Y. Doval, M. A. Alonsoa, and C. Gómez-Rodríguez. 2016. LyS at SemEval-2016 Task 4: Exploiting neural activation values for Twitter sentiment classification and quantification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. 79–84.

A. B. Warriner, V. Kuperman, and M. Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* 45, 4 (2013), 1191–1207.

J. Wang, L. C. Yu, K. R. Lai, and X. Zhang. 2016a. Locally weighted linear regression for cross-lingual valence-arousal prediction of affective words. *Neurocomputing* 194, 271–278.

J. Wang, L. C. Yu, K. R. Lai, and X. Zhang. 2016b. Community-based weighted graph model for valence-arousal prediction of affective words. *IEEE/ACM Trans. Audio, Speech and Language Processing* 24, 11 (2016), 1957–1968.

J. Wang, L. C. Yu, K. R. Lai, and X. Zhang. 2020. Tree-structured regional CNN-LSTM model for dimensional sentiment analysis. *IEEE/ACM Transactions on Audio Speech and Language Processing* 28, 581–591.

W. L. Wei, C. H. Wu, and J. C. Lin. 2011. A regression approach to affective rating of Chinese words from ANEW. In *Proceedings of the 4th International Conference on Affective Computing and Intelligent Interaction (ACII'11)*. 121–131.

C. Wu, F. Wu, Y. Huang, S. Wu, and Z. 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 (IJCNLP'17)*. 42–52.

Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL/HLT'16)*, 1480–1489.

Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. *arXiv preprint arXiv:1906.08237*.

L. C. Yu, L. H. Lee, S. Hao, J. Wang, Y. He, J. Hu, K. R. Lai, and X. Zhang. 2016. 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 (NAAC-HLT'16)*. 540–545.

L. C. Yu, J. Wang, K. R. Lai, and X. Zhang. 2018. Refining word embeddings using intensity scores for sentiment analysis. *IEEE/ACM Transactions on Audio Speech and Language Processing* 26, 3 (2018), 671–681.

L. C. Yu, J. Wang, K. R. Lai, and X. Zhang. 2020. Pipelined neural networks for phrase-level sentiment intensity prediction. *IEEE Transactions on Affective Computing* 11, 3 (2020), 447–458.

S. Zhu, S. Li, and G. Zhou. 2019. Adversarial attention modeling for multi-dimensional emotion regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL'19)*. 471–480.

Received January 2021; revised June 2021; accepted September 2021