Inferring Affective Meanings of Words from Word Embedding

Minglei Li, Qin Lu, Yunfei Long, and Lin Gui

Abstract—Affective lexicon is one of the most important resource in affective computing for text. Manually constructed affective lexicons have limited scale and thus only have limited use in practical systems. In this work, we propose a regression-based method to automatically infer multi-dimensional affective representation of words via their word embedding based on a set of seed words. This method can make use of the rich semantic meanings obtained from word embedding to extract meanings in some specific semantic space. This is based on the assumption that different features in word embedding contribute differently to a particular affective dimension and a particular feature in word embedding contributes differently to different affective dimensions. Evaluation on various affective lexicons shows that our method outperforms the state-of-the-art methods on all the lexicons under different evaluation metrics with large margins. We also explore different regression models and conclude that the Ridge regression model, the Bayesian Ridge regression model and Support Vector Regression with linear kernel are the most suitable models. Comparing to other state-of-the-art methods, our method also has computation advantage. Experiments on a sentiment analysis task show that the lexicons extended by our method achieve better results than publicly available sentiment lexicons on eight sentiment corpora. The extended lexicons are publicly available for access.

Index Terms—Affective lexicon, sentiment, emotion, word embedding, regression

1 Introduction

As the Internet and social media are becoming so popular, web text is becoming one of the most important channels for people to express their opinions, mental state, and communicate with each other. Affective meaning refers to emotion, sentiment, personality, mood and attitude expressed through text [1]. In this work, we refer to the term affective to be specific to emotion and sentiment. Affective computing from text has many potential applications, such as the analysis of consumer opinions on a company's products [2], automatic recommendation systems for movies, books, music or pictures based on current user's emotions [3], detection of people who have potential suicide risks based on social media [4], stock market prediction based on public opinions [5], product aspect extraction [6], sarcasm detection [7], personality detection [8], and intelligent human-computer interaction systems that can express and detect the affective states of human beings [9], etc.

The most important resource for affective computing is a comprehensive affective lexicon, in which words are annotated with affective meanings. The affective meaning of a word can be represented using different methods. Earlier works represent affective meanings of words by discrete affective labels, such as *positive*, *negative*, *happiness*, *sadness*, *anger* [10], [11], [12], etc. Another method is to represent affective meaning by the more comprehensive multi-dimensional representation models , such as the valence-arousal-dominance model (VAD) [13] and the evaluation-potency-activity model (EPA) [14]. Theoretically speaking, discrete affective labels can always be mapped to certain points in a multi-dimensional affective space [15]. Sentiment indicated by polarities can be viewed as a one dimensional affective model. For example, it is equal to the valence dimension in VAD or the evaluation dimension in EPA.

Compared to discrete emotion labels or one dimensional sentiment, multi-dimensional affective representation is more comprehensive because it can capture more fine-grained information compared to the discrete and the one dimensional models. According to the Affective Control Theory (ACT), each concept in an event has a transient affective meaning which is context dependent in addition to cultural, behavior and other background information [16]. Multi-dimensional models allow for more interaction between a sequence of words so that more context information can be included in affective computing of text. For example, the same noun *champion* may have different affective state in two different events: *The little boy defeated the champion* and *The champion defeated the little boy*. The difference of the affective states cannot be inferred through single sentiment dimension but it can be distinguished through multi-dimensional EPA affective lexicons based on the ACT [16]. However, multi-dimensional affective lexicons as NLP resources are limited because most available ones are based on manual annotation, such as the ANEW lexicon of VAD based on manual

- *M. Li, Q. Lu, and Y. Long are with the Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong. E-mail: {csmli, csluqin, csylong}@comp.polyu.edu.hk.*
- *L. Gui is with the College of Mathematics and Computer Science, Fuzhou University, Fuzhou, Fujian, China. E-mail: guilin.nlp@gmail.com.*

Manuscript received 19 Jan. 2017; revised 20 June 2017; accepted 28 June 2017. Date of publication 3 July 2017; date of current version 5 Dec. 2017. (Corresponding author: Minglei Li.)
Recommended for acceptance by E. Cambria, A. Hussain, and A. Vinciarelli.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TAFFC.2017.2723012

TABLE 1
List of Popular Discrete Emotion Categorizations

Author	Num	Basic Emotions
Ekman [24]	6	anger, disgust, fear, joy, sadness, surprise
Parrot [25]	6	anger, fear, joy, love, sadness, surprise
Frijda [26]	6	desire, happiness, interest, sorrow, surprise, wonder
Plutchik [26]	8	acceptance, anger, anticipation, disgust, fear, joy, sadness, surprise
Tomkins [27]	9	anger, contempt, disgust, distress, fear, interest, joy, shame, surprise
Ortony [28]	22	fear, joy, distress, happy-for, gloating, hope, pity, pride, relief, resentment, satisfaction, etc.
Xu [29]	7	anger, disgust, fear, joy, like, sadness, surprise

annotation[17], the extended ANEW lexicon based on crowdsourcing [18], the Chinese valence-arousal lexicon based on manual annotation [19], the EPA lexicon based on manual annotation [14]. Obviously, manual annotation is not scalable and it limits the use of multi-dimensional models in real applications. Only if automatic methods can be used to learn the representations of affective meanings of words, the more comprehensive multi-dimensional models can have a wider practical use. Word embedding based graph propagation method is used as an automatic method to predict the valence-arousal ratings from seed words [20]. However, word embedding is normally trained to obtain the general meaning of words, which can include denotative meaning, connotative meaning, social meaning, affective meaning, reflected meaning, collocative meaning and thematic meaning [21]. In other words, directly computing word similarity captures the general meanings of words rather than the affective meanings specifically. Words that have similar denotative meanings may be associated with different affective meanings. For example, *"father"* and *"dad"* have the same denotative meaning, yet they are associated with different affective meanings; *"father"* is more formal and detached whereas *"dad"* is more personal and dear affectively. Another type of automatic method uses regression models to extend affective lexicons. Specific methods include (1) a linear regression model based on manually defined features from a knowledge base[22], which is limited by the manually prepared features; (2) the linear regression weighted on the the semantic similarity between a target word and the seed words[23], which is limited by the accuracy of the semantic similarity.

In this work, we propose a regression method to infer various affective meanings from word embedding based on the assumption that different features in word embedding may contribute differently to a particular affective dimension and one feature in word embedding may also contribute differently to different affective dimensions. The method treats word embedding as word features and learns meaning specific weights to each feature when mapping embedding to different affective dimensions. Consequently, the method learns one regression model for each affective dimension based on the seed words to predict the affective meaning of a new word provided that its word embedding is available. We perform extensive experiments on inferring different affective meanings, including sentiment, valence, arousal, dominance, evaluation, potency, activity, imagery, and also other meanings including perceptual sense of words, concreteness of words. Evaluations show that:

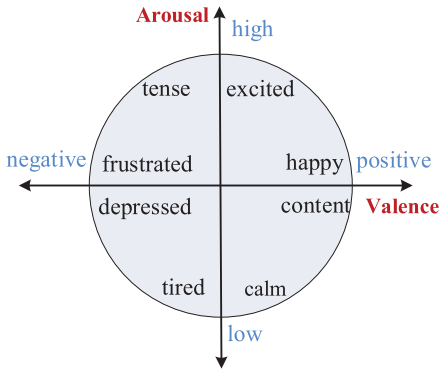

Fig. 1. Two dimensional valence-arousal (VA) affective model.

1) Our method achieves the state-of-the-art performance, outperforms all the baseline methods on several affective lexicons in affective space and lexicons in other semantic space.
2) Our method is rating scale insensitive, which means that our method does not require the rating range to be bipolar and there is no need to transform unipolar ratings to bipolar ratings.
3) Our method is computationally more efficient than the baseline methods, especially compared to propagation based methods.
4) Several affective lexicons with about million of words are built and one experiment using the built sentiment lexicon shows that lexicons based on word embedding perform better than previously available sentiment lexicons.

The rest of the paper is organized as follows: Section 2 describes related works, including affective models, lexicon generation methods, and word embedding models. Section 3 introduces our proposed method for inferring affective meanings. Section 4 performs extensive experiments on various affective lexicons to validate the effectiveness of our proposed method. Section 5 concludes this paper.

2 Related Works

2.1 Affective Model

Affective meaning includes emotion, sentiment, trait, mood, and attitudes, etc. Current research in affective computing mainly studies sentiment and emotion. Sentiment is measured by positive or negative polarities. Emotion can be considered as fine-grained sentiment. Affective meaning can be represented either by discrete categories or a set of values in continuous scales of some multi-dimension models. In the former representation, different categories are proposed. Table 1 lists several proposed emotion categorizations.

There are several multi-dimensional models including the valence-arousal model (VA) [13] as shown in Fig. 1; the evaluation-potency-activity model [30] as shown in Fig. 2; the hourglass model of emotion [31] which represents the affective state in four independent dimensions: pleasantness, attention, sensitivity and aptitude; the Pleasure-Arousal-Dominance

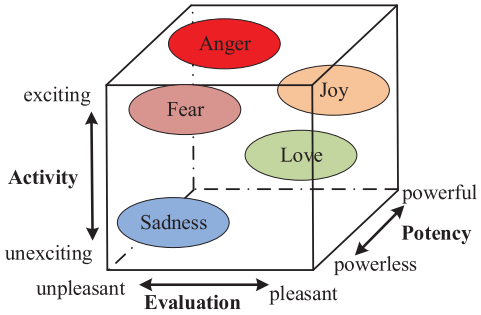

Fig. 2. Three dimensional evaluation-potency-activity (EPA) affective model.

(PAD) [32]; the two continuous dimensions of evaluation and activation [33]; the four dimensions of evaluation-pleasantness, potency-control, activation-arousal, and unpredictability [34]; the three dimensions of serotonin, dopamine and noradrenaline based on neuroscience [35]. Compared to the discrete affective models, the dimensional models can capture more information and are more suitable for computation because the interaction information between different dimensions can be captured.

2.2 Affective Lexicon Generation

Based on the affective models, affective lexicons are built either using a discrete affective model or a dimensional model. In this paper, we will only focus on dimension based lexicons. Since sentiments can be described by a one dimensional model, we also include methods for obtaining sentiment lexicons. Theoretically speaking, methods to obtain a sentiment lexicon can be extended to obtain other affective dimensions.

Affective lexicons can be obtained either by manual annotation or automatic methods. *Manual annotation* can obtain high-quality lexicons. Manually annotated sentiment lexicons include the General Inquirer (GI) [10], MPQA [36], the twitter sentiment lexicon based on crowdsourcing [37], [38], VADER based on crowdsourcing [39], etc. Manually annotated multi-dimensional affective lexicons include ANEW, CVAW, DAL, EPA and ANGST, among others. The ANEW lexicon based on the VAD model [17] which contains 1,034 English words. The extended ANEW lexicon contains about 13,965 English words annotated through crowdsourcing. The CVAW lexicon based on the VAD model [19] contains 1,653 traditional Chinese words annotated in the valence and arousal dimensions. The Dictionary of Affect in Language (DAL) lexicon annotated in the dimensions of pleasantness-activation-imagery contains 8,742 terms [33]. The EPA lexicon annotated in the evaluation-potency-activity dimensions [16] contains about 4,505 English terms. The ANGST lexicon annotated in the valence-arousal-dominance-imageability-potency dimensions contains 1,003 German words [40].

Automatic methods to obtain affective lexicons are focused mainly on the sentiment dimension because current research works are mostly on sentiment analysis [41], [42], [43]. In terms of methodology, there are mainly three approaches. The *first* approach uses statistical information between a target word and the seed words. For example, sentiment polarity intensities are calculated based on point-wise mutual information (PMI) between a target word and the positive seeds and negative seeds, respectively [37], [44]. Similarly, PMI is used to build discrete emotion lexicon based on naturally annotated hashtags in twitter [45]. The *second* approach is based on the label propagation method which first builds a word graph and then label propagation is performed to infer the affective values of unseen words from the seed words. For example, a graph can be built based on the semantic relationship in WordNet and the label propagation is performed to infer the EPA values [46] and sentiment polarity [47]. A knowledge based graph is confined by the coverage of the knowledge base. A word graph can also be built from a text corpus based on the cosine similarity of words represented by their contexts words and then graph propagation is performed to infer the sentiment polarity of unseen words [48]. Word embedding is also used to compute the cosine similarity between words to build the word graph and PageRank algorithm is employed to infer the valence-arousal ratings of unseen words [20]. Similarly, a word graph is constructed using cosine similarity of word embedding to infer sentiment polarities [43]. The *third* approach represents a word as a vector and then map this vector to some sentiment value or categories based on a regression model or a classifier. This approach mainly include (1) representing words by manual defined features based on some knowledge base and performing linear regression on the features [22]; (2) representing words as word embeddings obtained automatically and using a classifier [49] or linear regression [50] to obtain sentiment labels or scores; (3) mapping word embedding into sentiment space through a transformation matrix that minimizes intra-group distance in each sentiment category and maximizes inter-group distance without considering the actual values of the seed words [51].

2.3 Word Representation

In a conventional word representation, a word is first converted to a symbolic ID. Its feature set are then transformed into a vector using a one-hot representation. One-hot encoding is a high dimensional vector representation with only one dimension as 1 and all the other dimensions as 0 for one word, and the dimension size is the size of the vocabulary. This kind of representation cannot capture the semantic relations between different words. Another method is to represent a word using a low dimensional dense vector, also called word embedding, which can encode the semantic meaning of words and thus comparisons can be easily made. For example, using word embedding, we can make the approximation: $vec(king) - vec(queen) = vec(man) - vec(woman)$ [52].

Various approaches have been proposed to learn dense word vectors, which can be divided mainly into count based approaches and prediction based approaches [53], both of which are based on the distributional hypothesis that words occur in similar context tends to have similar meanings [54]. A count based method constructs a word-context co-occurrence statistic matrix and then perform matrix factorization to obtain the final word embedding. Features used include point-wise mutual information, positive point-wise mutual information (PPMI), and log of co-occurrences, etc. Based on the matrix factorization, various algorithms have been proposed, such as decomposition of the matrix into two low dimensional matrices [55], Singular Value Decomposition (SVD) [56], probabilistic matrix and tensor factorization

[57], low rank approximation [58]. The prediction based method directly predicts the context given the target word by maximizing the conditional probability of the context words given the target or vice versa [52]. More comprehensive studies use more kinds of contexts and knowledge base are explored to improve word embedding including the use of cross-lingual context [59], word definition context, knowledge base context [60], morphology context [61], and word embedding from multi-views or multi-resources [62], [63].

3 Proposed Method

In this work, we want to make good use of the semantic meanings encoded in word embedding to predict the affective meanings of words. This will help to build valuable lexical resources for affective computing using more comprehensive affective models. The basic idea of our proposed approach is to use regression models to learn the affective meaning in each affective dimension. For a multi-dimensional affective model having m dimensions, the objective is to learn m regression models that are suited for the m affective dimensions. Our method is based on the assumption that word embedding has encoded the general semantic meaning into the dense vector and a certain dimension in word embedding contributes differently to different affective meanings. We consider our approach as a general learning method by using word embedding and regression using a set of seed words, which will be referred to as the *Regression on Word Embedding* approach, labeled as *RoWE*.

3.1 Distributed Word Embedding

The first step in our approach is to build a high-quality feature representation for words using a vector space model (VSM), which represents a word through a low dimensional vector, also called word embedding or word vector [64]. As introduced in Section 2.3, they are mainly count based and prediction based prediction methods for obtaining word embedding. According to a comprehensive study done by [56], both methods can obtain similar information. In other words, they are basically equivalent although fine tuning may be needed. However, the prediction based method has lower computation cost because it does not need to perform matrix factorization over a large co-occurrence matrix. Thus, in this work, we only use the prediction based method to obtain word embedding.

The prediction based method is based on the neural network and one of the most widely used models is Skip-Gram with Negative Sampling (SGNS) [52]. Given a corpus with vocabulary V and the extracted word-context pair set D, let $p(D=1|w,c)$ be the probability that (w,c) comes from D and let $p(D=0|w,c)$ be the probability that (w,c) does not. The basic assumption of SGNS is that the conditional probability of $p(D=1|w,c)$ should be high if c is the context of word w in a window and low otherwise. Let $\vec{w}$ denote the vector representation of w, and $\vec{c}$ denote the vector of c. Then, $p(D=1|w,c)$ is computed as

$$p(D=1|w,c)=\sigma(\vec{w}\cdot\vec{c})=\frac{1}{1+e^{-\vec{w}\cdot\vec{c}}}, \tag{1}$$

where $\vec{w}$ and $\vec{c}$ are the word embedding and context embedding in our model, respectively. Both $\vec{w}$ and $\vec{c}$ are the model parameters to be learned. The basic idea behind is that if word w and context c co-occur, their corresponding vectors should have close correlation, modeled by $\vec{w}\cdot\vec{c}$. The objective of negative sampling is to minimize the conditional probability

$$p(D=1|w,c_N)=\sigma(\vec{w}\cdot\vec{c_N}), \tag{2}$$

where c_N denotes the negative context of w, namely, context that does not co-occur with word w. The method randomly samples negative context c_N of w from V_W. Let P_D be the empirical unigram distribution where

$$P_D(c)=\frac{\#(c)}{|D|}. \tag{3}$$

Combining Formulas (1) and (2), the objective for each word-context pair can be translated into maximizing

$$log\sigma(\vec{w}\cdot\vec{c})+k\cdot\mathbb{E}_{c_N\sim P_D}[log\sigma(-\vec{w}\cdot\vec{c_N})],$$

where k is the number of negative samples. For a given training corpus with a set of words V_W, the final objective function for the whole corpus is

$$\begin{aligned}J=\sum_{w\in V_W}\sum_{c\in V_W}\#(w,c)(&log\sigma(\vec{w}\cdot\vec{c})\\&+k\cdot\mathbb{E}_{c_N\sim P_D}[log\sigma(-\vec{w}\cdot\vec{c_N})]).\end{aligned} \tag{4}$$

The obtained $\vec{w}$ and $\vec{c}$ are the word embedding and context embedding, respectively. The performance of the embedding heavily relies on the hyper-parameters, as shown in [56]. Because finding the optimal word embedding is not our focus, we simply use the recommended settings from [56] for the SGNS model. Note that any kind of learning model for word embedding can be used in our framework including matrix factorization based word embedding [55], ensemble based word embedding [65], etc.

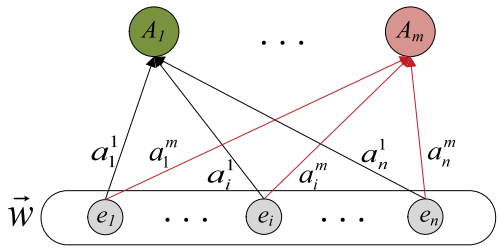

Fig. 3. The proposed regression method for affective representation learning based on word embedding.

3.2 Regression Method for Affective Meanings Prediction

Fig. 3 shows a general learning method of using linear regressions from word embedding to obtain affective meanings of words. In the training phase, each seed word s as a training sample, has known word embedding $\vec{s}$ which is a vector of size n, and its affective meaning is defined in m dimensional space. A word embedding and annotated affective meanings pair consists of one training sample. Given sufficient such pairs, we can learn a regression model for every affective dimension A_j where j is in the range of $[1\dots m]$. Based on the regression model, we can then predict the affective value of a new word based on its word

embedding. Consequently, we can extend an existing affective lexicon automatically.

Given a seed, s, and its word embedding $\vec{w}^s = [e_1^s, e_2^s, \ldots, e_n^s]$, we need to learn the mapping function f_j for the jth affective dimension

$$f_j(\vec{w}^s) = g_j(a_1^j e_1^s + a_2^j e_2^s + \cdots + a_n^j e_n^s), \tag{5}$$

where a_i^j is the weight of feature i, g_j is the mapping function. When f_j is a scalar value, g_j can be the identity function and this model becomes a typical linear regression model. When f_j takes categorical labels, g_j can be a logistic function and this model becomes a typical logistic regression model. f_j can be any kind of affective meanings, such as valence, arousal, dominance in the VAD model, or evaluation, potency, activity in the EPA model, or a simple positive/negative label.

Let V denote the set of seeds. The objective function for regression learning of each affective dimension j is then defined as follows:

$$\min_{\vec{a}} \sum_{s \in V} ||f_j(\vec{w}^s) - y_j^s)||_2^2 + \alpha R(\vec{a}^j), \tag{6}$$

where $R(\vec{a}^j)$ is the regularization part on the weight vector $\vec{a}^j = [a_1^j, a_2^j, \ldots, a_n^j]$ and α is the regularization weight. When $\alpha = 0$, the model degrades to the ordinary least squares linear regression. When $\alpha \neq 0$ and $R(\vec{a}^j) = ||\vec{a}^j||_2^2$, the model degrades to the Ridge regression model. When $\alpha \neq 0$ and $R(\vec{a}^j) = ||\vec{a}^j||_1^1$, the model degrades to the Lasso regression model. Different regression models are evaluated in the experiments.

This model can be trained on existing affective lexicons. Once the model is learned, given the embedding of a new word, its corresponding affective meanings in m dimensions can be predicted using m regression models, respectively. The size of the constructed lexicon depends on the size of available word embedding, which is in principle unlimited because of the large amount of available text corpora.

4 Experiments and Analysis

In this section, we first perform a set of experiments to evaluate our method in inferring affective meanings under different affective models including the sentiment, the valence-arousal-dominance, the evaluation-potency-activity, the evaluation-activation-imagery (EAI). To further prove the generality of our proposed method, we also evaluate our method in inferring other word meanings, including the concreteness-abstractness, the perceptual strength in five senses of hearing, seeing, touching, tasting and smelling. The second set of experiments evaluate the complexity of different methods. The third set of experiments evaluate the effects of the seed words on different methods. The fourth set of experiments evaluate the effects of the embedding dimension size. The fifth set of experiments looks at the performance of different regression models and also examine the different embedding resources in terms their predictability on an existing lexicon. The last set of experiments evaluate the performance of the sentiment lexicons obtained by our method on a downstream sentiment analysis task.

TABLE 2
Summary of Lexicons Used in the Experiments

Lexicon	Num	Overlap Num	std	Affective Meaning	Range
GI	3,626	2,942	N	Sentiment	$\{-1, 0, 1\}$
SemEval2015	1,515	751	N	Sentiment	$[-1, 1]$
VADER	7,502	3,124	Y	Sentiment	$[-4, 4]$
ANEW	1,034	958	Y	VAD	$[1, 9]$
E-ANEW	13,915	11,364	Y	VAD	$[1, 9]$
CVAW	1,647	1,309	Y	VA	$[1, 9]$
EPA	4,505	2,901	Y	EPA	$[-4, 4]$
DAL	8,743	8,003	N	EAI	$[1, 3]$
Perceptual	1,001	826	Y	Five senses	$[0, 5]$
Concreteness	39,954	18,111	Y	Concreteness	$[1, 5]$

4.1 Inferring Affective Meanings

The first set of experiments is set up to explore the effectiveness of our proposed RoWE. The compared methods are listed below.

1) *PMI* [44]: This method learns the intensity value of a word through the pointwise mutual information with the seed words.
2) *qwn-ppv* [47] This method automatically generates a set of positive and negative seed words over WordNet [66]. Then a word graph is constructed from WordNet based on the relations in WordNet. PageRanking algorithm is used to obtain sentiment intensity of unseen words. Here we directly use the provided lexicons for comparison because it is not affected by the corpus as the lexicon is produced from WordNet.
3) *Web GP* [48]: This web-based graph propagation method constructs a weighted graph using cosine similarity of a word by a vector of co-occurrence with its context words. This method only keeps the 25 highest weighted edges for each node to reduce the effect of noise in the web data. The iteration number is set to 5.
4) *Wt-Graph* [20]: This method uses the cosine similarity of word embedding as the edge weights to construct a weighted word graph and then use the PageRank algorithm to obtain the affective meanings (valence and arousal).
5) Densifier [51]: This method learns an orthogonal transformation from the original embedding space to obtain task specific information in ultradense space, such as the one dimensional sentiment polarity space.
6) SentProp [43]: Similar to Wt-Graph [20], this method also employs cosine similarity of word embedding as the edge weights to construct a word graph and use random walk to obtain the affective meaning (sentiment polarity in their work).

All the above methods need to use some seed words to infer the affective meanings of unseen words. To compare fairly, all the methods in the evaluation use the same set of seed words, the same corpus, and the same test settings.

The gold answers used for this set of experiments is a list of affective lexicons which are chosen because they are manually annotated and thus are considered to have high quality. A summary of the lexicons used as gold answers is given in Table 2. The table lists the lexicon names (*Lexicon*),

their sizes (*Num*), the number of words in the lexicon which also appears in the word embeddings (*Overlap Num*), whether standard deviation of annotation is supplied or not (*std*), the affective model (*affective meaning*), and the annotation range (*Range*). *GI* [10] is a sentiment lexicon annotated with *positive, neutral, negative*. During prediction, we use class-mass normalization to give discrete labels as done in [43]. *VADER* [39] and *SemEval2015* [38] are sentiment lexicons annotated with intensity and *VADER* also contains standard deviation of the annotation process. *ANEW* [17] and *E-ANEW* [18] are manually annotated in the three dimensions of valence, arousal and dominance with values from 1 to 9 and *E-ANEW* is an extended version of *ANEW* through crowdsourcing. *CVAW* [19] is the Chinese version of ANEW but annotated only on the two dimensions of valence and arousal. *EPA* [14] is annotated in the two dimensions of evaluation and potency. *DAL*[33] (dictionary of affect in language) is annotated in the three dimensions of evaluation, activation and imagery (*EAI*). *Perceptual* [67], [68] is annotated with perceptual strength of a target word by feeling through five sensations. During annotation, each word is annotated through the question *"To what extent do you experience something being WORD"* (with *"WORD"* being the target word to be annotated). Underneath this question are five separate rating scales for each perceptual modality, labeled "by feeling through touch", "by hearing", "by seeing", "by smelling", and "by tasting". The participants were asked to rate the extent to which they would experience about the five senses, from 0 (not at all) to 5 (greatly) [67], [68]. *Concreteness* [69] is annotated on the degree of concreteness or abstractness of a word through crowdsourcing. Among those lexicons, only CVAW is Chinese and the others are English. We include *Perceptual* and *Concreteness* lexicon, which are actually not affective lexicons, to test the generalization ability of our method on inferring other word meanings.

Experiment Settings. For English lexicons, we train the 300 dimensional word embedding based on Wikipedia August 2016 dump with 3.1 billion tokens.[1] For Chinese, we train the 300 dimensional word embedding based on Baidu Baike corpus with 1.8 billion tokens[2] after performing word segmentation using the HIT LTP tool.[3] Both embeddings are trained using the SGNS model introduced in Section 3.1. The respective overlap sets between the embeddings and the lexicons are randomly split equally to form the training sets and the testing sets. For each experiment, we run *five times* and report the average result with standard deviation in the parenthesis. In addition, we use the relative standard deviation as a metric of the robustness of the methods. To satisfy the requirement of bipolar scale of some baselines (PMI, Web-GP, DENSIFIER, SENTPROP), we transform the affective scales to bipolar scale if needed. For example, ANEW, E-ANEW, and CVAW are mapped from $[1, 9]$ to $[-4, +4]$ linearly, DAL is mapped from $[1, 3]$ to $[-1, +1]$, Perceptual is mapped from $[0, 5]$ to $[-2.5, 2.5]$ and Concreteness is mapped from $[1, 5]$ to $[-2, 2]$. The final predicted values are mapped back to the annotation range. For the regression model in RoWE, Ridge regression is used in the scikit-learn tool with default parameters for the following experiments.

Evaluation Metrics. For the GI lexicon, which is a ternary classification task, we use the area under curve (AUC) and macro F-score as the evaluation metrics using the method in [43] to transform the predicted scalar values to sentiment labels.[4] For the other lexicons, which are continuous value prediction task, we use the following evaluation metrics:

1) Root mean squared error (RMSE) $RMSE = \sqrt{\sum_{i=1}^{n}(A_i - P_i)^2/n}$,
2) Mean absolute error (MAE) $MAE = \frac{1}{n}\sum_{1}^{n}|A_i - P_i|$,
3) Mean absolute percentage error (MAPE) $MAPE = \frac{1}{n}\sum_{1}^{n}\frac{|A_i - P_i|}{A_i} \times 100\%$, and
4) Kendall rank correlation coefficient τ $\tau = \frac{C-D}{C+D}$,

where A_i is the gold standard value, P_i is the predicted value, n is the total number of the test samples, $\bar{A}$ and $\bar{P}$ are the average value of A and P, C is the number of concordant pairs and D is the number of discordant pairs. The lower the values of RMSE, MAPE and MAE, and the higher the value of τ, the better the performance is. Note that the MAPE evaluation metric suffers from the zero-division problem. We do not report the MAPE result if the gold value contains 0. So, for lexicons whose values contain zero (SemEval2015, EPA, DAL, Perceptual), we do not use the MAPE metric because MAPE is sensitive to zero. In addition, for the lexicons with provided standard deviation on annotation, we also use a new evaluation metric defined as follows:

$$ac_{1\sigma} = \frac{1}{n}\sum_{i=1}^{n} g(\sigma_i - |A_i - P_i|), \tag{7}$$

where

$$g(x) = \begin{cases} 1: & x > 0, \\ 0: & otherwise. \end{cases}$$

σ_i is the annotated standard deviation. $\text{ac}_{1\sigma}$ indicates the percentage of correctly predicted samples within 1 standard deviation of the gold answers.

The lexicons can be divided into three types: the sentiment lexicons including GI, SemEval2015 and VADER, the multi-dimensional affective lexicons including ANEW, E-ANEW, CVAW, EPA and DAL, and other word meanings rather than affective meaning including Concreteness and Perceptual. The results are shown in the three sub-tables of Table 3. Table 3 a is for the sentiment lexicons, Table 3 b is for the multi-dimensional affective lexicons, and Table 3 c is for the concreteness and perceptual lexicons. The first dimension (valence or evaluation) of the multi-dimensional lexicons is the same as sentiment. So we include qwn-ppv for comparison on this dimension too. There is no result for qwn-ppv on CVAW because they are in different languages. To make the tables more readable, we only show the standard deviations of the five runs for the sentiment lexicons.

1. https://dumps.wikimedia.org/enwiki/latest/ Accessed May 17, 2017
2. http://www.nlpcn.org/resource/list/2 Accessed May 17, 2017
3. http://www.ltp-cloud.com/ Accessed May 17, 2017

4. Though we can directly predict discrete labels using logistic regression on word embedding, the baseline methods can only produce scalar value. To be consistent with the baselines, we also predict the scalar value using a linear regression model.

TABLE 3
Result on Inferring Affective Meaning

(a) Evaluation results on sentiment lexicons.

Method	GI		SemEval2015			VADER			
	AUC	Macro-F1	RM	MA	τ	RM	MA	τ	$ac_{1\sigma}$
PMI	51.14(0.68)	53.50(4.63)	2.25(1.52)	2.13(1.56)	-0.58(5.29)	2.85(0.15)	2.32(0.16)	0.78(0.56)	28.60(3.55)
Web-GP	51.13(0.94)	48.23(1.19)	0.70(0.05)	0.54(0.03)	-0.47(4.13)	1.79(0.01)	1.62(0.00)	-1.23(2.18)	20.76(0.28)
DENSIFIER	79.45(6.56)	70.50(6.27)	4.78(1.62)	3.77(1.20)	-22.87(7.87)	6.54(1.15)	5.30(0.91)	24.34(33.42)	9.83(1.61)
SENTPROP	72.26(5.66)	64.50(5.37)	0.56(0.05)	0.44(0.05)	16.28(6.71)	1.84(0.01)	1.64(0.01)	27.79(4.99)	22.25(0.51)
qwn-ppv	88.56(0.39)	81.83(0.67)	0.47(0.01)	0.38(0.01)	37.38(3.13)	1.77(0.01)	1.61(0.01)	43.67(0.95)	20.47(0.20)
Wt-Graph	95.05(0.25)	88.23(0.48)	0.47(0.01)	0.37(0.01)	47.20(1.65)	1.67(0.01)	1.52(0.01)	55.98(0.65)	22.61(0.44)
RoWE	**96.16(0.21)**	**89.43(0.65)**	**0.29(0.00)**	**0.23(0.00)**	**55.95(0.85)**	**0.96(0.01)**	**0.74(0.01)**	**62.01(0.44)**	**63.57(0.81)**

(b) Evaluation results on multi-dimensional affective lexicons.

Method	ANEW					E-ANEW					CVAW(Chinese)					EPA			DAL			
	RM	MA	MP	τ	$ac_{1\sigma}$	RM	MA	MP	τ	$ac_{1\sigma}$	RM	MA	MP	τ	$ac_{1\sigma}$	RM	MA	τ	RM	MA	MP	τ
	Valence					Valence					Valence					Evaluation			Evaluation			
PMI	3.6	3.1	80.2	-0.25	31.8	2.0	1.7	40.2	-0.27	55.8	2.2	1.9	48.9	-3.0	15.4	2.7	2.4	-0.77	2.3	2.3	130.7	0.57
Web-GP	2.0	1.8	45.1	-1.4	49.0	1.3	1.0	23.3	0.7	79.1	1.9	1.7	45.1	0.23	12.5	1.4	1.2	0.97	0.48	0.38	23.2	-0.11
qwn-ppv	2.0	1.8	45.8	40.8	49.0	1.3	1.0	23.2	28.8	79.2	-	-	-	-	-	1.4	1.2	31.5	0.44	0.34	21.4	19.1
DENSIFIER	6.9	5.6	137.6	-4.4	19.2	5.2	4.1	88.5	3.1	24.7	8.4	7.0	192.4	25.1	9.4	4.5	3.5	10.3	4.8	3.8	215.2	0.82
SENTPROP	2.0	1.7	47.3	17.8	52.7	1.3	0.99	24.8	17.0	80.4	1.9	1.7	49.3	43.6	12.4	1.3	1.1	20.4	0.75	0.66	41.8	8.9
Wt-Graph	1.9	1.7	43.3	52.9	54.2	1.2	0.96	22.7	44.9	81.0	1.7	1.5	38.5	59.9	12.8	1.3	1.0	42.4	0.43	0.33	20.6	36.5
RoWE	**1.2**	**0.91**	**22.0**	**60.4**	**82.1**	**0.83**	**0.65**	**14.4**	**53.4**	**93.4**	**0.83**	**0.64**	**16.7**	**65.4**	**58.2**	**0.88**	**0.68**	**51.3**	**0.34**	**0.27**	**15.3**	**40.8**
	Arousal					Arousal					Arousal					Activity			Activity			
PMI	2.5	2.3	49.2	-1.4	54.9	2.2	2.0	50.3	0.02	62.5	1.5	1.2	22.4	-1.5	57.2	2.2	2.0	-0.02	2.0	1.9	108.1	0.19
Web-GP	1.1	0.91	18.3	-0.53	93.9	1.2	1.0	28.5	0.1	91.9	1.4	1.1	19.6	0.42	59.7	0.87	0.67	-0.69	0.45	0.36	21.7	0.15
DENSIFIER	5.5	4.4	86.8	-10.7	35.3	5.5	4.4	108.4	6.6	36.8	6.3	5.2	95.5	-9.4	15.0	4.5	3.6	3.7	4.6	3.7	204.7	8.6
SENTPROP	1.1	0.93	20.8	21.3	94.5	1.6	1.4	38.0	10.0	82.8	1.2	0.96	18.6	12.6	64.8	0.74	0.58	0.9	0.71	0.63	39.0	3.3
Wt-Graph	1.0	0.84	17.6	40.2	96.1	0.89	0.71	17.8	32.7	97.7	1.2	0.95	19.0	39.2	66.0	0.75	0.59	34.4	0.39	0.31	18.9	28.7
RoWE	**0.83**	**0.66**	**13.7**	**43.5**	**97.9**	**0.74**	**0.58**	**14.5**	**38.1**	**99.1**	**0.87**	**0.69**	**13.5**	**48.9**	**80.3**	**0.6**	**0.47**	**39.3**	**0.33**	**0.26**	**14.9**	**34.9**
	Dominance					Dominance					/					Potency			Imagery			
PMI	1.5	1.2	24.8	-0.92	81.7	3.3	3.1	60.0	0.82	41.0						3.3	3.2	0.24	2.5	2.3	132.0	1.7
Web-GP	1.1	0.88	19.0	0.72	94.8	0.98	0.79	15.9	-0.26	95.1						1.0	0.85	0.16	0.64	0.53	32.2	0.75
DENSIFIER	5.4	4.3	90.7	3.6	29.7	5.0	4.0	80.0	-6.5	37.7						5.1	4.1	3.3	5.0	4.0	232.8	-2.2
SENTPROP	1.1	0.86	20.2	11.8	94.7	0.96	0.75	16.5	7.9	95.5						0.86	0.68	7.4	0.81	0.68	46.9	20.9
Wt-Graph	0.99	0.79	17.2	46.3	97.0	0.92	0.73	15.3	39.2	96.1						0.86	0.67	32.7	0.6	0.5	31.3	43.2
RoWE	**0.75**	**0.59**	**12.6**	**49.6**	**98.8**	**0.71**	**0.56**	**11.5**	**44.2**	**98.9**						**0.7**	**0.54**	**40.2**	**0.45**	**0.36**	**20.9**	**50.1**

(c) Evaluation results on other word meanings.

Method	Perceptual																Concreteness				
	Hearing				Tasting				Touching				Smelling				Seeing				
	RM	MA	τ	$ac_{1\sigma}$	RM	MA	τ	$ac_{1\sigma}$	RM	MA	τ	$ac_{1\sigma}$	RM	MA	τ	$ac_{1\sigma}$	RM	MA	τ	$ac_{1\sigma}$	RM	MA	MP	τ	$ac_{1\sigma}$
PMI	3.8	3.5	-0.11	9.4	1.7	1.3	0.92	40.8	1.9	1.7	1.2	47.2	3.0	2.8	-2.3	16.9	1.5	1.2	0.3	56.2	2.3	2.1	71.3	-0.46	32.9
Web-GP	1.5	1.3	0.01	54.3	2.2	2.0	-3.8	17.1	1.5	1.3	0.89	55.9	2.0	1.9	0.44	23.4	1.5	1.3	-0.13	47.3	1.0	0.89	31.0	-0.45	66.8
DENSIFIER	6.9	5.6	-1.0	17.7	8.8	6.9	14.4	6.8	6.1	4.9	-3.3	18.6	6.2	4.8	5.9	12.1	5.5	4.4	-7.4	17.8	5.9	4.8	178.3	18.5	15.1
SENTPROP	1.7	1.5	34.5	47.7	2.7	2.6	17.2	12.7	1.6	1.4	12.6	53.3	2.5	2.4	8.9	16.6	1.1	0.96	2.3	60.1	1.0	0.89	35.4	34.9	63.9
Wt-Graph	1.2	1.0	48.0	63.3	1.1	0.82	35.7	45.0	1.3	1.1	39.9	61.0	1.0	0.83	29.3	51.8	0.87	0.69	37.4	78.8	0.97	0.84	30.8	56.2	67.9
RoWE	**0.91**	**0.73**	**50.8**	**76.6**	**0.73**	**0.52**	**40.0**	**61.4**	**0.96**	**0.8**	**49.0**	**75.9**	**0.77**	**0.58**	**37.9**	**66.6**	**0.71**	**0.56**	**41.5**	**85.9**	**0.56**	**0.44**	**16.0**	**64.4**	**90.6**

RM for RMSE; MA for MAE; MP for MAPE.

Based on the results from these tables, we make five major observations. (1) RoWE outperforms the other methods with large margins on all the affective dimensions of all the lexicons under all the evaluation metrics. For example, on the GI lexicon, RoWE has a relative improvement of 1.2 percent on AUC and 1.3 percent on Macro-F1 over the state-of-the-art Wt-Graph method. On the ANEW lexicon, RoWE outperforms the state-of-the-art Wt-Graph method with relative improvement of 36.8, 47.1, 49.2, 14.2, and 51.5 percent for RMSE, MAE, MAPE, and the Kendall correlation coefficient τ metrics, respectively. On the touching dimension of Perceptual lexicon, RoWE achieves a relative improvement over Wt-Graph of 26.2, 27.3, 22.8, 24.4 percent under RMSE, MAE, τ and $ac_{1\sigma}$ respectively. (2) Among different evaluation metrics, rankings on RMSE, MAE and MAPE are similar. But, Kendall correlation coefficient are different. For example, the ranking for RMSE from best to worst is RoWE, Wt-Graph, PMI, SENTPROP, Web-GP and qwn-ppv, DENSIFIER. However, the ranking for τ is RoWE, Wt-Graph, qwn-ppv, SENTPROP, PMI, Web-GP, DENSIFIER. The ranking for $ac_{1\sigma}$ is RoWE, Wt-Graph, SENTPROP, Web-GP and qwn-ppv, PMI, DENSIFIER. This means that different methods may have their

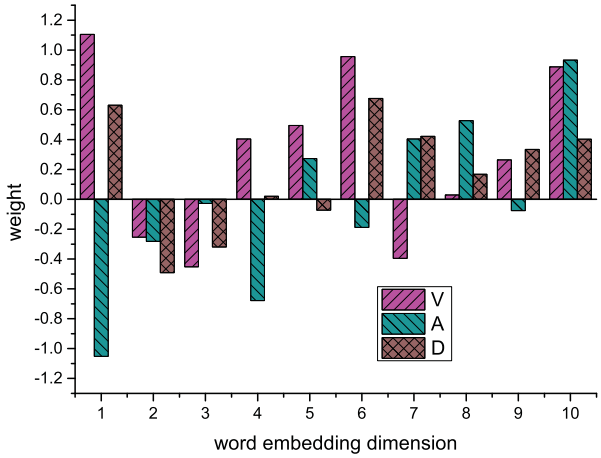

Fig. 4. The learned weights of different affecetive meanings for the ANEW lexicon.

merits under different performance measures. (3) To consider the different dimensions for the VAD lexicons, the performance on valence for τ is better than on arousal and dominance. However, it is opposite for $ac_{1\sigma}$, RMSE, MAE and MAPE. This may be because τ focuses on the ranking rather than value difference between the gold value and the predicted value, whereas the other evaluation metrics focus on the value difference between the gold value and the predicted value. (4) For the E-ANEW lexicon, which is annotated through crowdsourcing, the mean absolute errors (MAE) of our method are 0.65, 0.58, 0.56 on valence, arousal, dominance, respectively. This means that the predicted values are quite close to the manually annotated values. On the $ac_{1\sigma}$ metric, our method's performance achieves 93.4, 99.1, 99.0 percent on valence, arousal, dominance, respectively. This means that almost all the predicted values are in one standard deviation of the manually annotated mean value. (5) The standard deviations shown in parentheses of the sentiment lexicons indicate that RoWE has smaller relative standard deviations. In other words, RoWE is more robust and is less seed word sensitive.

In conclusion, our proposed RoWE method achieves the best result on all the lexicons under all the evaluation metrics, which validates our assumption that word embeddings do encode semantic information and the regression model can effectively decode the affective meanings from the embeddings by assigning different weights to different dimensions in the embedding. Fig. 4 shows a visualized weight values of $\vec{a}$ on the first ten dimensions in the vector space of word embedding to the three affective dimensions on ANEW lexicon for the VAD model. Note that the weights for the three affective dimensions can be quite different. For example, for the first vector in embedding, its corresponding affectives weights are 1.11, -1.05, and 0.63, respectively.

Table 4 lists some example words in the ANEW lexicon that are close in embedding space but not close in the valence dimension. In the table, the *word* column is the target word, the *G val* column is the gold valence value, *P val* is the predicted valence value, and the last column is the top 5 nearest words in embedding space based on cosine similarity. The value in the parenthesis is the predicted valence value. The words in bold are examples that are close in the embedding space but not close in the valence dimension.

TABLE 4
Example Words Close in Embedding Space, But Not Close in Predicted Affective Space

Word	G val	P val	Top 5 nearest words in embedding space
good	7.47	6.45	decent(5.94), **bad(3.34)**, excellent(7.35), **poor(3.32)**, commendable(7.19)
heaven	7.3	6.80	heavens(6.33), heavenly(6.80), **hell(4.74)**, god(6.54), afterlife(5.63)
clouds	6.18	5.66	cloud(5.00), mist(5.00), droplets(4.85), **dust(4.27)**, **overcast(4.54)**
cold	4.02	4.16	**warm(7.09)**, winters(5.27), colder(4.94), cool(6.34), freezing(4.24)
displeased	2.79	3.64	angered(3.34), unhappy(3.43), incensed(3.37), **pleased(6.40)**, apprehensive(3.79)

For example, the nearest word of *cold* is *warm* while their predicted valence value are 4.16 and 7.09 respectively. This validates that our method can distinguish the affective meanings through assigning different weights to the features in the embedding space.

4.2 Method Complexity

The complexity of different methods are shown in Table 5. In this table, N is the data sample size, d is the embedding dimension and k is the number of nearest neighbors used in Web-GP and SentProp. d and k are set as constants during experiment. The second column in the table indicates that the asymptotic complexities of PMI, Web-GP, Wt-Graph and SentProp grow quadratically with the data size, whereas the complexities of Densifier and our RoWE grows linearly with the data size. The third column in the table shows the complexity with constant coefficients d and k. Even though d and k do not have a role to play in Big O analysis, as shown in the second column, they do affect the efficiency of the implementations especially when data samples have limited size.

To further examine their run time efficiency, we also run an experiment to visually observe the difference in computing time by varying the data size from 1,000 to 11,000 using the E-ANEW lexicon and set the seed word number to 300. The remaining collection is used as test data. The hardware platform is a desktop computer with processor of Intel (R) Xeon (R) CPU E5-1620 and 64G RAM and during running each method, we close all the other programs. The result is shown in Fig. 5. Web-GP is not listed because its running time is too high ranging from about 23,900 to 38,000 (in micro seconds). The figure shows that RoWE requires the least running time. When the data size increase from 1,000 to 11,000, the the running time of RoWE changes from 11 to

TABLE 5
Complexity of Different Methods

Method	Asymptotic Complexity	Complexity with coefficient
PMI	$O(N^2)$	$O(N^2)$
Web-GP	$O(N^2)$	$O(N^2kd)$
Wt-Graph	$O(N^2)$	$O(N^2d)$
Densifier	$O(N)$	$O(Nd^3)$
SentProp	$O(N^2)$	$O(N^2kd)$
RoWE	$O(N)$	$O(Nd^2)$

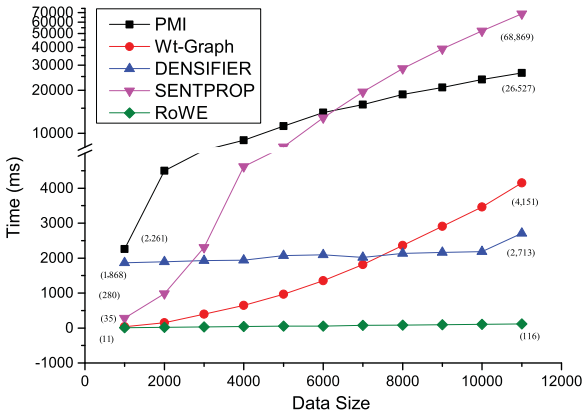

Fig. 5. The running time of different methods under different data size. We break the *y* axis at 5,000 to 6,000 to make the figure more readable. The numbers in parenthesis are the running time.

116 which basically translates a linear increase of 10.5 times. Although running time may be affected by actual implementations, this experiment can still reveal the computation advantage of RoWE over the other methods. In conclusion, RoWE has complexity advantage over the other methods.

4.3 The Effect of Seed Words

In this experiment, we explore the effects of seed word size using the ANEW lexicon. We change the size of the seed words from 10 to 800 with 30 as the step size and the remaining lexicon as the test data. Without loss of generality, we only measure the valence dimension in terms of $ac_{1\sigma}$. Result shown in Fig. 6 indicates that Web-GP, SentProp, and Wt-Graph methods achieve almost similar result and they are stable without much room to improve when more seed words are added. PMI and DNSIFIER, however, is not quite stable. RoWE has much better performance. It can also improve its performance when more seed words are used. Note that even with a small set of seed words (such as 100, which can be obtained easily through manual annotation or crowdsourcing), RoWE still achieves much better result.

4.4 Effects of Dimension Size of Word Embedding

In this experiment, we explore the effect of embedding dimension size. We train word embeddings with different dimension sizes on the Wikipedia corpus using the SGNS model and report the RMSE performance on the VADER lexicon and the VAD dimensions of the E-ANEW lexicon. The

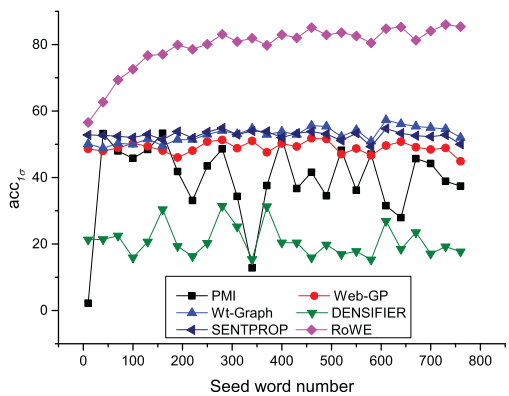

Fig. 6. The effect of seed word size on the ANEW lexicon.

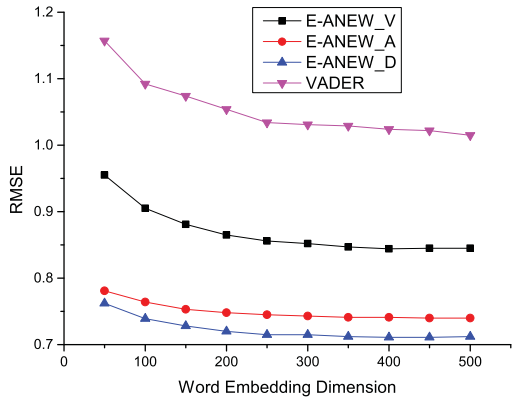

Fig. 7. The effects of embedding dimension.

result is shown in Fig. 7. Note that as the dimension increases from 50 to 300, the performance improves steadily. However, between 300 to 500, the curve is quite flat. Generally speaking, larger dimensions do bring better performance, but it would require more resources and computation power. To balance the performance and computation cost, we suggest to set the dimension between 300 to 400.

4.5 Effects of Regression Models and Embedding Methods for RoWE

In previous experiments, we use the Ridge regression model and the word embedding trained using the SGNS model. In principle, any regression model and word embedding method can be used in our proposed method. In practice, however, different regression methods and the actual embedding method may affect the overall performance. In this section, we explore the effects of the regression models and word embedding methods.

In principle different regression models can be used as explained in Section 3, such as linear regression, Ridge regression, BayesianRidge regression, ElasticNet regression, Lasso regression, as well as Support Vector Regression with linear kernel (SVM-Linear), Support Vector Regression with non-linear Gaussian kernel (SVM-RBF). We examine their performance in terms of $ac_{1\sigma}$, using the one-dimensional VADER lexicon. The size of the seed words changes from 10 to 600 with 30 as the step size and the remaining lexicon as the test data. All the models are based on scikit-learn[5] with default parameters. The result is shown in Fig. 8. Note that the SVR-Linear, the Ridge and the Bayesian Ridge achieve similar and much better result than the other regression models. This is because that Ridge regressions and SVR-Linear use norm 2 regularization on the weights to avoid overfitting. The linear regression model shows the U shape because of overfitting without regularization on the weight coefficients. SVR-Linear performs much better than SVR-RBF, which indicates that linear models are more suitable than non-linear model for inferring affective meanings from word embedding. Similar results are obtained under other evaluation metrics and other affective lexicons. Thus, we suggest to use SVM-Linear, Ridge or Bayesian Ridge regression models in our framework.

We conduct evaluation on different embedding resources. In addition to the embedding trained from Wikipedia,

5. scikit-learn.org/ Accessed May 17, 2017

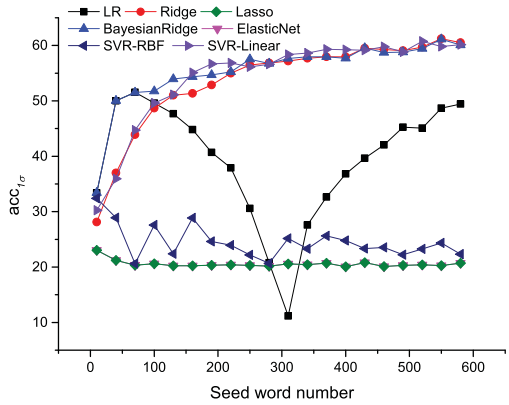

Fig. 8. The performance of different regression models on the VADER lexicon.

denoted as *wikiEmb* with size 204,981, as explained in Section 4.2, we also use the following public available embeddings that are obtained from different learning methods.

1) *Google embedding* (GoogleEmb) [52]: It is trained using the SGNS model as introduced in Section 3.1 from a news corpus of 10 billion tokens.[6] The embedding vocabulary size is 3,000,000.
2) *Glove 840B* (Glove) [55]: It is based on weighted matrix factorization on the co-occurrence matrix built from a corpus consisting of 840 billion tokens.[7] The embedding vocabulary size is 2,196,017.
3) *Meta-Embedding* (MetaEmb) [65]: This method ensembles different embedding sources to obtain the final meta-embedding.[8] The size is 2,746,092
4) *ConceptNet Vector Ensemble (CNVE)* [70]: This method combines word2vec, Glove with structured knowledge from ConceptNet [71].[9] The size is 426,572.
5) *MVLSA* (MVEmb) [62]: This method learns word embedding from multiple sources including text corpus, dependency relation, morphology, monolingual corpus, knowledge base from FramNet based on generalized canonical correlation analysis.[10] The size is 361,082.
6) *Paragram Embedding (ParaEmb)* [72]: This method learns word embeddings based on the paraphrase constraint from PPDB.[11] The size is 1,703,756.

We test the embeddings on the common set of 1,079 words in all the selected embedding resources and the VADER lexicon. Among the 1,079 words, we randomly select 50 percent as seed words and the other 50 percent as test words. We run each experiment 5 times and report the average performance with standard deviation in the parenthesis as shown in Table 7. Note that the knowledge based CVNE achieves the best result under all the evaluation metrics, which indicates that distilling knowledge base into embedding can improve the quality of word embedding. GoogleEmb performs slightly better than wikiEmb because GoogleEmb uses a much larger training corpus. Since evaluating embedding quality is not our focus, for the detailed discussion on the quality of embedding methods, we suggest the paper [56]. Other than MVEmb, which seems to be low in performance, all the other embeddings have comparable performance. Even though CVNE has the best performance in this experiment, it only indicates the usefulness of adding knowledge base information to a non-supervised training method. It does not by any means guarantee that CVNE is the best performer on a downstream task because the lexicon size is limited by the coverage of the knowledge base.

Table 6 shows the example words with the top 5 largest and top 5 smallest predicted values in each affective dimension under different affective models using the Ridge regression based on corresponding seed lexicons and CVNE embedding. Note that all the learned top words are quite reasonable. As sentiment indicators, ANEW-v, EPA-e has the same word *giving gift*. Several words do get listed in different lexicons such as *giving gift, make happy*. Note that our method is not limited to predict the affective meaning of words only. Phrase prediction is not a problem in general as long as phrase embeddings are given. Interestingly, on the Concreteness, the last word *istically* actually is the adverb suffix, which is quite abstract.

4.6 Downstream Task for Sentiment Classification

In this section, we evaluate the effectiveness RoWE through the performance of a downstream sentiment analysis task. In this experiment, we examine the effectiveness of the lexicons obtained from RoWE compared to baseline lexicons obtained from other methods including both manual ones and automatically obtained ones. The sentiment corpora used in the experiment are listed in Table 8. The baseline lexicons are all publicly available and are listed in Table 9. The list of lexicons is sorted according to their size. Note that the ANEW, VADER and E-ANEW are obtained manually or through crowdsourcing and the others are obtained automatically.

The setup of the experiment is to first use RoWE to extend the VADER sentiment lexicon using different embeddings introduced in Section 4.5. RoWE is trained using the intersection of the VADER lexicon and the respective embeddings. The size for each of the extended lexicon is different depending on the vocabulary of the embeddings. For a fair comparison, we use the same downstream sentiment classification method for all the different lexicons. We use the VADER method for sentiment classification [39] because it is a lexicon-based method using heuristic rules. We did not use any machine learning method to avoid the effects of other factors other than the evaluated lexicons. The VADER method can better reflect the quality of the evaluated sentiment lexicons. In the sentiment analysis task, we use F-score as the evaluation metric.

Table 10 shows the evaluation result and the best results are in bold. Note that all the lexicons obtained by using RoWE are listed in the second part of the table and the size of each obtained lexicon is included in parenthesis. In general, the embedding based lexicons perform better than the baseline lexicons. The ParaEmb lexicon, in particular, achieves the best result on all the sentiment corpora. In the baseline

6. https://code.google.com/archive/p/word2vec/ Accessed May 17, 2017
7. http://nlp.stanford.edu/projects/glove/ Accessed May 17, 2017
8. http://cistern.cis.lmu.de/meta-emb/ Accessed May 17, 2017
9. https://github.com/commonsense/conceptnet-numberbatch Accessed May 17, 2017
10. http://cs.jhu.edu/~prastog3/mvlsa/Accessed May 17, 2017
11. http://ttic.uchicago.edu/~wieting/Accessed May 17, 2017

TABLE 6
Example Words with Top 5 Largest and Smallest Predicted Affective Values Based on CVNE Embedding

VADER	ANEW-v	ANEW-a	ANEW-d	EPA-e	EPA-p	EPA-a	DAL-e	DAL-a	DAL-i	Concreteness
Examples words of top 5 largest predicted affective values										
giving gift	giving gift	insanity	paradise	giving gift	god	raver	giving gift	dangerous activity	neighbor's house	non powered device
making happy	making happy	gun	win	heaven	ceo	riot	making happy	climbing mountain	non powered device	opaque thing
excellentness	make happy	sex	positive attitude	make happy	christ	gunfight	make happy	playing snooker	own home	power shovel excavator
life of party	reading books	rampage	incredible	making happy	herculean strength	fighter	showing love	winning game	opaque thing	non agentive artifact
winning baseball game	positive attitude	tornado	self	positive attitude	pope	nightclub	enjoying day	playing cricket	single user device	single user device
Examples words of top 5 smallest predicted affective values										
hell with	stabbing to death	soothing	uncontroll- able	hell	coward	glum	mommick	scar	that degree	more equal
unpleasant person	life threatening condition	librarian	earthquake	murder	weakling	cemetery	unpleasant person	shadows	risibility	confessedly
hagridden	poor devil	dull	lobotomy	rape	high and dry	funeral	plague	elementary	in such way	hypostatize
abusive language	crybully	calm	alzheimers	unpleasant person	slave	mummy	plaguer	supplement	inhere	neuter substantive
hagride	abusive language	grain	dementia	rapist	powerless	graveyard	nidder	oxgang	in this	istically

lexicons, SentiWords performs the best. We want to point out that in both the baseline lexicons and lexicons obtained from RoWE, lexicon size is not the determiner for the best performance. Among the baseline lexicons, the best performer, SentiWords has only about 147K sentiment words whereas NNLexicon and Tang have about 184 K and 347 K respectively. The best performer ParaEmb is also not the largest in lexicon size. In fact, CVNE which is only 0.4 M in size has very good performance. Note that MetaEmb performs much worse than other embedding based lexicons. Further analysis indicates that although the size of MetaEmb is large, the overlap size of MetaEmb with the sentiment corpus vocabulary is quite low, For example, there are only 512 overlapping seeds out of 6,298 (10 percent) in the mpqa corpus compared to 6,193 of ParaEmb. Also, most of the words in MetaEmb are informal strings, such as *rates.download*, *now!download*. The general conclusion is that (1) the larger overlapping is generally good, but again it is not the determining factor; and (2) the high quality word embedding also helps even if its size is not large (as shown by CVNE).

TABLE 7
Evaluation of Different Embeddings on VADER Lexicon Using RoWE

Method	RMSE	MAE	τ	$ac_{1\sigma}$
wikiEmb	1.2(.02)	.96(.01)	49.9(1.1)	53.6(1.0)
GoogleEmb	1.1(.01)	.86(.01)	55.4(1.0)	57.6(1.5)
Glove	1.0(.02)	.80(.02)	59.4(1.2)	61.7(1.5)
CVNE	**.88(.01)**	**.69(.01)**	**66.0(.95)**	**67.3(1.2)**
MetaEmb	1.1(.03)	.86(.02)	56.4(1.3)	57.8(1.4)
MVEmb	1.3(.02)	1.0(.02)	42.4(1.0)	50.7(.31)
ParaEmb	1.0(.02)	.80(.02)	59.6(1.4)	60.8(1.4)

TABLE 8
Statistics of Sentiment Corpus

Corpus	num	pos num	vocab	avg words	Description
sem [73]	3,583	2,570	18,965	19.8	SenEval 2013
mR [39]	10,605	5,242	29,864	18.9	movie review
aR [39]	3,708	2,128	8306	16.5	Amazon review
nyt [39]	5,190	2,204	20,929	17.5	News
cr [74]	3,771	2,405	5,712	20.1	customer review
mpqa [75]	10,603	3,311	6,298	3.1	news
mr [76]	10,662	5,331	21,425	21.0	movie review
SST [77]	1,821	909	7,576	19.2	movie review

TABLE 9
Statistics of Baseline Sentiment Lexicons

Lexicon	size	Description
ANEW [17]	1034	manual annotation
VADER [39]	7,502	crowdsourcing annotation
E-ANEW [18]	13,915	crowdsourcing annotation
SenticNet4 [42]	50,000	propagation on ConceptNet
HashtagSenti [78]	54,129	statistics based on hashtag
senti140 [78]	62,468	statistics based on emoticon
qwn-ppv [47]	81,248	propagation on WordNet
SentiWordNet3 [79]	89,631	automatic based on WordNet
SentiWords [80]	147,305	ensemble on SentiWordNet
NNlexicon [81]	184,579	neural network prediciton
Tang [49]	347,446	representation learning

TABLE 10
Result on Downstream Sentiment Analysis Task

Lexicon(size in M)	sem	mR	aR	nyt	cr	mpqa	mr	SST
ANEW	0.71	0.56	0.55	0.49	0.62	0.27	0.54	0.57
VADER	0.83	0.66	0.71	0.57	0.78	0.63	0.66	0.70
E-ANEW	**0.85**	0.68	**0.74**	0.63	**0.79**	0.58	0.68	0.70
SenticNet4	0.79	0.66	0.69	0.59	0.74	0.57	0.66	0.68
HashtagSenti	0.81	0.62	0.66	0.53	0.71	0.41	0.62	0.66
senti140	0.82	0.68	0.65	0.60	0.68	0.55	0.68	0.70
qwn-ppv	0.76	0.63	0.69	0.57	0.74	0.45	0.63	0.66
SentiWordNet3	0.65	0.56	0.56	0.49	0.62	0.43	0.56	0.60
SentiWords	**0.85**	0.68	**0.74**	0.63	**0.79**	0.60	0.68	0.71
NNlexicon	0.77	0.64	0.68	0.53	0.73	0.55	0.64	0.67
Tang	0.83	0.63	0.63	0.53	0.66	0.54	0.63	0.68
wikiEmb(0.2 M)	0.84	0.68	**0.74**	0.62	0.78	0.66	0.68	0.69
GoogleEmb(3 M)	**0.85**	0.68	**0.74**	0.63	0.78	0.68	**0.69**	0.70
Glove(2 M)	**0.85**	**0.69**	**0.74**	**0.65**	**0.79**	0.69	**0.69**	0.71
CVNE(0.4 M)	**0.85**	**0.69**	**0.74**	0.63	0.78	0.68	**0.69**	0.70
MetaEmb(2.7 M)	0.73	0.47	0.49	0.43	0.48	0.04	0.49	0.47
MVEmb(0.36 M)	**0.85**	0.68	**0.74**	0.62	0.78	0.68	0.68	0.69
ParaEmb(1.7 M)	**0.85**	**0.69**	**0.74**	**0.65**	**0.79**	**0.70**	**0.69**	**0.72**

5 Conclusion

In this paper, we present a regression based method to automatically infer the affective meanings of words based on word embedding. We argue that word embedding not only carries general semantic meaning, but also meanings in some specific space, such as sentiments and affects which can be obtained through training. This framework first learns the word embedding through unsupervised way and then treat word embedding as the feature representation to train a Ridge regression model based on a small set of seed words. Our framework can infer different kinds of affective meanings in multi-dimensional models. A whole set of evaluations shows that our method achieves the state-of-the-art performance and outperforms all the baselines in both performance and computation cost. Existing lexicons can be easily extended through our method and experiment on downstream sentiment analysis task shows that the extended lexicon performs better than existing public sentiment lexicons on several sentiment corpora, which again indicates the effectiveness of the proposed method. We make the extended lexicons of different affective models publicly available.[12] Future work may include investigating on how to obtain higher quality word embeddings (especially incorporating text corpus and knowledge base) and how to apply the obtained multi-dimensional lexicons on affective computing for longer text.

Acknowledgments

This work is supported by HK Polytechnic University (PolyU RTVU and GRF PolyU 15211/14E). The work was done when Lin Gui was a research assistant with in the Hong Kong Polytechnic University.

References

[1] R. W. Picard, "Affective computing," MIT Media Lab, Cambridge, MA, USA, Tech. Rep. 321, 1995.

[2] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retrieval*, vol. 2, no. 1/2, pp. 1–135, 2008.

[3] E. Cambria, A. Hussain, and C. Eckl, "Taking refuge in your personal sentic corner," in *Proc. 5th Int. Joint Conf. Natural Language Process.*, 2011, pp. 35–43.

[4] B. Desmet and V. Hoste, "Emotion detection in suicide notes," *Expert Syst. Appl.*, vol. 40, no. 16, pp. 6351–6358, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417413003485

[5] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S187775031100007X

[6] S. Poria, E. Cambria, and A. Gelbukh, "Aspect extraction for opinion mining with a deep convolutional neural network," *Knowl.-Based Syst.*, vol. 108, pp. 42–49, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0950705116301721

[7] S. Poria, E. Cambria, D. Hazarika, and P. Vij, "A deeper look into sarcastic Tweets using deep convolutional neural networks," in *Proc. 26th Int. Conf. Comput. Linguistics*, 2016, pp. 1601–1612. [Online]. Available: https://arxiv.org/abs/1610.08815

[8] N. Majumder, S. Poria, A. Gelbukh, and E. Cambria, "Deep learning-based document modeling for personality detection from text," *IEEE Intell. Syst.*, vol. 32, no. 2, pp. 74–79, Mar./Apr. 2017. [Online]. Available: http://ieeexplore.ieee.org/abstract/document/7887639/

[9] J. Hoey, T. Schrder, and A. Alhothali, "Affect control processes: Intelligent affective interaction using a partially observable Markov decision process," *Artif. Intell.*, vol. 230, pp. 134–172, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S000437021500140X

[10] P. J. Stone, D. C. Dunphy, M. S. Smith, and D. M. Ogilvie, "The general inquirer: A computer approach to content analysis," *J. Regional Sci.*, vol. 8, no. 1, pp. 113–116, 1968. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-9787.1968.tb01290.x/f ull

[11] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word-emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, 2013.

[12] J. Staiano and M. Guerini, "Depeche mood: A lexicon for emotion analysis from crowd annotated news," in *Proc. Ann. Meeting Assoc. Comput. Linguistics*, 2014, vol. 2, pp. 427–433. [Online]. Available: http://aclanthology.info/papers/depeche-mood-a-lexicon-for-emotion-analysis-from-crowd-annotated-news

[13] J. A. Russell, "A circumplex model of affect," *J. Personality Social Psychology*, vol. 39, no. 6, 1980, Art. no. 1161.

[14] D. R. Heise, "Semantic differential profiles for 1,000 most frequent English words," *Psychological Monographs: General Appl.*, vol. 79, no. 8, 1965, Art. no. 1. [Online]. Available: http://psycnet.apa.org/journals/mon/79/8/1/

[15] R. A. Calvo and S. Mac Kim , "Emotions in text: Dimensional and categorical models," *Comput. Intell.*, vol. 29, no. 3, pp. 527–543, 2013. [Online]. Available: http://onlinelibrary.wiley.com/doi/10.1111/j.1467-8640.2012.00456.x/full

[16] D. R. Heise, "Affect control theory: Concepts and model," *J. Math. Sociology*, vol. 13, no. 1/2, pp. 1–33, 1987. [Online]. Available: http://dx.doi.org/10.1080/0022250X.1987.9990025

[17] M. M. Bradley and P. J. Lang, "Affective norms for English words (ANEW): Instruction manual and affective ratings," Center Res. Psychophysiology, University of Florida, Gainesville, FL, USA, Tech. Rep. C-1, 1999. [Online]. Available: http://www.uvm.edu/pdodds/teaching/courses/2009-08UVM-300/docs/others/everything/bradley1999a.pdf

[18] A. B. Warriner, V. Kuperman, and M. Brysbaert, "Norms of valence, arousal, and dominance for 13,915 English lemmas," *Behavior Res. Methods*, vol. 45, no. 4, pp. 1191–1207, 2013. [Online]. Available: http://link.springer.com/article/10.3758/s13428-012-0314-x

[19] L.-C. Yu, et al., "Building Chinese affective resources in valence-arousal dimensions," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technol.*, 2016, pp. 540–545. [Online]. Available: http://m-mitchell.com/NAACL-2016/NAACL-HLT2016/pdf/N16-1066.pdf

[20] L.-C. Yu, J. Wang, K. R. Lai, and X.-J. Zhang, "Predicting valence-arousal ratings of words using a weighted graph method," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2015, vol. 2, pp. 788–793. [Online]. Available: http://aclweb.org/anthology/P/P15/P15-2129.pdf

12. https://www.dropbox.com/sh/t6yy9yyrkj354rf/AAA49K2XimtcccawUszXJ7zLa?dl=0 Accessed May 17, 2017.

[21] G. N. Leech, *Semantics: The Study of Meaning*, 2nd ed. London, U.K.: Penguin Books, 1981.

[22] W.-L. Wei, C.-H. Wu, and J.-C. Lin, "A regression approach to affective rating of Chinese words from ANEW," in *Affective Computing and Intelligent Interaction*. Berlin, Germany: Springer, 2011, pp. 121–131.

[23] N. Malandrakis, A. Potamianos, E. Iosif, and S. S. Narayanan, "Kernel models for affective lexicon creation," in *Proc. 12th Annu. Conf. Int. Speech Commun. Assoc.*, 2011, pp. 2977–2980. [Online]. Available: http://sail.usc.edu/malandra/files/papers/interspeech2011.pdf

[24] P. Ekman, "Facial expression and emotion," *Amer. Psychologist*, vol. 48, no. 4, 1993, Art. no. 384.

[25] W. G. Parrott, *Emotions in Social Psychology: Essential Readings*. Hove, U.K.: Psychology Press, 2001.

[26] N. H. Frijda, *The Emotions*. Cambridge, U.K.: Cambridge Univ. Press, 1986.

[27] S. S. Tomkins, "Affect theory," *Approaches Emotion*, vol. 163, 1984, Art. no. 195.

[28] A. Ortony, *The Cognitive Structure of Emotions*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[29] L. Xu, H. Lin, P. Yu, H. Ren, and J. Chen, "Constructing the affective lexicon ontology," *J. China Soc. Sci. Tech. Inf.*, vol. 2, 2008, Art. no. 6.

[30] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum, *The Measurement of Meaning*. Champaign, IL, USA: Univ. Illinois Press, 1957.

[31] E. Cambria, A. Livingstone, and A. Hussain, "The hourglass of emotions," in *Cognitive Behavioural Systems*. Berlin, Germany: Springer, 2012, pp. 144–157.

[32] A. Mehrabian, "Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament," *Current Psychology*, vol. 14, no. 4, pp. 261–292, 1996.

[33] C. Whissell, "The dictionary of affect in language," *Emotion: Theory Res. Experience*, vol. 4, no. 113–131, 1989, Art. no. 94.

[34] J. R. Fontaine, K. R. Scherer, E. B. Roesch, and P. C. Ellsworth, "The world of emotions is not two-dimensional," *Psychological Sci.*, vol. 18, no. 12, pp. 1050–7, Dec. 2007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/18031411

[35] H. Lvheim, "A new three-dimensional model for emotions and monoamine neurotransmitters," *Med. Hypotheses*, vol. 78, no. 2, pp. 341–348, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306987711005883

[36] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proc. Human Language Technol. Conf. Conf. Empirical Methods Natural Language Process.*, 2005, pp. 347–354. [Online]. Available: http://aclanthology.info/papers/recognizing-contextual-polarity-in-phrase-level-sentiment-analysis

[37] S. M. Mohammad, S. Kiritchenko, and X. Zhu, "NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets," in *Proc. 7th Int. Workshop Semantic Evaluation.*, 2013, pp. 321–327.

[38] S. Rosenthal, P. Nakov, S. Kiritchenko, S. M. Mohammad, A. Ritter, and V. Stoyanov, "SemEval-2015 task 10: Sentiment analysis in Twitter," in *Proc. 9th Int. Workshop Semantic Evaluation.*, 2015, pp. 451–463. [Online]. Available: http://www.aclweb.org/website/old_anthology/S/S15/S15-2.pdf#page=493

[39] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," in *Proc. Eighth Int. Conf. Weblogs Social Media*, 2014. [Online]. Available: http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/viewPaper/8109

[40] D. S. Schmidtke, T. Schrder, A. M. Jacobs, and M. Conrad, "ANGST: Affective norms for German sentiment terms, derived from the affective norms for English words," *Behavior Res. Methods*, vol. 46, no. 4, pp. 1108–1118, Jan. 2014. [Online]. Available: http://link.springer.com/article/10.3758/s13428-013-0426-y

[41] B. Liu, "Sentiment analysis and opinion mining," *Synthesis Lectures Human Language Technol.*, vol. 5, no. 1, pp. 1–167, 2012.

[42] E. Cambria, S. Poria, R. Bajpai, and B. Schuller, "SenticNet 4: A semantic resource for sentiment analysis based on conceptual primitives," in *Proc. 26th Int. Conf. Comput. Linguistics*, 2016, pp. 2666–2677. [Online]. Available: http://www.sentic.net/senticnet-4.pdf

[43] W. L. Hamilton, K. Clark, J. Leskovec, and D. Jurafsky, "Inducing domain-specific sentiment lexicons from unlabeled corpora," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2016, pp. 595–605. [Online]. Available: http://aclanthology.info/papers/inducing-domain-specific-sentiment-lexicons-from-unlabeled-corpora

[44] P. D. Turney and M. L. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *ACM Trans. Inf. Syst.*, vol. 21, no. 4, pp. 315–346, 2003. [Online]. Available: http://nparc.cisti-icist.nrc-cnrc.gc.ca/npsi/ctrl?action=rtdoc&an=5210015

[45] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from Tweets," *Comput. Intell.*, vol. 31, no. 2, pp. 301–326, 2015.

[46] A. Alhothali and J. Hoey, "Good news or bad news: Using affect control theory to analyze readers' reaction towards news articles," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technol.*, 2015, pp. 1548–1558. [Online]. Available: http://aclanthology.info/papers/good-news-or-bad-news-using-affect-control-theory-to-analyze-readers-reaction-towards-news-articles

[47] I. San Vicente , R. Agerri, G. Rigau, and D.-S. Sebastin, "Simple, robust and (almost) unsupervised generation of polarity lexicons for multiple languages," in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2014, pp. 88–97. [Online]. Available: http://www.aclweb.org/anthology/E14-1#page=114

[48] L. Velikovich, S. Blair-Goldensohn , K. Hannan, and R. McDonald, "The viability of Web-derived polarity lexicons," in *Proc. Human Language Technol.: Annu. Conf. North Amer. Chapter Assoc. Comput. Linguistics*, 2010, pp. 777–785. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858118

[49] D. Tang, F. Wei, B. Qin, M. Zhou, and T. Liu, "Building large-scale Twitter-specific sentiment lexicon: A representation learning approach," in *Proc. 25th Int. Conf. Comput. Linguistics: Tech. Papers*, 2014, pp. 172–182. [Online]. Available: http://aclweb.org/anthology/C/C14/C14-1018.pdf

[50] S. Amir, R. Astudillo, W. Ling, B. Martins, M. J. Silva, and I. Trancoso, "INESC-ID: A regression model for large scale Twitter sentiment lexicon induction," in *Proc. 9th Int. Workshop Semantic Evaluation.*, 2015, pp. 613–618. [Online]. Available: http://aclweb.org/anthology/S15-2102

[51] S. Rothe, S. Ebert, and H. Schtze, "Ultradense word embeddings by orthogonal transformation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technol.*, 2016, pp. 767–777. [Online]. Available: http://aclanthology.info/papers/ultradense-word-embeddings-by-orthogonal-transformation

[52] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. 27th Annu. Conf. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[53] M. Baroni, G. Dinu, and G. Kruszewski, "Don't count, predict! A systematic comparison of context-counting versus context-predicting semantic vectors," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, 2014, pp. 238–247. [Online]. Available: http://anthology.aclweb.org/P/P14/P14-1023.pdf

[54] Z. S. Harris, "Distributional structure," *Word*, 1954. [Online]. Available: http://psycnet.apa.org/psycinfo/1956-02807-001

[55] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2014, pp. 1532–1543. [Online]. Available: http://nlp.stanford.edu/projects/glove/glove.pdf

[56] O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Trans. Assoc. Comput. Linguistics*, vol. 3, pp. 211–225, 2015. [Online]. Available: https://tacl2013.cs.columbia.edu/ojs/index.php/tacl/article/view/570

[57] J. Zhang, J. Salwen, M. R. Glass, and A. M. Gliozzo, "Word semantic representations using Bayesian probabilistic tensor factorization," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2014, pp. 1522–1531. [Online]. Available: http://www.aclweb.org/anthology/D14-1161

[58] S. Li, J. Zhu, and C. Miao, "A generative word embedding model and its low rank positive semidefinite solution," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2015, pp. 1599–1609. [Online]. Available: http://aclweb.org/anthology/D15-1183

[59] M. Faruqui and C. Dyer, "Improving vector space word representations using multilingual correlation," in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2014, pp. 462–471.

[60] Z. Wang, J. Zhang, J. Feng, and Z. Chen, "Knowledge graph and text jointly embedding," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2014, pp. 1591–1601. [Online]. Available: http://aclweb.org/anthology/D14-1167

[61] F. Sun, J. Guo, Y. Lan, J. Xu, and X. Cheng, "Inside out: Two jointly predictive models for word representations and phrase representations," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2821–2827. [Online]. Available: http://www.aaai.org/Conferences/AAAI/2016/Papers/15Sun11783.pdf
[62] P. Rastogi, B. Van Durme, and R. Arora, "Multiview LSA: Representation learning via generalized CCA," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Human Language Technol.*, 2015, pp. 556–566. [Online]. Available: http://aclweb.org/anthology/N15-1058
[63] S. L. Hyland, T. Karaletsos, and G. Rtsch, "A generative model of words and relationships from multiple sources," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016, pp. 2622–2629. [Online]. Available: http://www.aaai.org/Conferences/AAAI/2016/Papers/14Hyland12446.pdf
[64] P. D. Turney and P. Pantel, "From frequency to meaning: Vector space models of semantics," *J. Artif. Intell. Res.*, vol. 37, pp. 141–188, 2010.
[65] W. Yin and H. Schtze, "Learning word meta-embeddings," in *Proc. 54th Annu. Meeting Assoc. Comput. Linguistics*, 2016, pp. 1351–1360. [Online]. Available: http://aclweb.org/anthology/P16–1128
[66] T. Pedersen, S. Patwardhan, and J. Michelizzi, "WordNet:: Similarity - measuring the relatedness of concepts," in *Proc. 19th Nat. Conf. Artif. Intell. 16th Conf. Innovative Appl. Artif. Intell.*, 2004, pp. 1024–1025. [Online]. Available: http://dl.acm.org/citation.cfm?id=1614037
[67] D. Lynott and L. Connell, "Modality exclusivity norms for 423 object properties," *Behavior Res. Methods*, vol. 41, no. 2, pp. 558–564, 2009. [Online]. Available: http://link.springer.com/article/10.3758/BRM.41.2.558
[68] D. Lynott and L. Connell, "Modality exclusivity norms for 400 nouns: The relationship between perceptual experience and surface word form," *Behavior Res. Methods*, vol. 45, no. 2, pp. 516–526, 2013. [Online]. Available: http://link.springer.com/article/10.3758/s13428-012-0267-0
[69] M. Brysbaert, A. B. Warriner, and V. Kuperman, "Concreteness ratings for 40 thousand generally known English word lemmas," *Behavior Res. Methods*, vol. 46, no. 3, pp. 904–911, 2014. [Online]. Available: http://link.springer.com/article/10.3758/s13428-013-0403-5
[70] R. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4444–4451. [Online]. Available: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14972
[71] R. Speer and C. Havasi, "Representing general relational knowledge in ConceptNet 5," in *Proc. 8th Int. Conf. Language Resources Eval.*, 2012, pp. 3679–3686. [Online]. Available: http://redirect.subscribe.ru/_/-/www.lrec-conf.org/proceedings/lrec2012/pdf/1072_Paper.pdf
[72] J. Wieting, M. Bansal, K. Gimpel, K. Livescu, and D. Roth, "From paraphrase database to compositional paraphrase model and back," *Trans. Assoc. Comput. Linguistics*, vol. 3, pp. 345–358, 2015.
[73] P. Nakov, Z. Kozareva, A. Ritter, S. Rosenthal, V. Stoyanov, and T. Wilson, "SemEval-2013 task 2: Sentiment analysis in Twitter," in *Proc. 7th Int. Workshop Semantic Eval.*, 2013, pp. 312–320.
[74] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *Proc. 10th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2004, pp. 168–177. [Online]. Available: http://doi.acm.org/10.1145/1014052.1014073
[75] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources Eval.*, vol. 39, no. 2/3, pp. 165–210, 2005. [Online]. Available: http://dx.doi.org/10.1007/s10579-005-7880-9
[76] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proc. 43rd Annu. Meeting Assoc. Comput. Linguistics*, 2005, pp. 115–124. [Online]. Available: http://acl.ldc.upenn.edu/P/P05/P05-1015.pdf
[77] R. Socher, et al., "Recursive deep models for semantic compositionality over a sentiment treebank," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2013, pp. 1631–1642.
[78] X. Zhu, S. Kiritchenko, and S. M. Mohammad, "NRC-Canada-2014: Recent improvements in the sentiment analysis of Tweets," in *Proc. 8th Int. Workshop Semantic Eval.*, 2014, pp. 443–447. [Online]. Available: http://www.aclweb.org/anthology/S/S14/S14-2.pdf#page=463
[79] S. Baccianella, A. Esuli, and F. Sebastiani, "SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining," in *Proc. Int. Conf. Language Resources Eval.*, 2010, vol. 10, pp. 2200–2204. [Online]. Available: http://www.researchgate.net/profile/Fabrizio_Sebastiani/publication/220746537_SentiWordNet_3.0_An_Enhanced_Lexical_Resource_for_Sentiment_Analysis_and_Opinion_Mining/links/545fbcc40cf27487b450aa21.pdf
[80] L. Gatti, M. Guerini, and M. Turchi, "SentiWords: Deriving a high precision and high coverage lexicon for sentiment analysis," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 409–421, Oct.–Dec. 2016.
[81] D. Tin Vo and Y. Zhang, "Don't count, predict! An automatic approach to learning sentiment lexicons for short text," presented at the *54th Annu. Meet. Assoc. Comput. Linguistics*, Berlin, Germany, 2016.

Minglei Li receives the BE degree in mechanical engineering, in 2011 and the ME degree in mechanical and electrical engineering, in 2014 from the Huazhong University of Science and Technology, Wuhan, China. Currently, he is working toward the PhD degree in the Department of Computing, The Hong Kong Polytechnic University. His research interests include natural language processing, emotion analysis, computational linguistic, and applied machine learning.

Qin Lu is currently a professor with the Hong Kong Polytechnic University. Her main research works are in computational linguistics. That is, using computational methods to process Chinese text, extract useful information, and build Chinese NLP related resources. Her expertise is in lexical semantics, text mining, opinion analysis, and knowledge discovery.

Yunfei Long received the double bachelor's degree in both computer science and linguistics from JiLin University, Changchun, China, in 2013 and the Msc degree in cognitive science from the University of Edinburgh, United Kingdom, in 2015. He is currently working toward the PhD degree in the Department of Computing, The Hong Kong Polytechnic University. His current research interests include natural language processing, neural network, and social media analysis.

Lin Gui is a lecturer in the College of Mathematics and Computer Science, Fuzhou University. He received his BS degree from Nankai University and the MS, PhD degree from Harbin Institute of Technology. His research areas include natural language processing, sentiment analysis, emotion computation and machine learning.