Neurocomputing 210 (2016) 227–236

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Sentiment analysis for Chinese microblog based on deep neural networks with convolutional extension features

Xiao Sun [a,*], Chengcheng Li [a], Fuji Ren [a,b]

[a] School of Computer and Information, Hefei University of Technology, TunXi Road No. 193, 230009 Hefei, Anhui, China
[b] Faculty of Engineering, The University of Tokushima, 770-8506 Tokushima, Japan

ARTICLE INFO

Article history:
Received 15 April 2015
Received in revised form
24 September 2015
Accepted 16 February 2016
Available online 10 June 2016

Keywords:
DNN
Sentiment analysis
Microblog conversation
RBM
ClassRBM

ABSTRACT

Related research for sentiment analysis on Chinese microblog is aiming at the analysis procedure of posts. The length of short microblog text limits feature extraction of microblog. Tweeting is the process of communication with friends, so that microblog comments are important reference information for related post. A contents extension framework is proposed in this paper combining posts and related comments into a microblog conversation for features extraction. A novel convolutional auto encoder is adopted which can extract contextual information from microblog conversation as features for the post. A customized DNN (Deep Neural Network) model, which is stacked with several layers of RBM (Restricted Boltzmann Machine), is implemented to initialize the structure of neural network. The RBM layers can take probability distribution samples of input data to learn hidden structures for better high level features representation. A ClassRBM (Classification RBM) layer, which is stacked on top of RBM layers, is adopted to achieve the final sentiment classification label for the post. Experimental results show that, with proper structure and parameters, the performance of proposed DNN on sentiment classification is better than state-of-the-art surface learning models such as SVM or NB, which proves that the proposed DNN model is suitable for short-length document classification with the proposed feature dimensionality extension method.

1. Introduction

Microblog has been widely popularized in recent years. It acts not only as a way for interaction and communication among people, but also as a way to express individual emotions at work or in daily life. Some related studies measure the preferences and political orientations of microblogger through sentiment analysis of microblog text. The emotion polarity of bloggers might reflect his or her hobbies and interests [1–4]. Microblog sentiment classification emerges as a challenging task [5–8]. The general realization of emotion polarity detection commonly includes extraction of features [9–11] and selection of machine learning methods [12–15].

Extraction of features from text is a process that produces some vector representations [16–19] for essential features and characteristics of original observation contents (or texts). At present, there are commonly two types of feature extraction methods: literal contents based method [20–23] and external knowledge information based method [24–28]. The methods based on literal contents are to calculate the probabilities of character or word sequences in text under the precondition of target classification using statistics methods. For example, a N-gram system is designed to solve the web categorizing problem [16], which takes Chinese character as minimal unit of textual feature representation for Chinese, as Chinese word segmentation is considered to cause loss of semantic or syntactic information and some Chinese word segmentation could not handle out-of-vocabulary words which might lead to the problem of word meaning ambiguities. In contrast, most researchers adopt word level frequencies and features because word is the least unit of semantic representation [22,23]. Related algorithms are adopted to calculate the feature weight of a word, such as CHI, IG, and TF-IDF. These textual feature extraction methods are widely adopted. These methods mostly consider morphological connections between words but the meaning embedded in texts is ignored or seldom considered. Sentiment is semantic information embedded in text, so some deep learning method should be adopted to detect such deep semantic knowledge. The method based on external knowledge information is applied to analyze word sense, semantics and grammatical structure and so on [27,28]. The knowledge information mainly includes system sentiment word lexicons, expression rules, syntactic model, etc. The sentiment polarities of words are based

[*] Corresponding author.
E-mail address: sunx@hfut.edu.cn (X. Sun).

http://dx.doi.org/10.1016/j.neucom.2016.02.077
0925-2312/© 2016 Elsevier B.V. All rights reserved.

on subjective opinion under certain circumstances or in specific field, which means that a certain sentiment dictionary is usually domain related and not universal. Choi [17] proposed a novel method to transform existing sentiment lexicon for specific domain into a new one to reflect the characteristics of new domain more directly, so that it can be used in cross-domain sentiment analysis. Pang [18] and Saif [19] use sentiment words as additional features combined with domain characteristics, and they find that semantic word features produce better recall and F-score than word frequency features. Ye [25] presents an improved semantic oriented approach for sentiment classification of Chinese movie reviews. This semantic approach introduces two-word phrases patterns based on POS (parts-of-speech), which is primarily applied in English movie review classification. Because of different language expression structure of English and Chinese, the method does not have great effects on classification results. Hiroshi [22], Wilson [23] and Subrahmanian [24] find the dependency relationship among words through building syntactic parsing tree for a sentence. The parsing tree includes semantic structure for the whole sentence and grammatical role of words. The word features are adjusted by considering sentence modifiers and syntactic structure information, which are treated as classification features. By adopting semantic features, the approach is proved to have better performances than literal content-based feature extraction methods.

The selection of machine learning methods is the process of selecting a proper classifier with tuned parameters for specific task. From the structure aspect, there are mainly two kinds of machine learning methods: surface learning models and deep learning models. Surface learning models can be regarded as a model with single hidden layer. For example, SVM, Boosting and Logistic Regression etc. belong to surface machine leaning model. Ye [25] compares three surface machine learning algorithms (SVM, Naïve Bayes and N-gram model) for review sentiment classification and proves that all three approaches reached accuracies of at least 80% on specific dataset. In Abbasi's work [26], the utility of semantic and syntactic features is integrated according to the characteristics of target text and learned by EWGA (Entropy Weighted Genetic Algorithm) and SVM model. The results indicate high performance in their datasets. These surface models require large number of labeled experiment data, and the common characteristic of such models is the limited ability of complicated object function or data representation. Deep leaning model could learn complex object function through building a deep nonlinear multi-layer network structure. The deep learning model has brought lots of attentions recently as a hot research topic in many fields such as image and speech processing. Xavier Glorot [10] proposes a deep learning approach which shows linear classifiers trained with higher-level learned feature representation of reviews outperforming traditional surface learning methods. Deselaers [11] and Chen [5] testify the validity of deep learning on NLP tasks. The deep learning model is showing its abilities of learning deep structure knowledge such as semantic information embedded in text, so deep learning models could solve classification problem with relatively better performance than surface learning models.

In tasks of micriblog sentiment analysis, the brevity of a post limits the feature expression and the feature vector extracted is excessively sparse as the average length of Chinese microblog post is 13 Chinese Hanzi in average. Short length of a post caused that traditional feature extraction method is hard to build reasonable representation of a sentence for machine learning. This paper presents a convolutional content extension feature extraction method for Chinese microblog sentiment classification. Compared with traditional feature extraction approaches discussed above, the comments of a post are adopted to expand feature dimensionality for sentiment analysis of post in microblog. The proposed method combines post and comments to form a new microblog conversation while extracting microblog textual feature to extend the feature dimensionality and solve the feature sparseness problem. This paper proposes a feature auto-encoder, named Conversation to Sentence Convolutional Auto Encoder (ConCAE), to extract the context information of a post from microblog conversation. Furthermore, a specific DNN model is constructed by stacking a Classification RBM [3,13,14] layer and several RBM layers together. At last, we design some experiments to choose proper feature set and optimal structure (including parameters) for the proposed DNN model. In the experiments, some comparisons are also performed on some public corpus such as Sina weibo to prove the effectiveness of proposed methods for microblog sentiment analysis.

2. Extended feature extraction for short text in microblog

Chinese microblogging posts are short texts, from which features extracted are limited because of its brevity [30]. The commonly used method for long text classification such as bag of word might cause the problem of feature sparseness [31]. In order to solve such problem, this paper proposes a content extension method for feature extraction in Chinese microblog sentiment analysis problem. For a post, it might be followed by several comments. These comments are responses or references to the emotion of microblogger. The content of a post is the key feature for sentiment analysis and its comments are used as assistance features. We combine a post with its comments into a microblog conversation and extract sentiment related information from the microblog conversation by ConCAE. The content extension method employs filtering approach to obtain proper sized microblog conversation which is composed of the post and its fixed-sized comments. The first step is to capture the emotion and semantic information of words. It is supposed that every post and its comments is consisted by m word $\{w_1, w_2, \ldots, w_m\}$, the word information w_i is the integration of emotion information w_{ei} and corresponding semantic information w_{si}. We compose size-fixed word bag and expression information to represent the feature of a post and its comments. Then the post feature V_{post} and comment feature sets $\{V_{com}^1, V_{com}^2, \ldots, V_{com}^L\}$ can be obtained, where L is the number of comments. After word bag and expression information are obtained, the auto-encoding network is adopted to capture context information for the post with its comments.

2.1. Word-level feature extraction

The aim of word feature extraction is to acquire emotional, semantic and integration information of words [12]. The steps of word information extraction for a post are based on word segmentation and semantic analysis of the sentence. The emotional information of a word can be obtained from prior knowledge, such as a system lexicon or dictionary. Semantic information of each word in microblog is semantic role label, which can be obtained by analyzing the semantic structure of the microblog content.

2.1.1. Emotion information of words

The steps of emotion feature extraction are based on word segmentation and semantic analysis for the sentence. Word is the smallest meaning expression units in English and Chinese, while Chinese word expression is composed of Chinese Hanzi (single character) or words. Some Chinese Hanzi has no meanings. For Chinese microblog sentiment analysis, the first step is Chinese word segmentation and part-of-speech (POS) tagging. All

Table 1
Seed set of degree words and negative words.

Degree words	Negative words
太(too)非常(much)十分(completely)	没有(without)非(not)无(none)勿(do not)
特别(especially)尤其(particularly)	不用(no need)不够(not enough)
相当(rather)越来越(more and more)	不(no)未曾(have not)未尝(not ever)
更(more)稍微(slightly)极其(extremely)	从不(never ever)从未有过(never have)
几乎(almost)最(most)很(very).	尚未(not yet)绝非(not at all)
	从来不(never)真(so)

emotional words in microblogs are marked by an extra emotional dictionary in which every word is labeled with an EMV (emotional degree score) which represents the emotion information of word [29].

2.1.2. Semantic information of words

The next step is semantic analysis of microblogs after Chinese word segmentation and POS tagging. These emotional words are considered with their modifiers in the whole sentence. The modifiers refer to the existences of negative words, degree words and emotional words. Negative words, which can change or reverse the polarity of emotional words, enhance or weaken emotional degree of the emotional word. We have collected 57 commonly used degree words and 37 negative words as seed words set for the experiments. The seed words set is further expanded by some word-embedding methods such as Neural Network Language Model (NNLM) [32] and word2vec package by Google [33].

All degree words and negative words in Table 1 are assigned with a prior weight in order to measure their impact on sentiment. The value of degree words is represented as degree impact value (DIV), and the value of negative words is represented as negative impact value (NIV). A dependency parser is adopted from Stanford NLP Parser toolkit [34] which could analyze the syntactic structure of a Chinese sentence. Through contracting the semantic tree of microblog post sentence, these degree words and negative words which might modify the expression of emotional words in the post can be labeled. The emotional value of an emotional word (w) will change. The final emotional value (FEMV) of an emotional word can be calculated according to the following equation:

$$w_{si}(w) = \prod_{i=1}^{m} DIV_i \prod_{j=1}^{n} NIV_j \tag{1}$$

where m is the number of degree words, which modifies w, and n is the number of negative words. At last, the final emotional value of emotional words would be adopted in word-level feature extraction.

2.1.3. Integration information of words

Semantic information might transform the polarity of emotional words, enhancing or weakening the original emotional degree of the emotional word, so there is a meaningful relation between emotion information and semantic information for each and every word.

$$w_i = w_{ei} \cdot w_{si} \tag{2}$$

Integration information takes emotion and semantic information of words into account for microblog sentiment analysis, which specially considers the expression feature of words under particular context. In Eq. (2), w_{ei} means emotional information and w_{si} means semantic information. w_{ei} and w_{si} are both scalars and the '$\cdot$' means scalar multiplication.

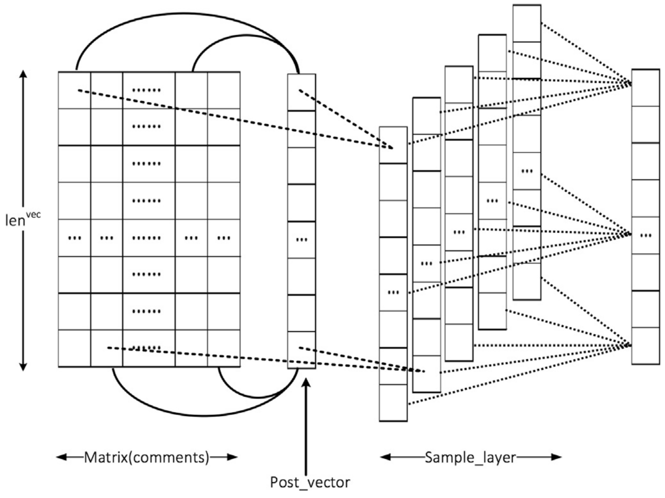


Fig. 1. Feature extraction structure for microblog conversation.

2.2. Microblog feature extraction

For a post in microblog, it could receive varying amounts of comments from other microbloggers or friends in microblog. In order to extract enough feature from short text of a post and avoid feature sparse problem. Some content (or feature) extension methods should be considered. The content extension means extension of length-limited post information by adopting size-fixed comments sets of the post in chronological order as a microblog conversation.

2.2.1. Microblog conversation

For sentiment analysis of each post, the post and its comments form a microblog conversation. We built a rule-based filtering method which could validate whether a comment is a reply to the post or not. After filtering, the last L comments of the post are selected to form the microblog conversation. We can get the post feature V_{post} and comments feature V_{com} according to Section 2.1. The comments feature sets $\{V_{com}^1, V_{com}^2, \ldots, V_{com}^L\}$ could be transformed into comments matrix M_{coms}, which could be seen on the left part of Fig. 1. The column of M_{coms} is L.

2.2.2. Context information of a post

Traditional methods of extracting context information of a long sentence are unifying sentences information around the sentence with certain windows size. The context information around a sentence usually has close semantic relations with each other. A microblog conversation (a post with its comments) belongs to group conversations and has the same characteristics as long sentences in common documents. For the task of context information extraction of a post, the context information mainly reflects semantic links between a post and its comments, without considering interactions among comments which has weak semantic relations with each other. The feature of a microblog conversation is a matrix and should be transformed into a vector for sentiment analysis. The paper settles this problem by proposing a ConCAE approach. In ConCAE, the idea of convolutional neural network for extracting features from images is adopted. The extracting process of ConCAE is described in Fig. 1, the ConCAE approach produces one dimensional feature vector from post and its comments of a microblog conversation and then combines them using sampling method to get a context information representation for the post.

Given a microblog conversation which contains a post and its L comments, we transform the microblog conversation into post feature vector V_{post} and comments feature vector sets $\{V_{com}^1, V_{com}^2, \ldots, V_{com}^L\}$, which can be treated as a matrix of comments M_{coms}. The input of ConCAE is composed of the post vector V_{post} and of comments features matrix M_{coms}, which is also can be treated as a matrix M_{con}. In matrix M_{con}, each row is a concatenated vector of V_{post} with each V_{com}^l in the following equation:

$$V_{con}^l = \left[V_{post};\, V_{com}^l\right] \tag{3}$$

L concatenated vector V_{con} can be obtained as input M_{con} for convolutional layer. The idea of convolutional neural network is adopted to transform the microblog conversation matrix M_{con} into a vector for the post with its L comments. The computing process of convolutional layer for each concatenation vector V_{con}^l is shown in the following equation:

$$v_f^l = f\left(W^0 V_{con}^l + b^0\right) \tag{4}$$

where W^0 is the weight matrix for input vector V_{con}^l of convolutional layer. $f(\cdot)$ is an activation function for output such as sigmoid. b^0 respects the bias terms. The convolutional layer shares the same weighting W^0 and b^0 with different input concatenation vector.

After the first convolutional layer, we can get a sequence of concatenation vector $\{V_f^1, V_f^2, \ldots, V_f^L\}$. The sample layer will sample the same position of these concatenation vectors by byte, and then different weights for sampling points could be co-trained when ConCAE is stacked with other supervised learning model such as SVM.

3. DBN model introduction

This paper builds a DNN model, which combines supervised learning with unsupervised learning, for Chinese microblog sentiment classification with extended multi-modality features. The DNN model is composed of a ClassRBM layer and several RBM layers. These RBM layers are used to achieve better representation of input data. The training process of these RBM layers is: first, each RBM layer is trained layer by layer as an autoencoder to obtain prior weights. The hidden layer of each RBM layer is used as the input of next RBM layer. The next RBM is trained as an autoencoder to obtain its prior weights like the former layer. Several RBM layers are acting as an unsupervised model. The deep structure of unsupervised model will try to learn the deep representation for input. ClassRBM is adopted as the top supervised level for classification. RBM and ClassRBM are co-trained to get the final weights or parameters. ClassRBM simplifies the process of classification without training extra classifier. Similarity structure between RBM and ClassRBM makes it easier to pass parameters between supervised and unsupervised layers. Comparisons have been performed between the performance of DBN and SVM, and the result shows that by choosing proper structure and parameters, DNN could perform better than traditional surface machine learning.

3.1. RBM and classRBM

RBM is a simplification of BM (Boltzmann Machine). RBM is a typical two-layer neural network: a hidden layer H and a visual layer V. The hidden layer and the visual layer are connected with each other. The units in the same layer are independent of each other. The process of RBM training can be described in the following procedure: the feature vector of visual layer maps to the hidden layer, these visual layer units are rebuilt by the hidden layer, these new visual units can remap to hidden layer and then RBM would get new hidden units. RBM deletes the connection among hidden units in order to shorten the iterative training time and accelerates the process of reaching thermal equilibrium. RBM is generally used to extract feature and optimize feature. Automatic feature extraction resolves the problem caused by ill-considered of artificial feature selection. In the process of feature extraction: the activations of these hidden units which are obtained by RBM training are used as new input features, or sometimes the combinations of activations and initial features can replace the original input data. Then some supervised learning would perform the final classification, such as SVM, EM and so on. RBM could also act as a tool to reduce the dimensionality for feature extraction here. RBM has another application: the training process of traditional neural network involves the iteration of weight matrix and offset. Inappropriate initial value might cause lack-training or overtraining which might easily lead to local minimum, leading to the loss of global minimum. The weight matrix and offset trained by RBM as the initial value of BP neural network could bring good results.

The commonly used RBM is Bernoulli–Bernoulli RBM which would satisfactorily handle binary and discrete feature space. But it is not applicable to continuous data. Meanwhile, the microblog feature introduced in the above sections is continuous. So Gaussian–Bernoulli RBM is adopted to solve the problem of continuous value input. Gaussian–Bernoulli RBM will change the binary variables of visual input layer to follow Gaussian distributions, so that it can convert the continuous random variables into binary random variables. Then Bernoulli–Bernoulli RBM can be used to handle binary input data. The first layer of multi-layer RBMs would adopt Gaussian–Bernoulli RBM and the rest layers are Bernoulli–Bernoulli RBM.

The energy function of Gaussian–Bernoulli RBM is defined as follows:

$$E(v, h) = -\sum_{i=1}^{V}\frac{(v^i - a^i)^2}{2\sigma^2} - \sum_{j=1}^{H} b_j h_j - \sum_{i=1}^{V}\sum_{j=1}^{H} w_{ij}\frac{v_i}{\sigma}h_j \tag{5}$$

The energy function of Bernoulli–Bernoulli RBM is:

$$E(v, h) = -\sum_{i=1}^{H} a_i v_i - \sum_{j=1}^{H} b_j h_j - \sum_{i=1}^{H}\sum_{j=1}^{H} w_{ij} v_i h_j \tag{6}$$

where v is the visual input feature $(v_1, \ldots, v_n)$, with the hidden units $h = (h_1, \ldots, h_m)$, σ are variances of input feature, and parameters $\theta = (a, b, W)$ need to be learned by training.

For both kinds of RBM layers, its input feature is the combinations of input features and hidden unit activations of the previous layer. $f(\cdot)$ is the activation function.

$$h^{(k)} = f\left(w^{(k)} v^{(k)} + b^{(k)}\right) \tag{7}$$

$$v^{(k+1)} = \left(v^{(k)}, h^{(k)}\right) \tag{8}$$

The step of feature handling is $v^{(0)} \to (h^{(0)}, v^{(0)}) \to (h^{(1)}, h^{(0)}, h^{(0)}) \cdots \to (h^{(k)}, \ldots, h^{(0)}, h^{(0)})$. Taking into account the cost of memory and calculation, the number of hidden layer units should be reduced.

The next step is self-training process of RBM based on CD (Contrastive Divergence) algorithm. First, the initial feature vector is assigned to the input layer, the next step is to calculate the conditional probability distribution of the hidden layer to input layer; after that, the conditional probability distribution of the input layer to hidden layer is computed in the same way, then the input units are reassigned, the above steps are repeated and stopped according to preset threshold.

ClassRBM is a three layers network: visual layer, hidden layer and input layer, which is used to calculate the conditional probability distribution of the label given input data, so as to fit the joint distribution of them. Besides directly achieving the label, ClassRBM also ensures that the features learning from the neural network are useful for classification. The difference between ClassRBM and RBM is that a sample class units Y is added. The input data of ClassRBM is composed of the sample feature representation vector and sample label vector $(0, 0, \ldots, 1, \ldots, 0)$. The first step is to redefine the energy function.

$$E(y, v, h) = -c^T v - d^T h - h^T W v - e^T \vec{y} - h^T U \vec{y} \tag{9}$$

where the number of parameters becomes $\Omega = (c, d, e, U, W)$. Mapping relations can be expressed as follows:

$$h = sigm(d + U\vec{y} + Wv) \tag{10}$$

where $sigm()$ is the same as sigmoid function $f()$. The conditional probability distribution among input layer and hidden layer or sample class could be obtained appropriately, but it is hard to calculate their mutual joint probability distributions.

$$p(h|y, v) = sigm(d + U\vec{y} + Wv) \tag{11}$$

$$p(y|h) = \frac{\exp(e_{y_*} + h^T U_{j\,y_*})}{\sum_{y_*} \exp(e_{y_*} + h^T U_{j y_*})} \tag{12}$$

The most commonly used training method for deep neural network is CD algorithm, which is also adopted in this paper. The stochastic gradient obtained by CD is shown in the following equation:

$$h_t = sigm(d + U_t \vec{Y_t} + W_t v_t) \tag{13}$$

$$h_{t+1} = p(h_t | y_t, v_t) \tag{14}$$

$$y_{t+1} = p(y_t | h_t) \tag{15}$$

$$v_{t+1} = p(v_t | h_t) \tag{16}$$

$$\Omega = \Omega - \lambda(E'_\Omega(y_t, v_t, h_t) - E'_\Omega(y_{t+1}, v_{t+1}, h_{t+1})) \tag{17}$$

where λ is the learning rate. The DNN model is co-trained with ConCAE to get the global optimal parameters for the whole framework.

3.2. The structure of DNN

Multilayer deep structure could solve the problem of complex function representation and reduce generalization ability of shallow structure [35–37]. Multilayer network clearly highlights the importance of feature learning and data representation. The features automatically learned by deep structure can better reflect the inner and deeper structure of input data, which is more beneficial for classification. Yet inadequate structure of deep neural network might bring negative effects, such as high computational complexity with plenty amount of arguments and over-fitting. The paper trains a DNN model to complete emotional classification problem for microblog.

The DNN model is composed of a single-layer ClassRBM and multi-layer RBMs. Fig. 2 shows the structure of the DNN. First, multi-layer RBMs are trained layer by layer to optimize the input features like traditional RBMs training process to get prior weights for the final co-training. The output of previous RBM hidden layer is adopted as the input data for next RBM input layer. The feature representation of data in original space is transformed into another space by each layer of training, which improves the ability of feature representation for the final sentiment classification. The next step is the training of ClassRBM to get its prior weights. Then, the RBMs layers and ClassRBM layers are co-trained to tune the weights and get final parameters for classification. The output features of the top layer of RBM are set as the input data for ClassRBM layer. The classification layer is trained in the way of supervised learning and the discriminative performance of the feature representation is monitored in real-time. Initial concrete feature vector is converted into the feature vector in the abstract space. This transformation of feature expression optimizes the sample parameter feature and streamlines the whole training process.

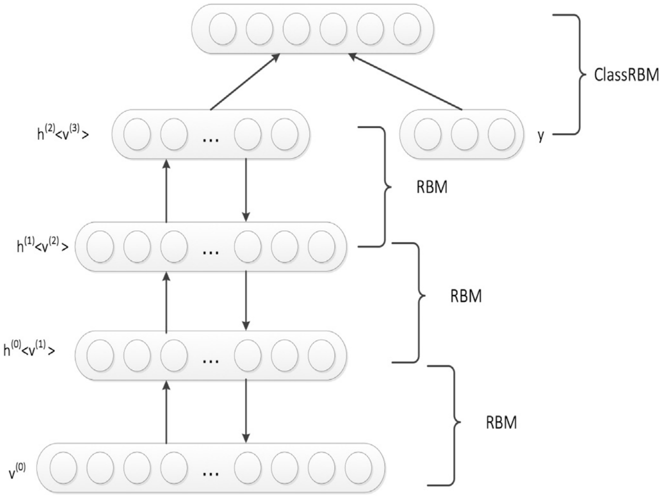


Fig. 2. The structure of DNN.

In deep learning, DNN is the deep neural network with cascading Boltzmann model. The task of each RBM layers is used to accomplish abstract representation of input feature. The unsupervised learning deep model is constructed by bottom-up training process. The major training concepts of top-layer ClassRBM are that append bias contribution parameter and corresponding contribution value to energy function and provide an independent architecture for unsupervised learning problem.

4. Experiments and results

In order to make fair comparisons and prove the effectiveness of proposed deep learning framework, three different corpus are adopted in two different language to validate the effect of content extension method described above: Twitter dataset,[1] Sina microblog sentiment corpus (SMSC)[2] and the corpus from The Fifth Chinese Opinion Analysis Evaluation (COAE2014) and the corpus from The Fifth Chinese Opinion Analysis Evaluation (COAE2014),[3] which includes fine grained sentiment labels for about 7000 examples in task four (micro-blog sentiment analysis) and the corpus are only labeled with two emotion labels (positive and negative). The Twitter datasets are from International workshop on semantic evaluation 2013 task 2, the twitter datasets include 1391 labeled twitters with comments [15–19]. The Sina microblog

[1] http://www.cs.york.ac.uk/semeval-2012/task2/index.php?id=data
[2] http://www.datatang.com/data/44059
[3] http://www.datatang.com/data/46475

Table 2
Experimental corpus.

Set	Dataset	Class	Train	Test
D1	Twitter database (in English)	Positive	455	106
		Negative	200	52
		Neural	460	118
D2	Sina microblog (in Chinese)	Positive	2523	504
		Negative	1760	367
		Neural	3180	423
D3	COAE2014 (in Chinese)	Positive	3069	445
		Negative	3427	474
		Neural	0	0

database mentioned is crawled from Sina microblog and manually labeled with three emotion labels (positive, negative or neural). We collected Sina microblog posts with its comments as analysis object and the number of comments for each microblog post is over 6. The details of the three different datasets are shown in Table 2.

To verify the effectiveness of proposed methods, experiments are designed for two purposes. The first is to prove that the method based on content extension feature extraction is more effective than single post feature extraction method especially for short text. The second aim is to prove that DNN, which is composed of RBMs and ClassRBM, with better selected features and fine-tuned parameters, could further improve the performance of classification than traditional surface learning models [20–23].

4.1. *Experiment 1: measuring the impact of content extension method*

This experiment is aiming at measuring the impact of content extension method for microblog post sentiment analysis. For this purpose, there are different experimental sets with several different combinations of features mentioned above on two datasets (D1 and D2). These optional features sets are shown in Table 3. There are two kinds of word-level features in Experiment 1: emotional information and integration information of words. For sentence-level feature, we adopted bag-of-word model to transform sentences into vectors. Post-only information and post-context combination information by ConCAE are adopted as Microblog features sets [24–27]. Therefore four different basic feature sets are shown in Table 3.

Table 4 shows the results of different feature combinations with machine learning models on two datasets. There are four different feature combinations with two kinds of word information and two types of post features. Meanwhile, SVM with optimal parameters (5-folder cross-violation) and one hidden layer DNN models (one layer RBM with 1500 hidden units) are adopted to perform the experiment. The size of system dictionary for bag-of-word is 2500. In order to compare the proposed method's performance with other models on the whole dataset, we adopted macro-average measuring method in Table 4. The precision (P), recall (R) and F-1 score are all calculated by macro-average method on three classes.

Table 3
Different feature sets.

Feature labels	Word-level feature sets
W1	Emotion-only information
W2	Integration information
Microblog feature sets	
P1	Post-only information
P2	Post-Context information

Table 4
Macro-average results of different combinations.

	SVM		KNN		NB		DNN	
Benchmarks	D1	D2	D1	D2	D1	D2	D1	D2
W1+P1								
P	0.651	0.635	0.572	0.588	0.643	0.651	0.621	0.618
R	0.624	0.655	0.589	0.567	0.618	0.619	0.605	0.619
F	0.637	0.649	0.58	0.577	0.63	0.635	0.613	0.618
W2+P1								
P	0.671	0.622	0.579	0.588	0.631	0.661	0.635	0.652
R	0.642	0.634	0.601	0.612	0.621	0.684	0.619	0.622
F	0.656	0.628	0.59	0.6	0.627	0.672	0.627	0.637
W1+P2								
P	0.684	0.681	0.601	0.603	0.619	0.634	0.711	0.699
R	0.671	0.694	0.579	0.588	0.611	0.608	0.641	0.634
F	0.677	0.687	0.59	0.595	0.615	0.621	0.674	0.665
W2+P2								
P	0.714	0.721	0.618	0.636	0.627	0.621	0.741	0.735
R	0.693	0.72	0.648	0.568	0.612	0.603	0.698	0.711
F	0.703	0.72	0.633	0.6	0.619	0.612	0.719	0.723

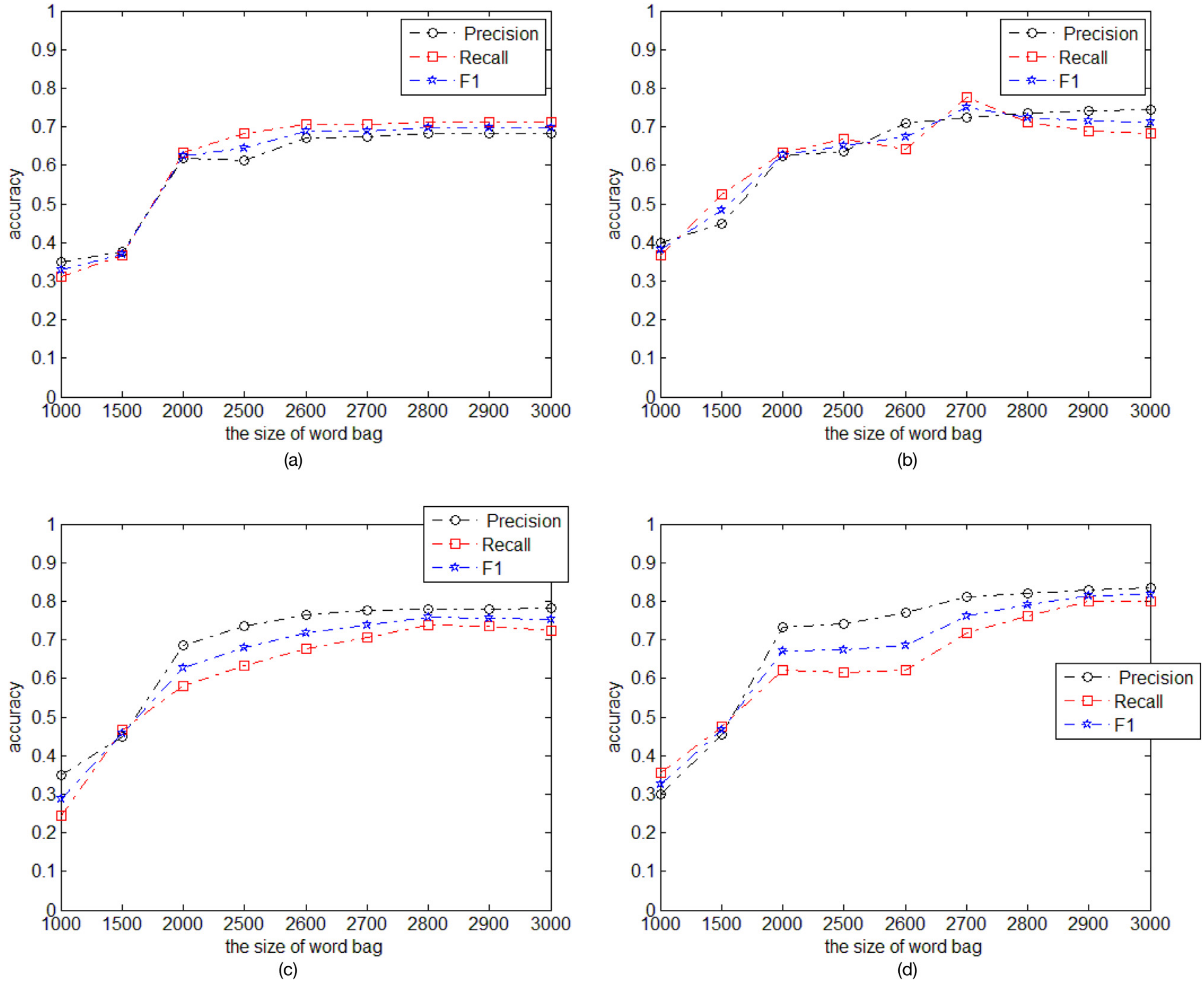
The first general result is adopted as experiment baselines, in which the semantic information is not involved for word-level features and context information is not involved for microblog features. The comparisons on different feature sets are also performed. The results show that the introduction of emotional word-level feature has no significant effects on the results, when only considering the emergences of emotional words and ignoring semantic information of words. Almost all machine learning models are improved on the accuracy of classification which might because the addition of context information. It is also shown that KNN algorithm is not sensitive to content extension. The classification results of NB model perform better when the feature extracting method only considers statistics information. Instead of modest gain which had been expecting, recognition rates dropped 2% with the introduction of context information.

From the results shown in Table 4, only emotional feature cannot represent the semantic information well. The optimal feature set combination is Integration information with Post-Context information of post. SVM and DNN model is more effective in sentiment classification task.

In the next experiment, the purpose is to select proper dimension of posts and comments features which is the size of system dictionary that bag-of-word method used. Four experiments are set to get the optimal dimension value:

(a) This experiment chooses second feature set (W2+P1) as initial features and the features are adopted in training SVM.
(b) This experiment combines integration information feature and microblog conversation feature as initial features which are adopted in training SVM.
(c) This experiment chooses second feature set (W2+P1) as initial features for training, the model is DNN, which includes a RBM layer and a ClassRBM layer.
(d) This experiment combines integration information feature and microblog conversation information feature as the initial feature. The DNN is composed of a RBM and a ClassRBM layer.

In these four experiments, for the sake of comparisons, we set the optimal parameters for machine learning models by 5-folder cross-violation. The dataset D2 is the experiment dataset. The optimal parameter of SVM in this experiment is ($\gamma = 3$, $coef = 1.5$). The number of hidden units of DNN is one layer RBM with 1500 hidden units. The P, R and F-1 are also calculated by macro-average



(a)

(b)

(c)

(d)

Fig. 3. The macro-average results with different size of word bag.

method on three classes to get the final performance on the whole database.

Fig. 3 shows the experimental results, as can be seen from these four sub-figures, general trends are mostly the same: the accuracy increases with the feature dimension value (from 1000 to 3000), accuracy grows fast when the dimension value is less than 2500 and slowly after 2600. When the value is under 1500, all the systems perform badly which means the feature vector cannot represent short microblog text well because of inadequate dimension. The reason why only one RBM layer is picked here is that the calculation complexity increases with the feature dimension value. Taking accuracy and computation complexity into account, the final experiment chose 2600 as the text-based modality feature dimension. Comparing Fig. 3(a), (b), (c) and (d), the performance of feature extension method is better than traditional feature extraction method. When the dimension value is limited to 1000, feature extension got more features than traditional feature extraction method to obtain better accuracy, which is reasonable for application.

To test the feature extension method above even further, we compared it with the rank-1 method based on sentiment phrase (SP) [27] in task 4 (micro-blog sentiment analysis) of The Fifth Chinese Opinion Analysis Evaluation (COAE2014 http://www.liip.cn/CCIR2014/). COAE2014 provides a famous open platform for researchers to compete on sentiment analysis for Chinese short text in social network. We use our deep model to perform experiment on task 4. As the training and testing data only contain microblog posts without comments, so we only adopt post-related feature sets in our deep model for this experiment. Table 5 shows the comparisons between our deep model with the best performance (medians results are also listed in the table) on task 4 of COAE2014. In Table 5, as there are two classes in COAE2014 corpus, we measure the performance and get the P, R and F-1 score on each class.

Table 5
The results on COAE2014.

	Positive			Negative		
Systems	P	R	F	P	R	F
Best	0.977	0.603	0.715	0.971	0.766	0.778
Medians	0.891	0.299	0.445	0.85	0.281	0.428
Our	0.862	0.730	0.791	0.727	0.89	0.800

Experiments results prove the effectiveness of our model comparing with sentiment phrase (SP) method, which shows the best performance in task 4 of COAE2014. Meanwhile, the results also show that comparing with SP algorithm, the proposed deep model showed better Recall and F1 on Negative samples than SP, even with only post-related features.

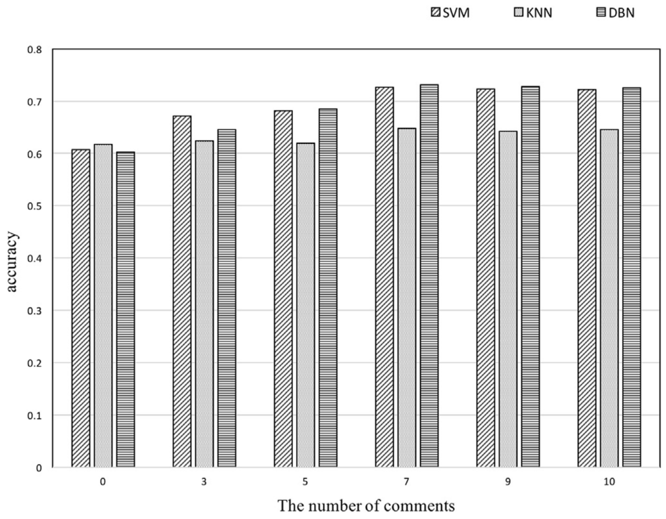


Fig. 4. The results with different size of comments.

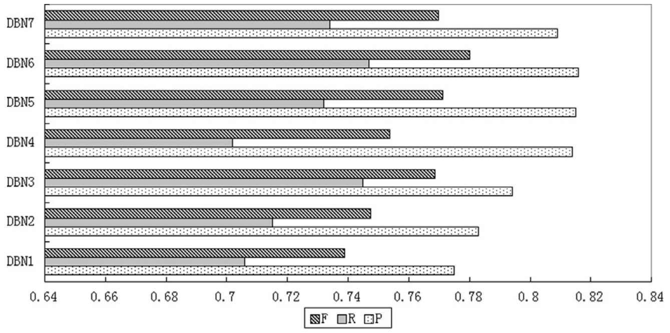


Fig. 6. The results of experiments on DNN with different depth.

For further proving the effectiveness of proposed content extension method, optimal context information of post from microblog conversation with different size of comments should be obtained. Fig. 4 show the results with different number of comments for the context information extraction of post on D2. Three models (SVM, NB and DNN) are adopted here. The best feature set obtained from the former experiment W2+P2 is adopted. We use the micro-average, which is the same as accuracy because there are more than two classes in D2 corpus. These results indicate a correlation between context information of post and the size of comments. As comment size increases, the accuracy of classification increases until the size reaches 7. The accuracy decreases with the size increase after the number over 9. The comments are filtered by time. According to the increment scale of the conversation, more noise data sets might be introduced which might directly bring on the decrease of classification results.

4.2. Experiment 2: verifying the validity of DNN model

For the second purpose, the proper shapes and depth of DNN should be determined. This experiment is performed on D2 corpus and micro-average score is adopted to measure the performances (Precision, Recall and F-1), which is the same as accuracy as there are three classes, on the whole corpus. The results for various DNN shapes (seen in Fig. 5) are referred by the number of hidden units in top level of DNN. The DNN structure adopted in this paper is one top classRBM layer. It can be seen from the results that the DNN shape has little impact on the performance of DNN. When the number of hidden units (of classRBM) exceeds 500, the accuracy remains almost the same around 76% and then DNN shape slightly impact the performance of DNN in the following curve. Several experiments of DNN model with different depth are designed to further verify the effectiveness of deep network model for classification and to choose the optimal depth of DNN. We set the maximum depth as 7. The numbers of hidden units from bottom up are 2000, 1500, 1000, 1000, 500, 500, 500 respectively.

It can be seen from Fig. 6 that precision maintains a steady growth trend with growth of DNN depth, although the recall has obvious undulatory. The precision gradually declines while the depth of RBM layers surpasses 4 layers. When the layers of RBM reaches 7, precision tends to be lower than DNN model with less layers, which might be caused by insufficient training data. The complexity of nonlinear functions that need to be constructed grows with the increasing complexity of DNN model. Accordingly, the energy losses in such functions increase. The feature information might be lost during optimization, which might in return cause the decreased accuracy.

Finally, four sets of experiment are designed to compare the performances of DNN model with traditional surface model. DNN (i) (i stands for the number of DNN layers), SVM and NB are used to train on dataset D2 respectively. All the three experiments are compared with different emotions (positive, neutral and negative). DNN chooses the optimal structure obtained in the former experiments, while SVM and NB are tuned with their optimal parameters. This experiment is also performed on D2 corpus and macro-average score is adopted to measure the final performances (Precision, Recall and F-1 score).

Table 6 shows the results of emotion classification. P is precision and R is recall. It can be seen from the table that, the overall performance of DNN is better than SVM and NB even though they have almost equally score on neutral emotion classification. In the three different kinds of emotion classification, the result of neutral emotion is the worst, which might be caused by fuzzy Chinese expressions and complex semantic pragmatics of language used in

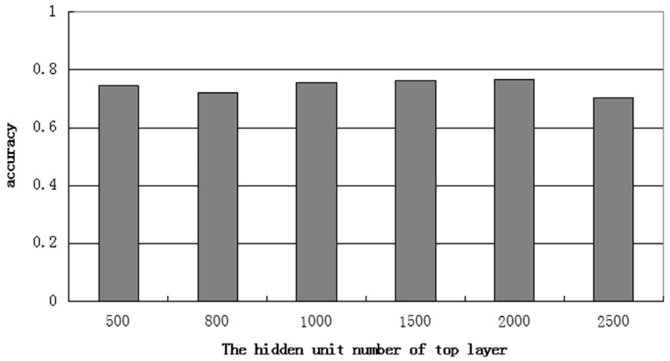


Fig. 5. DNN with different shapes (top hidden layer units).

Table 6
The result of emotion classification.

Labels	Benchmarks	SVM	DNN(1)	DNN(2)	NB
Positive	P	0.802	0.812	0.831	0.792
	R	0.693	0.732	0.773	0.701
	F1	0.747	0.77	0.786	0.747
Neutral	P	0.701	0.703	0.718	0.678
	R	0.682	0.673	0.645	0.632
	F1	0.691	0.715	0.705	0.654
Negative	P	0.793	0.827	0.831	0.754
	R	0.711	0.699	0.782	0.711
	F1	0.736	0.758	0.817	0.732

daily life, although there are some emotional words in neutral sentences. The accuracy of DNN(2) is better than SVM and DNN(1).

5. Conclusions and future work

In this paper, we proposed a deep belief nets (DNN) model with a content extension method ConCAE to solve features sparse problem caused by the limitation of length of short text in Chinese microblog for sentiment classification. The experimental results demonstrate that the proposed extended post feature extraction method is more effective than traditional feature extraction method, and the performance of deep learning with proper structure and fine-tuned parameter on sentiment classification could be better than traditional surface learning models such as SVM or NB. The performance could be further improved in the following aspects: all the relationships between bloggers could be calculated, which might also increase the computational complexity; meanwhile, the system time-consuming increases with the number of DNN layers during training, which could be optimized by adapted some faster training [38,39] and feature cutting algorithms [40–42].

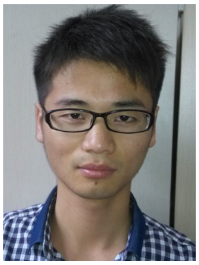
Acknowledgment

The work is supported by National Natural Science Funds for Distinguished Young Scholar (No. 61203315) and the Natural Science Foundation of Anhui Province (No. 1508085QF119), and also supported by the State Key Program of National Natural Science of China (No. 61432004). This work was partially supported by JSPS KAKENHI Grant No. 15H01712. This work was supported by the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (No. 201407345).

References

[1] Guohong Fu, Xin Wang, Chinese sentence-level sentiment classification based on fuzzy sets, in: Coling 2010, Beijing, August 2010 Poster Volume, 2010, pp. 312–319.
[2] Ch. Lin, Sh. Wang, Fuzzy support vector machines, IEEE Trans. Neural Netw. 13 (2) (2002) 464–471.
[3] H. Larochelle, M. Mandel, R. Pascanu, et al., Learning algorithms for the classification restricted Boltzmann machine, J. Mach. Learn. Res. 13 (2012) 643–669.
[4] A. Agarwal, B. Xie, I. Vovsha, O. Rambow, R. Passonneau, Sentiment analysis of twitter data, in: Proceedings of ACL 2011 Workshop on Languages in Social Media, 2011, pp. 30–38.
[5] Yu Chen, De-Quan Zheng, Tie-Jun Zhao, Chinese relation extraction based on deep belief nests, J. Softw. 23 (10) (2012) 2572–2585.
[6] A. Reyes, P. Rosso, Making objective decisions from subjective data: detecting irony in customers reviews, J. Decis. Support Syst. 53 (4) (2012) 754–760.
[7] A. Esuli, F. Sebastiani, SentiWordNet: a publicly available lexical resource for opinion mining, in: Proceedings of LREC 2006, 5th Conference on Language Resources and Evaluation, 2006.
[8] Hassan Saif, Yulan He, Harith Alani, Alleviating data sparsity for Twitter sentiment analysis, in: 2nd Workshop on Making Sense of Microposts, 2012.
[9] S. Bhuiyan, Social media and its effectiveness in the political reform movement in Egypt, Middle East Media Educ. (2011) 14–20.
[10] Glorot Xavier, Bordes Antoine, Bengio Yoshua, Domain adaptation for large-scale sentiment classification: a deep learning approach, in: Proceedings of the 28 International Conference on Machine Learning, Bellevue, WA, USA, 2011.
[11] T. Deselaers, S. Hasan, O. Bender, et al., A deep learning approach to machine transliteration, in: Proceedings of the Fourth Workshop on Statistical Machine Translation, Association for Computational Linguistics, 2009, pp. 233–241.
[12] Mnih Volodymyr, Larochelle Hugo, E. Hinton Geoffrey, Conditional restricted Boltzmann machines for structured output prediction, in: Proceedings of the Twenty-seventh Conference on Uncertainty in Artificial Intelligence (UAI'11), AUAI Press, Barcelona, Spain, 2011.
[13] Louradour, Larochelle Hugo, Classification of sets using restricted Boltzmann machines, in: Proceedings of the Twenty-seventh Conference on Uncertainty in Artificial Intelligence(UAI'11), AUAI Press, Barcelona, Spain, 2011.
[14] Larochelle Hugo, Erhan Dumitru, Courville Aaron, James Bergstra, Yoshua Bengio, An empirical evaluation of deep architectures on problems with many factors of variation, in: Proceedings of the Twenty-fourth International Conference on Machine Learning (ICML'07), 2007, pp. 473–480.
[15] L. Barbosa, J. Feng, Robust sentiment detection on twitter from biased and noisy data, in: 2010 Proceedings of COLING, 2010, pp. 36–44.
[16] A. Go, R. Bhayani, L. Huang, Twitter Sentiment Classification Using Distant Supervision, CS224N Project Report, Stanford, 2009.
[17] Y. Choi, C. Cardie, Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, vol. 2, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 590–598.
[18] B. Pang, L. Lee, S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, in: Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing, vol. 10, Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, pp. 79–86.
[19] H. Saif, Y. He, H. Alani, Semantic sentiment analysis of twitter, in: The Semantic Web—ISWC 2012, Springer, Berlin, Heidelberg, 2012, pp. 508–524.
[20] D. Barbagallo, C. Cappiello, C. Francalanci, et al., Semantic sentiment analyses based on reputations of web information sources, Appl. Semant. Web Technol. (2011) 325.
[21] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity in phrase-level sentiment analysis, in: Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2005, pp. 347–354.
[22] K. Hiroshi, N. Tetsuya, W. Hideo, Deeper sentiment analysis using machine translation technology, in: Proceedings of the 20th International Conference on Computational Linguistics, Association for Computational Linguistics, 2004, p. 494.
[23] T. Wilson, J. Wiebe, P. Hoffmann, Recognizing contextual polarity: an exploration of features for phrase-level sentiment analysis, Comput. Ling. 35 (3) (2009) 399–433.
[24] V.S. Subrahmanian, D. Reforgiato, AVA: adjective-verb-adverb combinations for sentiment analysis, IEEE Intell. Syst. 23 (4) (2008) 43–50.
[25] Q. Ye, Z. Zhang, R. Law, Sentiment classification of online reviews to travel destinations by supervised machine learning approaches, Expert Syst. Appl. 36 (3) (2009) 6527–6535.
[26] A. Abbasi, H. Chen, A. Salem, Sentiment analysis in multiple languages: feature selection for opinion classification in web forums, ACM Trans. Inf. Syst. (TOIS) 26 (3) (2008) 12.
[27] Vanzo Andrea, Croce Danilo, Basili Roberto, A context-based model for sentiment analysis in Twitter, in: Coling 2014, Dublin, Ireland, August 23–29, 2014, pp. 2345–2354.
[28] Zhen-jun, Pang, Li-bo Gao, Tian-fang Yao, Web Text Tendency Classification based on Sentiment Phrase, The Fifth Chinese Opinion Analysis Evaluation (COAE2014), 179-186.
[29] ⟨http://www.keenage.com/html/c_bulletin_2007.htm⟩, May 2015.
[30] Wang Meng, Liu Xueliang, Wu. Xindong, Visual classification by l1-hypergraph modeling, IEEE Trans. Knowl. Data Eng. (2015).
[31] Wang Jing, Wang Meng, et al., Online Feature Selection with Group Structure Analysis, TKDE 2015.
[32] Mikolov Tomas, Sutskever Ilya, Chen Kai, Corrado Greg, Dean Jeffrey, Distributed representations of words and phrases and their compositionality, in: Proceedings of NIPS, 2013.
[33] ⟨https://code.google.com/p/word2vec/⟩, May 2015.
[34] ⟨http://nlp.stanford.edu/downloads/lex-parser.shtml⟩, May 2015.
[35] Nie Liqiang, Wang Meng, Zhang Luming, Yan Shuicheng, Chua Tat-Seng, Disease inference from health-related questions via sparse deep learning, IEEE Trans. Knowl. Data Eng. (2015).
[36] Wang Meng, Gao Yue, Lu Ke, Rui Yong, View-based discriminative probabilistic modeling for 3d object retrieval and recognition, IEEE Trans. Image Process. 22 (4) (2013) 1395–1407.
[37] Nie Liqiang, Wang Meng, Gao Yue, Zha Zheng-Jun, Chua Tat-Seng, Beyond text QA: multimedia answer generation by harvesting web information, IEEE Trans. Multimed. 15 (2) (2013) 426–441.
[38] Wang Meng, Ni Bingbing, Hua Xian-Sheng, Chua Tat-Seng, Assistive Tagging: A survey of multimedia tagging with human–computer joint exploration, in: ACM Computing Surveys, vol. 4, no. 4, Article 25, 2012.
[39] Wang Meng, Hong Richang, Li Guangda, Zha Zheng-Jun, Yan Shuicheng, Chua Tat-Seng, Event driven web video summarization by tag localization and key-shot identification, IEEE Trans. Multimed. 14 (4) (2012) 975–985.
[40] Zha Zheng-Jun, Yu Jianxin, Tang Jinhui, Wang Meng, Chua Tat-Seng, Product aspect ranking and its applications, IEEE Trans. Knowl. Data Eng. 26 (5) (2014) 1211–1224.
[41] Yu Jian Xing, Zha Zheng-Jun, Wang Meng, Chua Tat-Seng, Aspect ranking: eliciting important product aspects from online consumer reviews, in: Annual Meeting of ACL (ACL), 2011.
[42] Li Zhisheng, Xiao Xiangye, Wang Meng, Wang Chong, Wang Xufa, Xie Xing, Towards the automatic categorization of yellow pages queries, ACM Trans. Internet Technol. 11 (4) (2012) 16–42.

Xiao Sun was born in 1980. He received the M.E. degree in 2004 from the Department of Computer Sciences and Engineering at Dalian University of Technology, and got his double doctor's degree in Dalian University of Technology (2010) of China and the University of Tokushima (2009) of Japan. He is now working as an associate professor in AnHui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine at Hefei University of Technology. His research interests include Natural Language Processing, Machine Learning and Human–Machine Interaction.



Chengcheng Li was born in 1989. He received the M.E. degree in 2015 from the School of Computer and Information at Hefei University of Technology. His research interests include Natural Language Processing, Machine Learning and Human–Machine Interaction.

Fuji Ren was born in 1959, in China. He received the B. E, M.E. from Beijing University of Posts and Telecommunications, Beijing, China, in 1982 and 1985, respectively. He received the Ph.D. Degree in 1991 from Faculty of Engineering, Hokkaido University, Japan. He worked at CSK, Japan, where he was a chief researcher of NLP from 1991. From 1994 to 2000, he was an associate professor in the Faculty of Information Sciences, Hiroshima City University. He became a professor in the faculty of engineering, the University of Tokushima in 2001. His research interests include Natural Language Processing, Artificial Intelligence, Language Understanding and Communication, and Affective Computing. He is a member of the IEICE, CAAI, IEEJ, IPSJ, JSAI, AAMT and a senior member of IEEE. He is a Fellow of The Japan Federation of Engineering Societies. He is the President of International Advanced Information Institute.