

An improved algorithm for sentiment analysis based on maximum entropy

Xin Xie¹  · Songlin Ge¹ · Fengping Hu² · Mingye Xie³ · Nan Jiang¹

© Springer-Verlag GmbH Germany 2017

Abstract Sentiment analysis is an important field of study in natural language processing. In the massive data and irregular data, sentiment classification with high accuracy is a major challenge in sentiment analysis. To address this problem, a novel maximum entropy-PLSA model is proposed. In this model, we first use the probabilistic latent semantic analysis to extract the seed emotion words from the Wikipedia and the training corpus. Then features are extracted from these seed emotion words, which are the input of the maximum entropy model for training the maximum entropy model. The test set is processed similarly into the maximum entropy model for emotional classification. Meanwhile, the training set and the test set are divided by the K -fold method. The maximum entropy classification based on probabilistic latent semantic analysis uses important emotional classification features to classify words, such as the relevance of words and parts of

speech in the context, the relevance with degree adverbs, the similarity with the benchmark emotional words and so on. The experiments prove that the classification method proposed by this paper outperforms the compared methods.

Keywords Semantic analysis · Maximum entropy · Probabilistic latent semantic analysis

1 Introduction

The narrow sense of emotional analysis (sentiment analysis) refers to analyze and mine view, emotion, attitude, etc., from text data by computer. While the generalized emotional analysis includes image, video, voice, text and other multi-modal information emotional calculation. Simply put, the goal of emotional analysis is to establish an effective analytical method, model and system to analyze the emotional information in input information, such as perspective, attitudes, subjective views or emotions.

Internet, Internet of Things and other tertiary industries become the latest driver of economic development. In these applications, emotional analysis is one of the important forces to promote its development and progress, especially in the public opinion management, business decision-making, big data analysis and other tasks. For example, in the field of Internet public opinion analysis, the emotional analysis technology can be used to get the majority views of Internet users for specific events, keep abreast of public opinion trends, and take correct action to achieve effective and orderly social management. In the field of counterterrorism, through the analysis of extreme emotions on social media, potential terrorists can be found. In the field of business decision-making, through the emotional analysis and opinion mining of massive user comments, reliable feedback information of user

Communicated by V. Loia.

✉ Xin Xie
xienew@gmail.com
Songlin Ge
masterx@126.com
Fengping Hu
hufengping1968@126.com
Mingye Xie
xiemingye@gmail.com
Nan Jiang
jiangnan@ecjtu.edu.cn

¹ School of Information Engineering, East China Jiaotong University, Nanchang, People's Republic of China

² School of Civil Engineering, East China Jiaotong University, Nanchang, People's Republic of China

³ School of Information Science Technology, East China Normal University, Shanghai, People's Republic of China

can be obtained to understand the advantages and disadvantages of the product. At the same time, a deep understanding of the real needs of users can achieve precision marketing. In addition, emotional analysis has also been successfully applied to the stock market forecast, box office forecast, election results forecast and other scenes. All of these fully reflect the emotional analysis has a huge role in all walks of life.

Previous efforts on opinion mining focused on sentiment classification on chapter and sentence levels. Pang et al. (Pang et al. 2002; Pang and Lee 2004) firstly made a series of studies about polarity classification. Three classifiers, Naive Bayes Model (NBM) (Cheeseman and Stutz 1996), Maximum Entropy Model (MaxEnt) (Berger et al. 1996), and Support Vector Machines (SVM) (Burgess 1998), were mainly used (Pang et al. 2002). Graph-based minimum-cut approach was adopted to identify the subjectivity and objectivity of sentences in Pang and Lee (2004). Ni et al. (2007) employed NBM, SVM, and Roechio's algorithm to make text sentiment classification. Information Gain and CHI were also used to select features (Ni et al. 2007).

Of course, the current emotional analysis technology is not perfect, still face many difficulties and problems. The main contributions of this paper are summarized as follows:

1. In this paper, we propose a method to calculate the similarity of words based on probabilistic latent semantic analysis. The method can solve the problem of word meaning and can calculate the semantic similarity of words more accurately than the mutual information. What is more, it has higher precision and recall rate, which is a better feature extraction method in emotion classification.
2. The maximum entropy classification based on probabilistic latent semantic analysis uses the important emotion classification features such as the relationship between words and parts in the context of words and the degree of relevance with degree adverbs, and similarity of reference emotion words. This classification method has achieved the ideal classification effect.
3. Combining the characteristics of language and being based on emotional word recognition, this paper puts forward a Sentence Recognition Method of fusing multi-feature weights such as emotional gain, degree adverbs, negative words and so on.

The structure of the rest of the paper is as follows. In section 2, we list some related work about PLSA algorithm and sentiment analysis of the application areas. In Sect. 3, we introduce PLSA and maximum entropy algorithm. And then an improved sentiment analysis algorithm is described in detail in Sect. 4. Section 5 shows modeling solving and inference. Experimental results and analysis is in Sect. 6, and a lot of experiments prove that the classification method proposed

by this paper outperforms the compared methods. Finally, in Sect. 7, we conclude the paper.

2 Related works

Probabilistic latent semantic analysis (PLSA) (Hofmann 2001) and extended work played a significant role in opinion mining and other research. Many researchers have made a lot of contributions in this field.

Wasilewski and Hurley (2016) examined a number of input aspect models and evaluated the impact that different models have on the framework. In particular, they propose a constrained PLSA model that allows for interpretable output, in terms of known aspects, while achieving greater performance than the explicit co-occurrence counting method used in previous work. MA et al. introduce a document-specific context probabilistic latent semantic analysis (DCPLSA) model for speech recognition in Haidar and O'Shaughnessy (2015). Zhang et al. (2015) proposed a multimodal multimedia retrieval model based on probabilistic latent semantic analysis (pLSA) to achieve multimodal retrieval. Extensive experiments results demonstrate the effectiveness and efficiency of the proposed model. Xu et al. (2009) propose a novel supervised dual-PLSA which estimate topics with many kinds of observable data, i.e., labeled and unlabeled documents, supervised information about topics. Experiments show the dual-PLSA has a very fast convergence. Du et al. perform a unified action recognition framework-based probabilistic latent semantic analysis (PLSA) in Du et al. (2016), which has a effective performance. To improve the precision of classification, Huang et al. proposed a new method which incorporates spatial information coming from neighbor words and topics' position into pLSA in Huang et al. (2012). Zhong and Miao (2014) present a graph regularized multimodal GM-pLSA (GRMMGM-pLSA) model to incorporate such correlation between multimodal continuous words into the process of model learning. Experiments on YouTube videos show the effectiveness of their proposed model. Wang et al. (2015a) proposes a novel vehicle color classification method which uses the concept of probabilistic latent semantic analysis (pLSA) to overcome the problem of sparse representation in data classification. Vehicle color classification is demonstrated in this paper to prove the superiority of the new classifier. Chen et al. (2007) propose a method based on probability latent semantic analysis (PLSA) to analyze web pages that are of interest to the user and the user query co-occurrence relationship, and utilize the latent factors between the two co-occurrence data for building user profile. The experimental results showed that their approach was more effective than the other typical approaches to construct user profile. Hong et al. compare these various variants of PLSA approaches with unimodal PLSAs in Hong et al.

(2015), which use either audio, visual or text features only. The experimental results show not only that one of the triple-model PLSAs achieves the highest precision, but also that social tags (text features) play an important role for classifying movies genres.

Sentiment analysis has been extensively studied since 2002 by Pang et al. (2002), especially in the emotional tendencies of online commentary. Now, it is becoming more and more popular in academia and industry. Here are the latest developments in Sentiment analysis:

Lipenkova (2015) present a pipeline for aspect-based sentiment analysis of Chinese texts in the automotive domain, which demonstrates how knowledge about sentence structure can increase the precision, insight value and granularity of the output. And the input to the pipeline is a string of Chinese characters; the output is a set of relationships between evaluations and their targets. In order to prevent the negative influence from the wrong knowledge by distinguishing highly credible knowledge, Chen et al. (2015) propose to integrate into knowledge transfer a knowledge validation model. Their experimental results demonstrate the necessity and effectiveness of the model. You et al. (2015) first design a suitable CNN architecture for image sentiment analysis. They obtain half a million training samples by using a baseline sentiment algorithm to label Flickr images. To make use of such noisy machine labeled data, they employ a progressive strategy to fine-tune the deep network. Furthermore, they improve the performance on Twitter images by inducing domain transfer with a small number of manually labeled Twitter images. They have conducted extensive experiments on manually labeled Twitter images. The results show that the proposed CNN can achieve better performance in image sentiment analysis than competing algorithms. Nguyen and Shirai (2015) use sentiments on social media to build a model to predict stock price. A new feature which captures topics and their sentiments simultaneously is introduced in the prediction model. In addition, a new topic model TS LDA is proposed to obtain this feature. The results show that incorporation of the sentiment information from social media can help to improve the stock prediction. Wang et al. (2015b) proposed a novel unsupervised sentiment analysis (USEA) framework for social media images. Their approach exploits relations among visual content and relevant contextual information to bridge the “semantic gap” in prediction of image sentiments. Zhang et al. (2016) extend related idea by proposing a sentence level neural model to address the limitation of pooling functions, which do not explicitly model tweet-level semantics. And their model has high accuracy. Wang et al. (2016) proposes a novel visual sentiment analysis approach with deep coupled adjective and noun neural networks. Specifically, to reduce the large intra-class variance, it learns a shared middle-level sentiment representation by jointly learning an adjective and a noun deep neural

network with weak label supervision. Cheng et al. (2017) study a novel problem of unsupervised sentiment analysis with signed social networks. In particular, they incorporate explicit sentiment signals in textual terms and implicit sentiment signals from signed social networks into a coherent model SignedSenti for unsupervised sentiment analysis. Empirical experiments on two real-world datasets corroborate its effectiveness. You et al. (2017) study the impact of local image regions on visual sentiment analysis. Their proposed model utilizes the recent studied attention mechanism to jointly discover the relevant local regions and build a sentiment classifier on top of these local regions.

Great progress has been made in sentiment analysis in the recent years, but the result is not so ideal when dealing with the problems “Polysemy” and “More word one meaning”. In this paper, we proposed the maximum entropy classification based on probabilistic latent semantic analysis, which is a better feature extraction method in emotion classification. What is more, it has higher precision and recall rate than other models.

3 An improved sentiment analysis algorithm

3.1 Introduction of PLSA

Probabilistic latent semantic analysis (PLSA) was proposed by Hoffman in 1999. At present, PLSA model has been successfully applied to information filtering, text classification, information retrieval and many other aspects. PLSA, developed on the basis of LSA, maps word vectors and document vectors to a low-dimensional space by the probability model. The PLSA is achieved by re-estimating LSA’s statistical maximum likelihood, with a more solid mathematical basis and easy to use data generation model, reducing the semantic ambiguity of words and documents, making the semantic relations between documents more clear.

The processing object of PLSA is a co-occurrence data pair or a binary tuple, such as two-tuple (user-object), (document-word). Just as LSA analysis, PLSA is still based on the covariance matrix of binary data. But the difference is that PLSA introduces hidden semantic class variables between binary co-occurrence data, which correspond to new semantic spaces. And probabilistic models map words and documents simultaneously to this hidden class variable in semantic space. The semantic space is the key to PLSA, which associates each pair of binary data with hidden class variables. In PLSA, the probability distribution of a word in different hidden class spaces can be described as multiple semantics of the word, and the words of the same semantic space can be understood as synonyms. So PLSA overcomes the difficulties of analyzing polysemy and synonym.

In PLSA, d is a document, and w is a word of document. Given a documents set $D = \{d_1, d_2, \dots, d_n\}$ and a set of words $W = \{w_1, w_2, \dots, w_m\}$, the text set can be expressed by co-occurrence matrix of documents and words $A = [n(d_i, w_j)]_{|D| \times |W|}$, where $n(d_i, w_j)$ is the frequency of word w_j in the document d_i . In the matrix A , each row represents a document, and each column represents a word. Every pair (d_i, w_j) is an observation. Then an unobserved hidden variable z is introduced to make the documents and words conditional independent of each other. The z belongs to the hidden variable set $Z = \{z_1, z_2, \dots, z_k\}$, where k value depends experience. If the k is too large, it easy to introduce noise. And the k is too small to classify correctly. The mapping of documents and words to hidden variables is shown in the figure.

The PLSA model further introduces the following probabilities. $p(d_i)$ is the probability that a word appears in document d_i . $p(w_j|z_k)$ is conditional probability of word w_j under the condition of hidden class variable z_k determination. $p(z_k|d_i)$ is probability distribution of document d_i in latent semantic space. Using these definitions, the word-document co-occurrence generative model can be defined as follows:

1. selecting document d_i from the document set with probability $p(d_i)$.
2. selecting hidden class variable z_k from the document d_i with probability $p(z_k|d_i)$.
3. after determining the hidden class variable z_k , generating word w_j with probability $p(w_j|z_k)$.

A pair of observations (d_i, w_j) can be got, and the hidden class variable not observed z_k is discarded. The data generation process can be expressed by the following joint probability formula.

$$\begin{aligned}
 p(d_i, w_j) &= p(d_i)p(w_j|d_i) \\
 p(w_j|d_i) &= \sum_{k=1}^K p(w_j|z_k)p(z_k|d_i) \\
 p(d_i, w_j) &= \sum_{k=1}^K p(z_k)p(d_i|z_k)p(w_j|z_k) \quad (1)
 \end{aligned}$$

This generation model can be described in Fig. 1.

3.2 The use of the maximum entropy model

The learning process of the maximum entropy model is the process of solving the maximum entropy model, which can be formalized as a constrained optimization problem.

Given a training data set $T = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$ and feature function $f_i(x, y)$, $i = 1, 2, \dots, n$, the

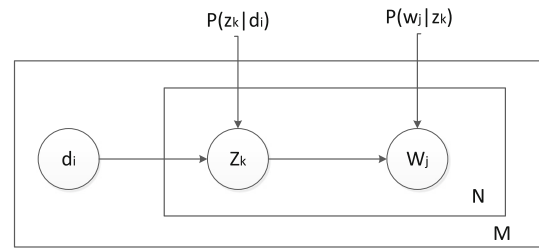


Fig. 1 PLSA graph model

learning of the maximum entropy model can be formalized as a constrained optimization problem:

$$\begin{aligned}
 \max H(P) &= - \sum_{x,y} \tilde{P}(x) P(y|x) \log P(y|x) \\
 \text{s.t } E_p(f_i) &= E_{\tilde{p}}(f_i), i = 1, 2, \dots, n \\
 \sum_y P(y|x) &= 1 \\
 Z_w(x) &= \sum_y \exp \left\{ \sum_{i=1}^n w_i f_i(x, y) \right\} \quad (2)
 \end{aligned}$$

$Z_w(x)$ is called a normalization factor; $f_i(x, y)$ is feature function; w_i is the weight of the feature. The model $P_w = P_w(y|x)$ represented by equations is the maximum entropy model. And w is parameter vector of the maximum entropy model.

4 Maximum entropy-PLSA model

Based on the analysis of previous emotional analysis techniques, this paper proposes a semantic tendency calculation method based on probabilistic latent semantic analysis (PLSA) technique and uses the maximum entropy classification algorithm to classify the emotion words.

At present, many emotional analysis studies take the importance of feature to the emotional classification into account to select feature, but they ignore the semantic relations between lexical features and the influence of context on word meaning. While PLSA technology can map text to low-dimensional latent concept semantic spaces to obtain a reinterpretation of the text, and the representation of the text in this space can better reflect the semantic similarity between texts. The maximum entropy classification method can flexibly define the eigenfunctions and can better estimate the probability distribution of unknown words in the context of known context constraints. Taking into account the semantic relations among words, the context of words and the importance of various types of words on the classification of emotional words, this paper uses probabilistic latent semantic analysis (PLSA) to reduce the high dimension of the

text features, construct a mixed multi-feature features function and use the maximum entropy classification algorithm for the classification of emotional words.

The maximum entropy classification algorithm consists of the following two parts:

- (1) Generating the model's parameter file. It includes feature extraction and training parameters. Feature extraction is based on to the selected feature template, generating a file for training parameters. It mainly uses the algorithm to calculate the emotional word semantic features. Training parameters is based on the selected feature template, generating values of parameters and storing in the file.
- (2) Discriminating sentiment word. It is the word segmentation and part-of-speech (POS) tagging to corpus, filtering out the candidate emotional words. For each candidate emotional word, first look for it in emotional dictionary. If it presents, mark it. Otherwise, we calculate the probability that a word belongs to a certain kind of emotional tendency according to the selected feature template, the value of parameter, and specific context. calculate the probability of a certain kind of emotional tendency. And then we select the class with largest probability, marking the corresponding emotional word tendencies, with its probability as the emotional confidence. These results of various types are sorted in descending order by emotional confidence. The words with big emotional confidence selected as emotional words. The new emotional words by manual verification are added into the existing emotional dictionary, which is the emotional dictionary to identify the following emotional sentence. Confidence is the degree of confidence to make judgments. In this paper, results of various types are sorted in descending order by emotional confidence, which facilitates the analysis of the experimental results and the subsequent recognition of emotional sentences.

Figure 2 shows that we first use the probabilistic latent semantic analysis (PLSA) to extract the seed emotion words from the Wikipedia and the training corpus. Then Features are extracted from these seed emotion words, which are the input of the maximum entropy model for training the maximum entropy model. The test set is processed similarly into the maximum entropy model for emotional classification. Meanwhile, the training set and the test set are divided by the K-fold method. The maximum entropy classification based on probabilistic latent semantic analysis uses important emotional classification features to classify words, such as the relevance of words and parts of speech in the context, the relevance with degree adverbs, the similarity with the benchmark emotional words and so on.

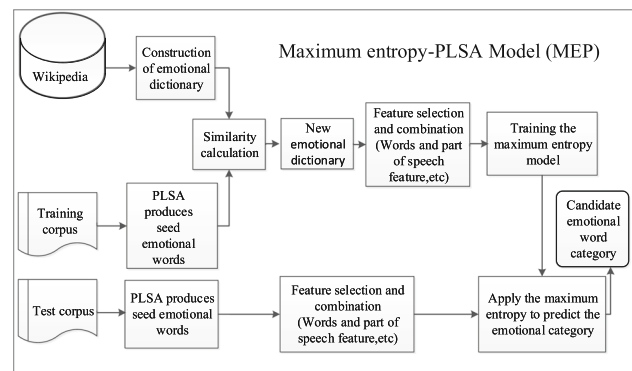


Fig. 2 MEP overall flow chart

4.1 Construction of emotional dictionary

Emotional seed is the words with absolute emotional meaning. These words are either the author's own emotions, or the inner mood of the characters in the text. For mining of these emotional seed words, this paper uses Wikipedia's dictionary resources. Wikipedia is a common knowledge base, which is based on the concept of a word as a descriptive object to reveal the relationship between concepts and attributes possessed by concepts. Emotional analysis of the phrase set contains the evaluation of words and emotional words and is divided into two categories. In this paper, we choose the words with absolute meaning as the emotional seeds of emotion analysis. We think these seeds can show the emotional tendency of the characters in almost all the contexts, so the confidence degree of these words is 1.

In addition, the emotional vocabulary is regarded as the basic emotion dictionary of sentiment analysis, and divided commendatory and derogatory tables. The confidences of all emotional seed are 1; the rest words' confidences are set 0.8. However, due to the limit of emotional words' number in the dictionary, it is necessary to further expand the emotional dictionary. In a large number of corpus, we find that the emotional words with the same tendency often appear at the same time, and the emotional words with the different tendency do not appear together. According to these two phenomena, this paper proposes an algorithm based on probabilistic latent semantic analysis and maximum entropy model. We obtain the words with high similarity to the seed words from the large corpus, and use the maximum entropy classification method to judge the affective tendency. In order to find more emotional words and phrases, this paper adds emotional words distinguished by classifier into the emotional dictionary. And after artificial check, we add the emotional words with high confidence into the emotional seed to search for emotional words iteratively.

Degree adverb

Based on the degree words in the emotional analysis vocabulary and the results of part-of-speech (POS) tagging to corpus,

this paper collates the commonly used degree adverb, such as “very”, “very”, “then”, “rather” and so on. In this paper, the degree adverbs are divided into three levels. The first level of degree adverbs weaken the emotional strength of the modified words, such as “a little”, “slightly”. The second level of adverbs enhances the modified emotional intensity, such as “much”. The third degree of adverbs greatly enhances the emotional intensity of the modified words, such as “very”. Three levels of adverbs enhance the emotional intensity of the modified word multiply.

Expressive verbs

The emotional words are often accompanied by verbs like “experience” or “expression”. So this article collects these words together based on emotional analysis of the words in the vocabulary and the results of part-of-speech (POS) tagging to corpus, such as “feel”, “show”, “feel” and so on.

Interjection

In the past information classification and clustering method, the interjection is removed as stop words. But people often have a lot of modal words when post comments or express emotions, such as “ah”, “what”, “friends”, “yeah”. These words can help people express their feelings tendencies. Thus, this vocabulary can be a distinguishing feature of subjective and objective texts. By way of artificial annotation during the training corpus, this paper collects the interjection for emotional analysis of sentences.

Conjunctions

The complex sentence is composed of two or more clauses, which are grammatical units that are structurally similar to single sentences without complete sentences. The clauses of a complex sentence can be a parallel relationship, a turning relationship, a progressive relationship, and so on. These relationships have a very important role in the judgment of the sentimental sentence. So by way of artificial annotation during the training corpus, this paper collects the associated vocabulary of the complex sentence. There are three types of conjunctions:

- (i) Progressive words, such as “and”, “not to mention”, “even” and so on;
- (ii) Tied words, such as “also”, “and”, “and” and so on;
- (iii) Turning words, such as “but”, “but”, “but”, “but” and so on.

Negative vocabulary

In determining the initial emotional score of a sentence, the negation of the sentence can lead to the elimination or even change of the emotion. Therefore, this paper collects a negative vocabulary from the results of the training notes. If there are negative words in a sentence, the sentence needs to do appropriate adjustment in a certain emotional tendencies. Common negative words are “no”, “not”, “never” and so on.

4.2 Similarity calculation

The coordinates of the words ω_i and ω_j in the k -dimensional space are v_i and v_j , respectively, where $v_i = [x_{i,1}, \dots, x_{i,n}]$ $v_j = [x_{j,1}, \dots, x_{j,n}]$ The semantic similarity of the two words ω_i and ω_j can be calculated by angle cosine of the coordinate vector v_i and v_j in the k -dimensional space:

$$\text{Similar}(\omega_i, \omega_j) = \cos(v_i, v_j) = \frac{\sum_{k=1}^n x_{ik}x_{jk}}{\sqrt{\sum_{k=1}^n x_{ik}^2} \sqrt{\sum_{k=1}^n x_{jk}^2}} \quad (3)$$

Through the formula (3), we can get the semantic similarity between the candidate emotion words and the seed set of emotional words.

4.3 Feature selection and combination

One of the advantages of the maximum entropy model is that it can flexibly select the feature set for a specific task. In the recognition of emotional words, this paper defines the feature space as follows, by considering some factors affecting judgments, such as context information of the candidate affective words, associated words and degree of statistical information.

- (1) The emotional characteristics of words in the candidate affective words.
- (2) The context and feature of the affective words.
- (3) Statistics the correlation degree of the candidate affective words, degree adverbs and the expressive verbs. Statistical similarity between the candidate affective words and known emotional words.

These features introduce knowledge of transcendental areas, whose main reference is the previous summary of the emotional vocabulary, emotional verbs and the degree of adverb. the main reference is the previous summary of emotional words, expressive verbs and degree adverb.

According to the above-defined feature space, the feature template in the model is shown in Table 1.

The following is a detailed description of some feature template calculations:

- (1) Words and part of speech features

w_0 refers to the candidate emotional word, and pos_0 is w_0 's part of speech. w_{-1} , w_{-2} are the first two words of the candidate emotional word in the sentence, and pos_{-1} and pos_{-2} are their part of speech, respectively. w_{+1} , w_{+2} are the last two words of the candidate emotional word in the sentence, and pos_{+1} and pos_{+2} are their part of speech, respectively.

- (2) Word feature similarity

Table 1 Atomic features template

Atomic template	Feature description
$w_0; w_{+1}; w_{+2}; w_{-1}; w_{-2}$	Words
$pos_0; pos_{+1}; pos_{+2}; pos_{-1}; pos_{-2}$	Part of speech
S_{char}	Sentiment orientation based on word features
C_{adv}	The relevance of candidate emotional words and degree adverbs
C_v	The relevance of candidate emotional words and expressive adverbs
S_w	The relevance of candidate emotional words and benchmark emotional words
$S_{w-conver}$	The relevance of candidate emotional words after the conversion and benchmark emotional words

Assuming that the occurrence frequency of a character C in the i th words of the given positive and negative emotional words set V_{pos} , V_{neg} are m_{pos-i} , m_{neg-i} , then the membership of the character C to the positive emotional words set is expressed as:

$$p(c \in V_{pos}) = \frac{\sum_i m_{pos-i} * \lambda_i}{\sum_i m_{pos-i} * \lambda_i + \sum_i m_{neg-i} * \lambda_i} * \frac{\sum_i m_{pos-i}}{|V_{pos}|} \quad (4)$$

where the weight parameter $\lambda_i = \frac{S_i}{Len_i}$. S_i is the emotional confidence of the i th word in the emotional set, and Len_i is the length of the i th word. Similarly, we can get C ' membership belonging to another set.

If a candidate emotion word w is composed of the characters $c_1 - c_n$, then the membership of candidate word w on the word attribute of the set V_{pos} can be expressed as:

$$P(w \in V_{pos}) = \sum_{i=1}^n P(c_i \in V_{pos}) \quad (5)$$

Similarly, we can get the candidate emotion word' membership belonging to another set.

(3) Correlation with degree adverb C_{adv}

If a word and a particular word appear together many times in a given size window, then the word is considered to have a strong semantic correlation with the particular word. The window size is the number of words between two words, and the window size is 6 in this paper. The co-occurrence of word w_i and word w_j can be calculated by formula. The greater the value, the greater the correlation.

$$cooccur(w_i, w_j) = \log_2 \left(\frac{p(w_i \& w_j)}{p(w_i) * p(w_j)} \right) \quad (6)$$

where $p(w_i \& w_j)$ is the co-occurrence probability of two words in the window size. $p(w_i)$ and $p(w_j)$ are the probabilities of the words w_i and w_j in the document, respectively.

The correlation C_{adv} between the candidate emotional word w and the degree adverb is the maximum of the co-

occurrence of the word w and degree adverb adv_i in degree adverb table, which is

$$C_{adv} = \max_{i=1, \dots, n} cooccur(adv_i, w) \quad (7)$$

(4) Correlation with expressive verbs C_v

The correlation between the word w and the word v_j in expressive verbs table is expressed by their co-occurrence $Connect(v_j, w)$. And the correlation of the candidate emotion word w and the expressive verbs C_v is the maximum of $Connect(v_j, w)$.

$$C_v = \max_{j=1, \dots, m} Connect(v_j, w) = \max_{j=1, \dots, m} cooccur(v_j, w) \quad (8)$$

(5) Similarity features with baseline emotional words set S_w

The semantic similarities $Similar(pos_i, w)$, $Similar(neg_j, w)$ of the candidate emotion word and the sets of positive and negative baseline emotional words are calculated by the algorithm LSA. Then the similarity S_w between the word w and baseline emotional words is calculated according to the following formula.

$$S_w = \frac{\sum_{i=0}^k Similar(pos_i, w)}{k} - \frac{\sum_{j=0}^p Similar(neg_j, w)}{p} \quad (9)$$

(6) Similar features with the convert form of the known emotional words $S_{w-conver}$

The training sample (x, y) is instantiated for the above templates to obtain specific feature functions. The training sample set instantiates all the templates to get the feature set. In this paper, the maximum entropy model is used to fuse the feature functions defined by the above feature templates, and the weight of each feature function is obtained through a large number of corpus training stations.

The context information of the text will affect the output of the random process. For example, in the emotional word

recognition, the tendency of the unknown candidate emotional word in the text may be related to its context. In this paper, we use the feature function to represent the predictive effect of the context information, the semantic relations between words, etc., on the tendency of unknown emotional words. The feature function is usually expressed as a binary function, such as:

$$f(x, y) = \begin{cases} 1, & \text{sign} = x \text{ and } \text{context} = y \\ 0, & \text{other} \end{cases} \quad (10)$$

Given a word and its context c_i , we want to get the maximum entropy model satisfying all constraints. That is:

$$T^* = \arg \max P(t_i | c_i) \quad (11)$$

Where t_i is emotional tendency of w , and $P(t_i | c_i)$ is the entropy of the word w_i marked as t_i under the condition of the context c_i .

When the unknown probability distribution is equal, the entropy is the largest. And the distribution is the required distribution.

The context conditions are transformed into a set of feature functions, and the weight of the feature function f_i is represented by λ_i . In this paper, we estimate a probability model by calculating the linear combination of feature functions $\sum_i \lambda_i f_i(t, c)$. And the maximum entropy solution can be expressed as follows:

$$H(t | c, \lambda) = \frac{\exp \sum_i \lambda_i f_i(t, c)}{\sum_{t'} \exp \sum_i \lambda_i f_i(t', c)} \quad (12)$$

5 Model solving and inference

5.1 PLSA solution

From the above, we can see that the conditional probability distributions $p(z_k | d_i)$ and $p(w_j | z_k)$ should be estimated to train generative model from data set. Therefore, we need to obtain the likelihood function.

$$L = \sum_{i=1}^M \sum_{j=1}^N n(d_i, w_j) \log p(d_i, w_j) \quad (13)$$

In order to find the conditional probability distribution in the latent semantic model, a common method is to estimate the parameters of the maximum likelihood function by the expectation maximization algorithm (EM algorithm). EM algorithm is an iterative process that consists of two steps.

E step: according to the estimation of the current parameters, the posterior probability of the hidden variables is calculated.

$$P(z_k | d_i, w_j) = \frac{\hat{P}(z_k) \hat{P}(z_k | d_i) \hat{P}(w_j | z_k)}{\sum_{l=1}^K \hat{P}(z_l) \hat{P}(z_l | d_i) \hat{P}(w_j | z_l)} \quad (14)$$

Where \hat{p} represents the parameter value estimated in last iteration of the EM algorithm, which is initialized to a random value.

M step: The probability distributions $p(d_i | z_k)$, $p(w_j | z_k)$ and $p(z_k)$ are updated using the posterior probability calculated in E step.

$$\begin{aligned} P(w_j | z_k) &= \frac{\sum_{i=1}^M n(d_i, w_j) P(z_k | d_i, w_j)}{\sum_{i=1}^M \sum_{l=1}^N n(d_i, w_l) P(z_k | d_i, w_l)} \\ P(d_i | z_k) &= \frac{\sum_{j=1}^N n(d_i, w_j) P(z_k | d_i, w_j)}{\sum_{l=1}^M \sum_{j=1}^N n(d_l, w_j) P(z_k | d_l, w_j)} \\ P(z_k) &= \frac{\sum_{i=1}^M \sum_{j=1}^N n(d_i, w_j) P(z_k | d_i, w_j)}{\sum_{i=1}^M \sum_{j=1}^N n(d_i, w_j)} \end{aligned} \quad (15)$$

Both of E step and M step are carried out interactively and iteratively until the likelihood function reaches the maximum or the magnitude of the likelihood function changes within a certain threshold so as to avoid training overfitting. Finally, the matrices $P(d | z)$ and $P(w | z)$ are obtained, and the matrix $P(z | d)$ can be obtained by Bayes formula.

Iteration is the process of solving the problem in a numerical analysis by finding a series of approximate solutions from an initial estimate. The method used to implement this process is collectively referred to as iterative method. The EM algorithm belongs to the iterative method, and its algorithm complexity is mainly related to the size of the problem, which can be represented by $O(2 * T)$, and T is the number of iterations.

5.2 Maximum entropy solution

The maximum entropy model is based on the maximum entropy principle to evaluate parameters for each feature. And each parameter corresponds to a feature in order to establish the required model.

In Algorithm 5.1 f_1, f_2, \dots, f_n are called Feature functions. $\tilde{p}(x, y)$ is empirical distribution, and w_i is the weight of the feature. The model $P_w = P_w(y | x)$ represented by equations is the maximum entropy model. And w is parameter vector of the maximum entropy model. We get the maximum entropy optimization model $p_w(y | x)$ through algorithm Improved Iterative Scaling (IIS). IIS algorithm also belongs to iterative method. Its algorithm complexity is mainly related to the size of the problem, which can be represented by $O(n * T)$, n is the number of parameters, and T is the number of iterations.

Algorithm 5.1 Improved Iterative Scaling**Require:**

Feature functions f_1, f_2, \dots, f_n ;
 empirical distribution $\tilde{p}(x, y)$;
 model $p_\omega(y|x)$.

Ensure:

Optimal parameter values ω_i^* ;
 optimal model p_{ω^*} .
 // begin procedure:
 1. Start with $\omega_i = 0$ for all $i \in \{1, 2, \dots, n\}$
 2. Do for each $i \in \{1, 2, \dots, n\}$:
 a. Let δ_i be the solution to

$$\sum_{x,y} \tilde{p}(x)p(y|x)f_i(x,y)\exp(\delta_i f_i^\#(x,y)) = E_{\tilde{p}}(f_i)$$

 where

$$f_i^\#(x,y) \equiv \sum_{j=1}^n f_j(x,y)$$

 b. Update the value of δ_i according to : $\omega_i \leftarrow \omega_i + \delta_i$ 3. Go to step2 if not all the δ_i have converged
 The key step in the algorithm (2a), the computation of increments δ_i that solve (46).
 If $f_i^\#(x,y)$ is constant($f_i^\#(x,y) = M$ for all x,y ,say), then δ_i is given explicitly as

$$\delta_i = \frac{1}{M} \log \frac{E_{\tilde{p}}(f_i)}{E_p(f_i)}$$

If $f_i^\#(x,y)$ is not constant, then δ_i must be computed numerically. A simple and effective way of doing this is by Newton's method. This method computes the solution δ_i^* of an equation p3 iteratively by the recurrence
 // Newton iterative method to get parameters

$$\delta_i^{(k+1)} = \delta_i^{(k)} - \frac{g(\delta_i^{(k)})}{g'(\delta_i^{(k)})}$$

with an appropriate choice for $\delta_i^{(0)}$ and suitable attention paid to the domain of g .
return $p_\omega(y|x)$.

6 Experimental results and analysis

In this paper, we use two corpus to verify the validity of the MEP model. At first, we use the same data set as Brody did in Brody and Elhadad (2013), which originates from Ganu et al. (2009). This data set consists of restaurant review corpus. Similar to their methods, we manually annotate 100 sentences for training the MAE model. When pre-processing the data, we remove stop words and use Stanford POS Tagger (Toutanova 2004) to tag the data set. We can get the corpus from this internet site (Brody and Elhadad 2009). Besides, we use another data set which provided by Cornell University. This data set consists of film reviews, of which 1000 are positive and negative attitudes. In addition, there are 5331 sentences marked with negative emotions, marked with the subjective and subjective sentences of 5000 sentences. The current film critics are widely used in a variety of particle size, such as words, sentences and chapter-level emotional analysis. We can get the corpus from this internet site (Pang and Lee 2002). The maximum entropy model and the PLSA model are implemented using the associated maximum entropy toolkit (Zhang 2015) and the PLSA maximum entropy toolkit (JyFantas 2014).

In this paper, we use the precision and recall rate to evaluate the performance of emotional word recognition and

classification experiments. For a result of the emotional orientation marker sequence, the precision rate refers to the correct proportion of the results of the given marker sequence, and the recall rate refers to the ratio of the result of the given marker sequence to the actual correct sequence of the mark sequence. These two indicators are derived from information retrieval, which are more commonly used indicators in natural language processing tasks. The accuracy indicates the correctness of the emotional analysis model, and the recall rate indicates the integrity of the emotional analysis model.

Given a test set of text sequences $T = (t_1, t_2, t_3, \dots, t_n)$, through the model after analysis by the emotional analysis model, an emotional orientation marker sequence $S = (s_1_{c1}, s_2_{c2}, \dots, s_m_{cm})$ is obtained. In the case of word-level emotion analysis, s_i represents the emotional word identified in the test set. In the sentence level of emotional analysis, s_i represents the emotional sentence identified in the test set. c_i is the emotional tendencies of s_i , including positive, negative and neutral three categories. N_r is the number of words that belong to class C and are marked as class C in emotional orientation marker sequence. N_w is the number of words that not belong to class C but are marked as class C . N_l is the number of words that belong to class C but are not marked as class C . The recall rate and precision rate are defined as follows:

$$R = \frac{N_r}{N_r + N_l}$$

$$P = \frac{N_r}{N_r + N_w}$$

There is a slight negative correlation between accuracy and recall. In order to evaluate the system well, people define the harmonic mean of these two indicators as a comprehensive evaluation index F -measure. F -measure can correctly reflect the effect of the emotional analysis model in the balance of precision and recall. its formula is as follows.

$$F\text{-measure} = \frac{2 * P * R}{P + R} \quad (16)$$

6.1 Experiments on restaurant review corpus

Table 2 shows the sample result of food of MEP model. For different aspects of food, we enumerate its positive sentiment or negative sentiment classified by our model MEP. For example, the word “missed” is positive sentiment for “cake”. The word “delicious” expresses the praise of “food”. While, about “rice”, the “cooked” is negative sentiment. The word “horrible” expresses a strong dislike for “sushi”. The positive words classified by MEP account for 81.2% of the total, while the negative words only account for 18.8%.

Tables 3 and 4 show the result of experiment. Multinomial Naive Bayes (NB), Maximum Entropy (MaxEnt) and

Table 2 The sample result of food of MEP model

Positive (81.2%)		Negative (18.8%)	
Aspect	Sentiment	Aspect	Sentiment
Cake	Missed	Rice	Cooked
Food	Delicious	Sushi	Horrible
Dessert	Tasty	Chicken	Raw
Dish	Authentic	Chow	Ordinary
Chocolate	Fresh	Pancakes	Disgusting
Sauce	Decent	Coffee	bad
Wine	Tasty	Burger	Oil
Bread	Satisfying	Turkey	Bland

Table 3 The average precision, recall and *F*-measure of different size emotional word

Algorithms	Precision (%)	Recall (%)	<i>F</i> -measure (%)
LSTM	84.28	88.47	84.28
DCNN	84.32	88.54	86.38
SAE	83.88	88.88	86.31
NB	83.03	87.30	85.11
SVM	83.69	88.78	86.16
MaxEnt	84.14	89.65	86.81
MEP	85.21	90.34	87.70

Table 4 The average precision, recall and *F*-measure of different size test corpus

Algorithms	Precision (%)	Recall (%)	<i>F</i> -measure (%)
LSTM	86.08	90.15	88.07
DCNN	86.12	89.97	86.03
SAE	85.08	90.27	87.59
NB	84.63	89.24	86.87
SVM	86.29	89.74	87.98
MaxEnt	86.34	90.11	88.18
MEP	87.11	91.42	89.21

Support Vector Machine (SVM) are used as baseline methods. We also experiment on another deep learning technique called stacked auto-encoders (SAE) (Gehring et al. 2013) with TF-IDF representation.

From Table 3, we find that for the emotional words, both LSTM and DCNN have higher accuracies compared to other classifiers. The precision of these methods is 84.28 and 84.32, respectively. And the MEP proposed in this paper outperforms all baselines. The precision, Recall and *F*-measure of MEP are 85.21, 90.34 and 87.70, respectively.

Table 4 shows the precision of the test corpus. From the table, it is found that MaxEnt has higher precision than other baselines, and its precision is 86.34. The SVM, DCNN and LSTM have general performance. What is more,

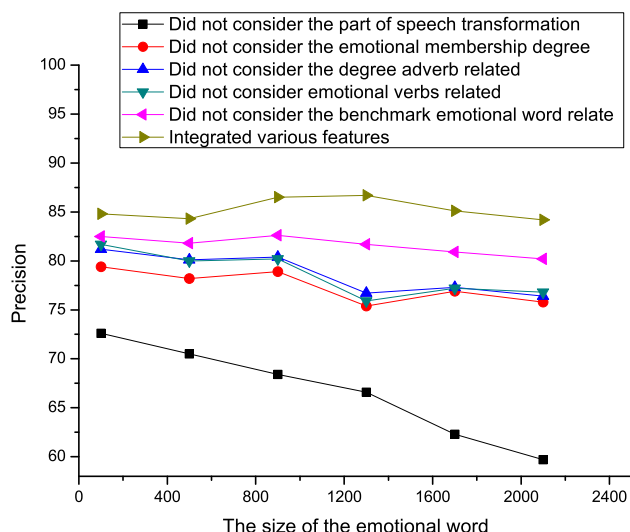
Table 5 Classification results based on different similarity calculation methods

	Precision (%)	Recall (%)	<i>F</i> -measure(%)
LSA	82.51	85.71	84.08
MI	80.63	81.34	80.98
PLSA	85.6	90.5	87.9

the precision, Recall and *F*-measure of MEP are 87.11, 91.42 and 89.21, respectively, which are highest in these methods.

6.2 Experiments on film review corpus

In this paper, the maximum entropy classification is based on the feature function of probabilistic latent semantic analysis. Therefore, we compare three kinds of semantic similarity calculation methods, latent semantic analysis, probabilistic latent semantic analysis and mutual information (MI). The experimental results are shown in Table 5. We can see that the effect of PLSA is quite good. It can be seen from the experimental results that the semantic similarity method can be better than the calculation method of the mutual information in the precision and recall rate of the classification by probabilistic latent semantic analysis. The feature function is an important part of the maximum entropy classification. The good feature function can improve the efficiency of the classifier and obtain better classification results. The effect of different eigenfunctions on the maximum entropy classification is different. Therefore, this paper compares the effect of different feature function combinations on the classification of maximal entropy.

**Fig. 3** The precision of different size of the emotional word

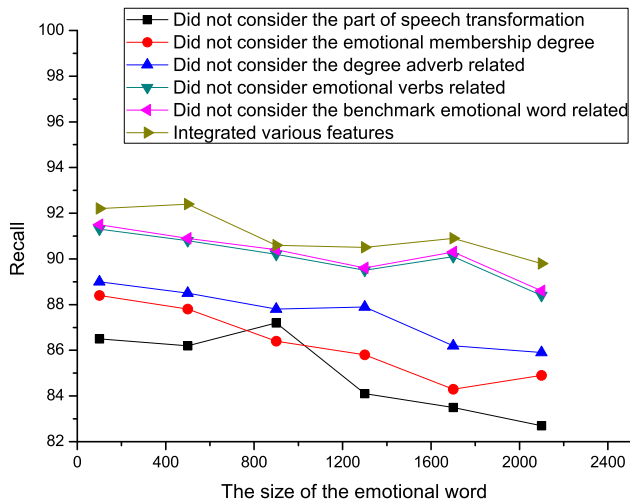


Fig. 4 The recall of different size of the emotional word

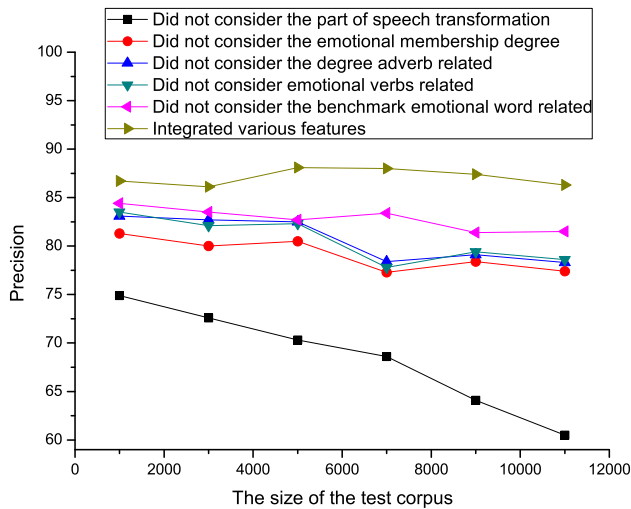


Fig. 5 The precision of different size of the test corpus

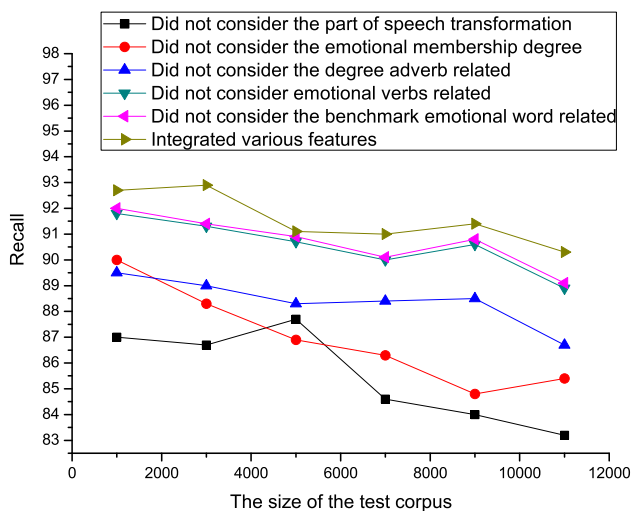


Fig. 6 The recall of different size of the test corpus

The x -axis of Fig. 3 represents the number of emotional words from the results in the descending order by emotional word confidence. From Fig. 3, it can be seen that the similarity feature function has a significant effect on improving the recall rate of emotion word recognition. It can be seen from Fig. 3 that the precision of the various features is higher than that of a particular feature.

The abscissa of Fig. 4 represents the number of emotional words from the results in the descending order by emotional word confidence. From Fig. 4, it can be seen that the similarity feature function has a significant effect on improving the precision of emotion word recognition, which is likely that the correlation between degree adverbs and emotional verbs plays a major role in improving precision. The similar feature based on the convert form has less effective on improving the precision of the recognition of emotional words.

In Fig. 5, we find that the similarity of the word-based transformation is obvious to improve the recall rate of the emotional word recognition. The relevance features of the candidate word and the degree adverb and the emotional verb also partially enhance the recall rate. The similarity feature function between the candidate words and the benchmark emotional words based on reduces the recall rate of some emotion words. The similar features based on the convert form have little effect on improving the precision of emotional word recognition.

From Fig. 6, we find that the recall rate is declining. The probability of unlisted words gradually increased; the probability of misclassification of the classifier also increased. The main reason is that as the number of sentences in the test corpus increases, the probability of unlisted words gradually increased, and the misclassification probability of the classifier is increased.

Based on the above results, the judgment of the emotional sentence based on the recognition of the emotional word has a certain precision Rate and recall rate, and combined with the characteristics of the maximum entropy model in the field of emotional analysis has a very good effect.

7 Conclusion and prospect

In this paper, the classification theory is applied to the identification of two kinds of emotional words. By fusing multi-features, we put forward a method to recognition and classification of emotion sentences. Combining with the characteristics of the candidate words, context, the coexistence of adverbs and other characteristics, we construct a characteristic function and train the maximum entropy model to identify the new emotion words in the corpus. In this paper, we propose a Maximum entropy-PLSA Model. In this model, we use the probabilistic latent semantic analysis (PLSA) to

extract the seed emotion words from the Wikipedia and the training corpus. Then Features are extracted from these seed emotion words, which are the input of the maximum entropy model for training the maximum entropy model. The test set is processed similarly into the maximum entropy model for emotional classification. Meanwhile, the training set and the test set is divided by the K -fold method. The maximum entropy classification based on probabilistic latent semantic analysis uses important emotional classification features to classify words, such as the relevance of words and parts of speech in the context, the relevance with degree adverbs, the similarity with the benchmark emotional words and so on. The experiments prove that the classification method proposed by this paper has an ideal classification effect.

Acknowledgements This work is supported by the National Natural Science Foundation, under Grant Nos. 61762037, 61640217, 41402290, 61462028, Science and Technology Support Program of Jiangxi Province, under Grant No. 20151BBE50055, and Science and Technology Project supported by education department of Jiangxi Province under Grant No. GJJ150541, and Nanchang City Sensor Network and Compressed Sensing Knowledge Innovation Team under Grant No. 2016T75.

Compliance with ethical standards

Conflict of interest The authors declare that there is no conflict of interest regarding the publication of this paper.

References

- Berger AL, Pietra VJD, Pietra SAD (1996) A maximum entropy approach to natural language processing. *Comput Linguist* 22(1):39–71
- Brody S, Elhadad N (2009) Restaurant review corpus. <http://people.dbmi.columbia.edu/noemie/ursa>
- Brody S, Elhadad N (2013) An unsupervised aspect-sentiment model for online reviews. In: Human language technologies: conference of the North American chapter of the Association of Computational Linguistics, Proceedings, June 2–4, 2010. Los Angeles, California, USA, pp 804–812
- Burges CJC (1998) A tutorial on support vector machines for pattern recognition. *Data Min Knowl Discov* 2(2):121–167
- Cheeseman P, Stutz J (1996) Bayesian classification (autoclass): theory and results. Fayyad U.m.etc. *advances in Knowledge Discovery & Data Mining Aaai*, pp 153–180
- Chen Q, Wenjie Li Y, Lei XL, He Y (2015) Learning to adapt credible knowledge in cross-lingual sentiment analysis. *ACL* 1:419–429
- Cheng K, Li J, Tang J, Liu H (2017) Unsupervised sentiment analysis with signed social networks. In: *AAAI*, pp 3429–3435
- Chen D, Wang D, Yu G, Yu F (2007) A PLSA-based approach for building user profile and implementing personalized recommendation. In: *Advances in data and web management*. Springer, pp 606–613
- Du K, Shi Y, Lei B, Chen J, Sun M (2016) A method of human action recognition based on spatio-temporal interest points and PLSA. In: 2016 international conference on industrial informatics-computing technology, intelligent technology, industrial information integration (ICIICIT). IEEE, pp 69–72
- Ganu G, Elhadad N, Marian A (2009) Beyond the stars: improving rating predictions using review text content. In: International workshop on the web and databases, WEBDB (2009) Providence. Rhode Island, USA, June
- Gehring J, Miao Y, Metze F, Waibel A (2013) Extracting deep bottleneck features using stacked auto-encoders. In: 2013 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 3377–3381
- Haidar MA, O'Shaughnessy D (2015) Document-specific context PLSA language model for speech recognition. In: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 5326–5330
- Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. *Mach Learn* 42(1):177–196
- Hong HZ, Hwang JI (2015) Multimodal PLSA for movie genre classification. In: International workshop on multiple classifier systems. Springer, pp 159–167
- Huang F, Jing X, Sun S, Lu Y (2012) Incorporate spatial information into PLSA for scene classification. In: International conference on trustworthy computing and services. Springer, pp 170–177
- JyFantas (2014) PLSA. <https://github.com/JFantasy/plsa>
- Lipenkova J (2015) A system for fine-grained aspect-based sentiment analysis of Chinese. In: *ACL (system demonstrations)*, pp 55–60
- Nguyen TH, Shirai K, Velcin J (2015) Modeling based sentiment analysis on social media for stock market prediction. In: The meeting of the association for computational linguistics and the international joint conference on natural language processing of the Asian Federation of natural language processing
- Ni X, Xue GR, Ling X, Yu Y, Yang Q (2007) Exploring in the weblog space by detecting informative and affective articles. In: *Proceedings of the 16th international conference on World Wide Web*. ACM, pp 281–290
- Pang B, Lee L (2002) Movie review data. <http://www.cs.cornell.edu/people/pabo/movie-review-data>
- Pang B, Lee L (2004) A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*. Association for Computational Linguistics, p 271
- Pang B, Lee L, Vaithyanathan S (2002) 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, pp 79–86
- Toutanova K (2004) Stanford log-linear part-of-speech tagger. <http://nlp.stanford.edu/software/tagger.shtml>
- Wang SY, Hsieh JW, Yan Y, Chen LC, Chen DY (2015a) PLSA-based sparse representation for vehicle color classification. In: 2015 12th IEEE international conference on advanced video and signal based surveillance (AVSS). IEEE, pp 1–6
- Wang Y, Wang S, Tang J, Liu H, Li B (2015b) Supervised sentiment analysis for social media images. In: *IJCAI*, pp 2378–2379
- Wang J, Fu J, Xu Y, Mei T (2016) Beyond object recognition: visual sentiment analysis with deep coupled adjective and noun neural networks. In: *IJCAI*, pp 3484–3490
- Wasilewski J, Hurley N (2016) Intent-aware diversification using a constrained PLSA. In: *ACM conference on recommender systems*, pp 39–42
- Xu WR, Liu DX, Guo J, Cai YC et al (2009) Supervised dual-PLSA for personalized SMS filtering. In: *Asia information retrieval symposium*. Springer, pp 254–264
- You Q, Jin H, Luo J (2017) Visual sentiment analysis by attending on local image regions. In: *AAAI*, pp 231–237
- You Q, Luo J, Jin H, Yang J (2015) Robust image sentiment analysis using progressively trained and domain transferred deep networks. *arXiv preprint arXiv:1509.06041*
- Zhang L (2015) A maximum entropy modeling toolkit for python and C++. <https://github.com/lzhang10/maxent>

- Zhang Y, Yuan Y, Guoren W (2015) A multimodal multimedia retrieval model based on PLSA. In: Web information system and application conference, pp 33–36
- Zhang M, Zhang Y, Vo DT (2016) Gated neural networks for targeted sentiment analysis. In: AAAI, pp 3087–3093
- Zhong C, Miao Z (2014) Modeling correlation between multi-modal continuous words for PLSA-based video classification. In: 2014 IEEE international conference on image processing (ICIP). IEEE, pp 4304–4308