**An improved algorithm for sentiment analysis based on maximum entropy**

**1 Introduction**

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.

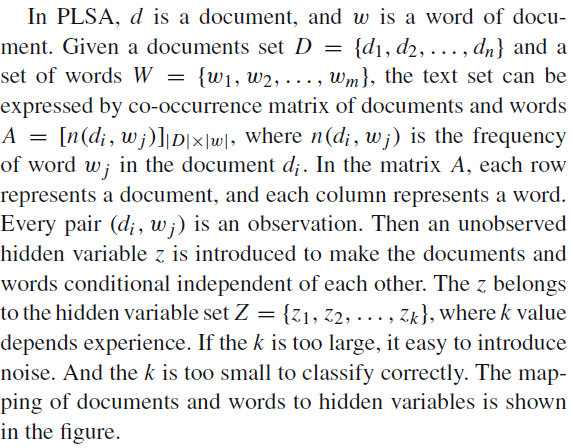
3. Combining the characteristics of language and being based on emotional word recognition, this paper puts forward a Sentence Recognition Method of fusing multifeatured weights such as emotional words, degree adverbs, negative words and so on.

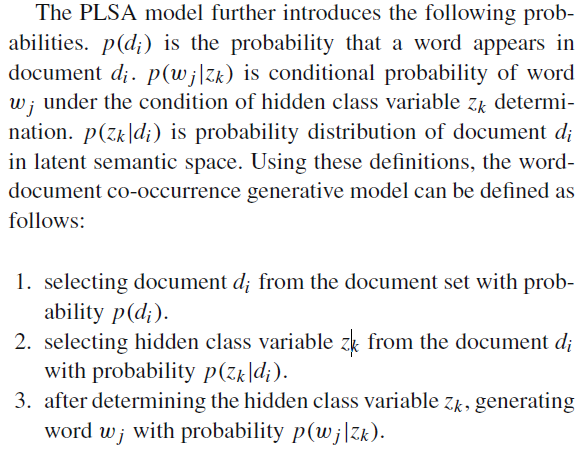
**2 Related works**

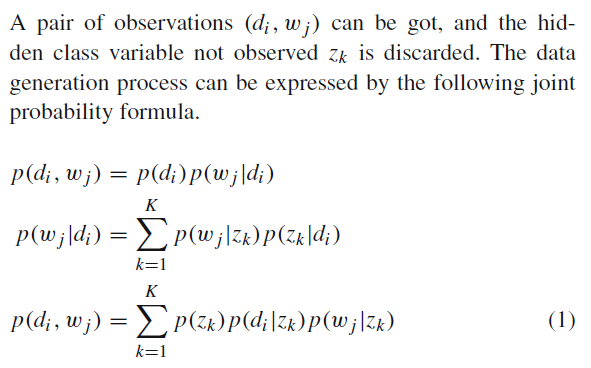
In this paper, we proposed the maximum entropy classification based on probabilistic latent semantic analysis, which is a better feature extraction method in emotion classification. What is more, it has higher precision and recall rate than other models.

**3 An improved sentiment analysis algorithm**

**3.1 Introduction of PLSA**

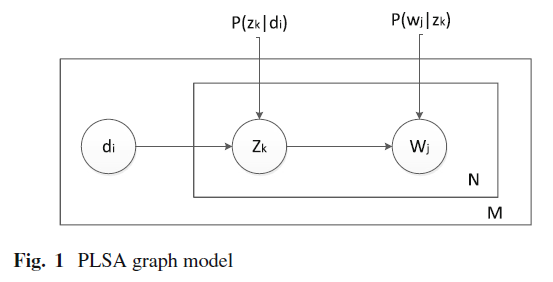


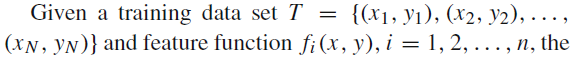


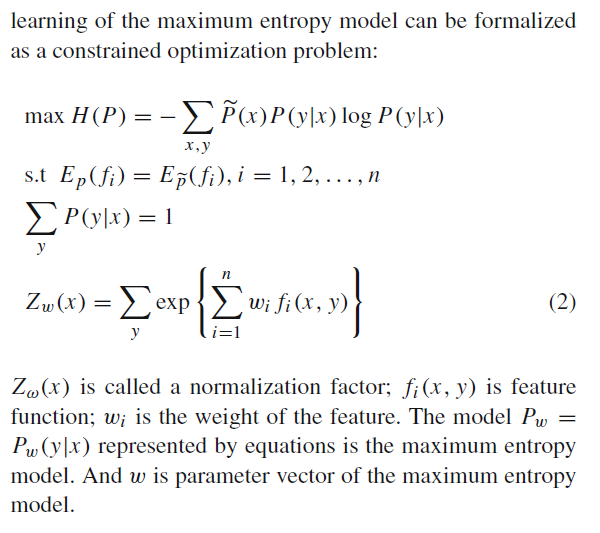


**3.2 The use of the maximum entropy model**

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.







**4 Maximum entropy-PLSA model**

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.

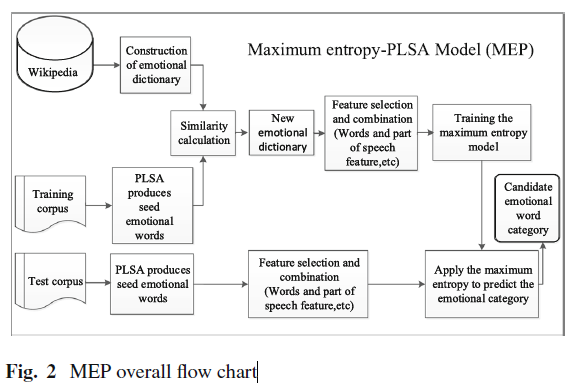
The maximum entropy classification method can flexibly define the eigenfunctions and can better estimate the probability distribution of unknown words in the context of known context constraints.

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.



**4.1 Construction of emotional dictionary**

Emotional seed is the words with absolute emotional meaning.

*Degree adverb*

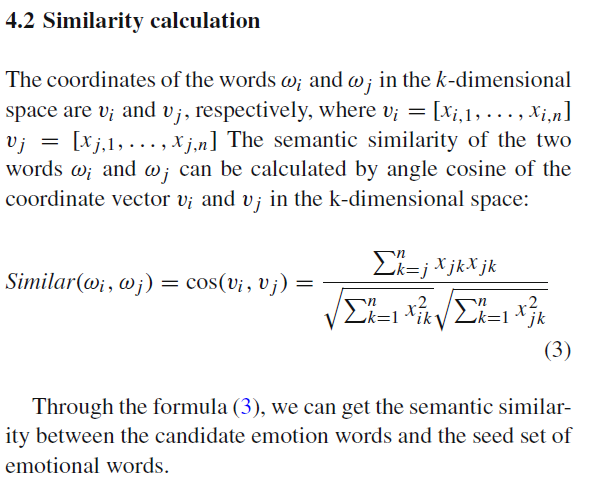
*Expressive verbs*

*Interjection*

*Conjunctions*

*Negative vocabulary*

**4.2 Similarity calculation**



**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.

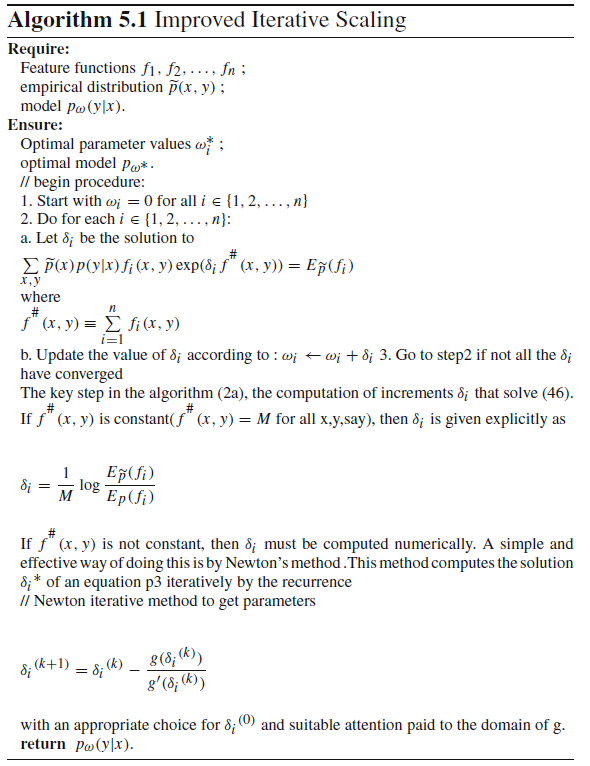
1. Words and part of speech features
2. Word feature similarity
3. Correlation with degree adverb Cadv
4. Correlation with expressive verbs *Cv*
5. Similarity features with baseline emotional words set *Sw*
6. Similar features with the convert form of the known emotional words *Sw*−*convert*

**5 Model solving and inference**

**5.1 PLSA solution**

**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.



**6 Experimental results and analysis**

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

**6.1 Experiments on restaurant review corpus**

**6.2 Experiments on film review corpus**

**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.