**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.

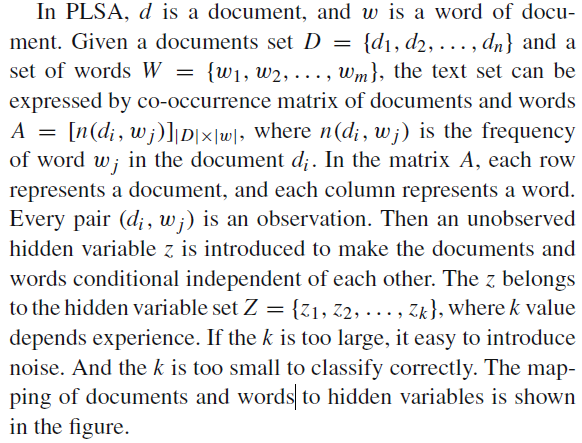
3. Combining the characteristics of language and being based on emotional word recognition, this paper puts forward a Sentence Recognition Method of fusing multifeatured weights such as emotional words, degree adverbs, negative words and so on.

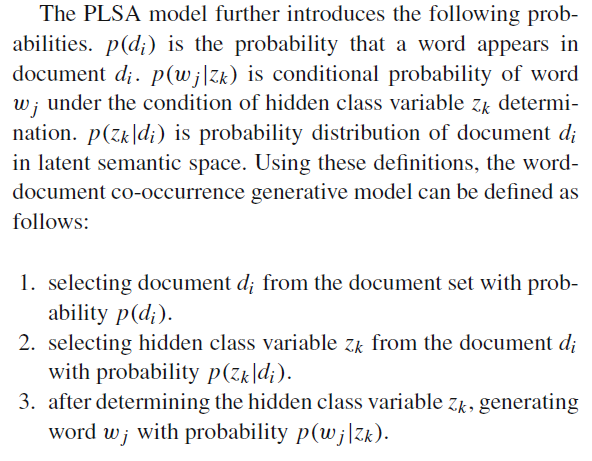
**2 Related works**

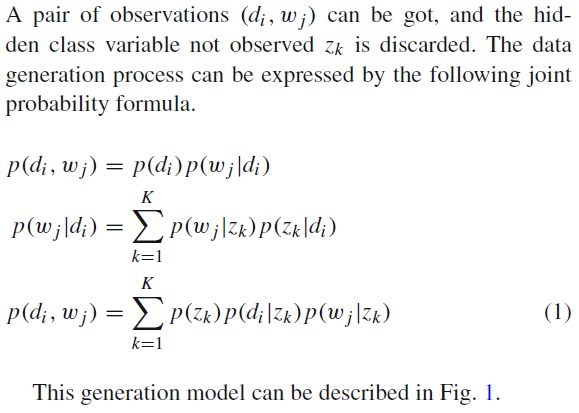
**3 An improved sentiment analysis algorithm**

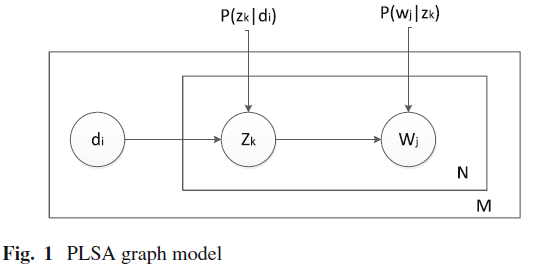
**3.1 Introduction of PLSA**

The semantic space is the key to PLSA, which associates each pair of binary data with hidden class variables.





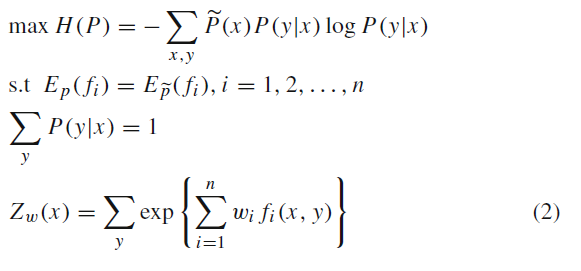




**3.2 The use of the maximum entropy model**

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*), . . . , (xN , yN )*} and feature function *fi (x, y), i* = 1*,* 2*, . . . , n*, the learning of the maximum entropy model can be formalized as a constrained optimization problem:



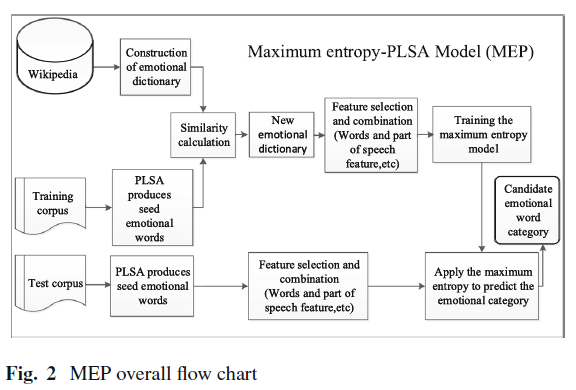
*Zω(x)* is called a normalization factor; *fi (x, y)* is feature function; *wi* is the weight of the feature. The model *Pw* = *Pw(y*|*x)* represented by equations is the maximum entropy model. And *w* is parameter vector of the maximum entropy model.

**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.

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.

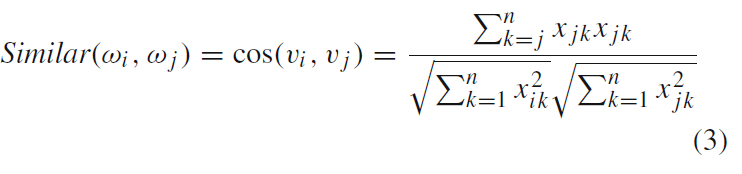


**4.1 Construction of emotional dictionary**

Emotional seed is the words with absolute emotional meaning. Emotional analysis of the phrase set

contains the evaluation of words and emotional words and is divided into two categories.

**4.2 Similarity calculation**



**4.3 Feature selection and combination**

**5 Model solving and inference**

**5.1 PLSA solution**

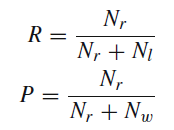
**5.2 Maximum entropy solution**

The maximum entropy model is based on the maximum entropy principle to evaluate parameters for each feature.

**6 Experimental results and analysis**

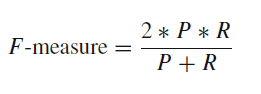
In this paper, we use the precision and recall rate to evaluate the performance of emotional word recognition and classification experiments.

The recall rate and precision rate aredefined as follows:



*F*-measure can correctly reflect the effect of the emotional analysis model in the balance of

precision and recall.



**6.1 Experiments on restaurant review corpus**

**6.2 Experiments on film review corpus**

**7 Conclusion and prospect**