Unsupervised Multi-class Sentiment Classification Approach

Liwei Xu*, Jiangnan Qiu**

Dalian University of Technology, Faculty of Management and Economics, Dalian 116024, China,
*<xuliwei@mail.dlut.edu.cn>, **<qiujn@dlut.edu.cn>

Liwei Xu is a PhD candidate at the Faculty of Management and Economics, Dalian University of Technology, China. She received her BS and MS degrees in information management and information system and computer software and theory from Northeast Forestry University. Her papers have received best paper awards at the China Association for Information Systems Conference. Her current research interests include machine learning, emergency management, sentiment analysis, and big data analysis.

Jiangnan Qiu is a professor at the Institution of Information management and information systems, Dalian University of Technology, China. He received his BS, MS and PhD degrees in engineering mechanics, computer software and theory, and management science and engineering from Dalian University of Technology. His work has been published in *Decision Support Systems*, *Journal of Knowledge Management*, and *Behaviour & Information Technology*, among others. His research interests are emergency management and big data analysis.

Xu, Liwei and Jiangnan Qiu. 2019. "Unsupervised Multi-class Sentiment Classification Approach." *Knowledge Organization* 46(1): 15-32. 64 references. DOI:10.5771/0943-7444-2019-1-15.

Abstract: Real-time and accurate multi-class sentiment classification serves as a tool to gauge public user experiences and provide a decision-making basis for timely analysis. In the field of sentiment classification, there is an urgent need for an accurate and efficient multi-class sentiment classification method. With the aim to overcome the drawbacks of the existing methods, we propose a novel, unsupervised multi-class sentiment classification method called Gaussian mixture model of multi-class sentiment classification (GMSC). Based on the Gaussian mixture model (GMM), the GMSC consists of the following essential phases: first, combining a dictionary with microblog texts to calculate and construct the feature matrix of sentiment for each sample; second, introducing a dimension reduction method to avoid the influence of a sparse feature matrix on the results; third, modeling the multi-class sentiment classification procedure based on GMM; and lastly, computing the probability distribution of different categories of sentiment by using GMM to partition sentiments in microblogs into distinct components and classify them via a Gaussian process regression. The results indicate the GMSC approach's accuracy is better and manual tagging time is reduced when compared to semi-supervised and unsupervised sentiment classification methods within the same parameters.

Received: 30 May 2018; Revised: 29 November 2018; Accepted 20 December 2018

Keywords: sentiment, classification, model, method, unsupervised

1.0 Introduction

With the rapid development of social media, people are increasingly using platforms such as Sina Microblog and Twitter to share their views about and experiences with a particular product, policy, or event. Considering the economic value of user-generated content (Ghose and Ipeirotis 2009; Goh, Heng and Lin 2013), there is a growing trend of using user reviews to promote a product (Goh, Heng and Lin 2013). Moreover, reviews, comments, tweets, microblogs, and other user-generated content express the user's sentiment or attitude about a product, which decision makers can reference to improve the design and quality of products. Therefore, user-generated content is considered an important source of information in sentiment analysis applications for decision making. Meanwhile, due to the growing interest in determining the exact sentiment within a text, sentiment classification has become an active area of research (Dai et al. 2012).

User-generated content poses different challenges due to the unstructured nature of online texts. Currently, the methodologies commonly used in sentiment classification can be categorized into supervised, semi-supervised (Sindhwani and Melville 2008; Zhou, Chen and Wang 2010) and unsupervised (Peng and Shih 2010). The first two methods have the disadvantage of requiring manual labeling, which may increase the time cost, especially in terms of big data analyses (Cambria et al. 2013). Thus, in big data research contexts, unsupervised sentiment classification methodology is popular. However, it still faces

several challenges: 1) weak generalization of the classifier (Lin, He and Everson 2010); 2) simple categorization of sentiment (Liu, Bi and Fan 2017a and 2017b); and, 3) high time complexity.

First, the issue of low accuracy arises due to weak generalization capabilities and a crude pre-tagging process. Most existing methods employ different superimposed ways to expand the training set of pre-labeled samples, thereby reducing the classification ability of unknown data for an overall model. The second issue is the limited classes of sentiment type. Most of the aforementioned methods only consider a sentiment as positive or negative without full consideration of sentiment diversity and multi-class classification of a specific sentiment. Some scholars (Liu, Bi and Fan 2017a and 2017b) have pointed out the excessive simplicity of classifying online texts as displaying only positive or negative sentiments while multi-class sentiments (e.g., happy, good, sad, anger, fear, disgust, surprise) can provide more information to companies whose relevant department can use them to modify their advertisements, product design, and so on, which leads to more effective decisions. The third challenge is decreasing the complexity of the sentiment classification methodology by optimizing the parameters. Many classification tasks lead to a long run time, which decreases the efficiency of acquiring real-time information.

To fill these research gaps, this paper proposes a novel unsupervised multi-class sentiment classification approach based on the Gaussian mixture model (GMM) called Gaussian mixture model of multi-class sentiment classification (GMSC). Because GMM is a generative and unsupervised classification methodology, which differs from general sentiment classification approaches, such as naïve Bayesian and support vector machine, it can take advantage of an unlabeled dataset and suits multi-class classification issues. GMM is also a non-parametric probabilistic method, i.e., it does not need any parameters set in advance; meanwhile, an algorithm can optimize the parameters. Moreover, varied historical data can be used to obtain the parameters of an adaptive training model (Qiao et al. 2015) for which different categories of datasets have better flexibility. Therefore, GMM can solve the issues of weak generalization ability, high time complexity, etc., and it suits multi-class sentiment classification of user-generated content.

A novel modeling approach, GMSC consists of the following essential phases: first, combining a dictionary with user-generated content to calculate and construct the feature matrix of the sentiment for each sample. Second, introducing a dimension reduction method to prevent the influence of a sparse feature matrix on the results. Third, a modeling multi-class sentiment classification procedure based on GMM. Lastly, calculating the probability distribution for different categories of sentiment by using GMM to partition sentiments into distinct components and classify them via a Gaussian process regression. The proposed method of GMSC uses different categories of training samples to obtain parameters for each Gaussian process of the model, which has a stronger generalization ability for data when combining a density peaks clustering (DPC) algorithm (Rodriguez and Laio 2014) to initialize parameters and an estimation maximum (EM) algorithm to optimize parameters as a gradual iterative process. Therefore, its time complexity can be reduced.

The paper is organized as follows: Section 2 presents approaches to sentiment classification in recent research. Section 3 describes the basic concepts and framework of the model. The proposed approach is detailed in Section 4, while the experimental process and results analysis are presented in Section 5. Finally, Section 6 concludes the study and outlines avenues for future research.

2.0 Related work

Sentiment classification involves several steps. First is extracting sentiment features. A commonly used methodology is the "bag of words" model, which has the disadvantages of a giant vector space dimension and a sparse matrix. As the first step of this study, we chose several sentiment vectors based on the lexicon to construct the bag of words model, which may decrease the giant vector space dimension. Fernández-Gavilanes et al. (2018) also used a lexicon to extract sentiment features and further construct an unsupervised sentiment classification model. Lin, He and Everson (2010) illustrated that the joint sentiment topic (JST) model is an unsupervised sentiment classification model, and it relies on a sentiment lexicon as a basis to extract features. And using a bag-of-words model based on a lexicon has another advantage; microblogs can be represented as different sentiment dimensions, which is more comprehensive. In the second step of sentiment classification, some researchers reduce the feature dimensions and then use the classification methodology. Each text can be classified as one main type of sentiment.

Sentiment classification methodologies can be categorized into two types: semantic orientation and machine learning methodology.

1) Semantic orientation: This type of method compares words in the sample with a sentiment lexicon to judge sentiment tendency (Hogenboom et al. 2014; Taboada et al. 2011; Wang et al. 2014). Wang et al. (2012) analyzed the sentiment tendency of comments on the 2012 U.S. presidential election by using Twitter Firehose and expertly curated rules and keywords to get a full and accurate picture of the online political landscape. Agarwal et al. (2011) presented the results for sentiment analysis on

Twitter data by introducing POS-specific prior polarity features and exploring the use of a tree kernel to preclude the need for tedious feature engineering, while Jiang et al. (2011, 151) incorporated target-dependent features and also considered Tweets related to the one question by utilizing graph-based optimization, which can improve the accuracy of target-dependent Twitter sentiment classification. The main drawback of these approaches is the inability to deal with domain- and context-specific orientations, the continuous appearance of new words (Melville, Gryc and Lawrence 2009), changes in expression patterns, and complicated language-processing problems (Sarvabhotla, Pingali and Varma 2001). Even so, they can be used to pre-tag because of their simple operation.

2) Machine learning: This method requires constructing a machine-learning model (Serrano-Guerrero et al. 2015), typically by inputting training samples into a classification model (Mai 2004a and 2004b; Café and Souza 2017; Albrechtsen and Pejtersen 2003), and entering the test sample produces results (Boiy and Moens 2009). These methodologies fall into three main types: supervised, semi-supervised, and unsupervised (Ouyang et al. 2014).

    1. Supervised: This classification model involves manually pre-tagging training datasets (Kim, Howland and Park 2005; Tang, Tan and Cheng 2007). Pang and Lee (2004) were the first to employ three existing supervised learning methods (i.e., naive Bayes, maximum entropy and support vector machine) to classify movie reviews as positive or negative. Later on, Pang and Lee (2008) improved on accuracy through the use of efficient techniques to find minimum graph cuts to classify texts into subjective or objective. Since the work of Pang and Lee (2004 and 2008), various models and features have been proposed to improve classification. For example, Xu, Ding and Wang (2007) utilized naive Bayesian and maximum entropy methods to classify the sentiment of Chinese news and reviews; they proved machine learning can achieve good results. Li and Huang (2010) combined stacking sentimental classification and finally overcame the dependence dilemma of classification methods in the field. Despite the method's superior classification performance and popularity, it can hardly process a training dataset in a timely manner (Tang, Tan and Cheng 2007).
    2. Semi-supervised: This method combines unlabeled data with labeled training data (often small-scaled) to improve the model (Silva et al. 2016; Serrano-Guerrero et al. 2015; Zhang, Xu and Wan 2012). Tan, Wang and Cheng (2008) integrated lexicon-based and corpus-based approaches by using an unsupervised technique to label examples for a supervised classifier. However, the latter stage does not involve adopting any strategy to take full advantage of unlabeled data. Other attempts to adopt a self-training strategy such as Li et al. (2011) investigated semi-supervised learning for imbalanced sentiment classification and generate different views from random feature subspaces. Wan (2011) applied a co-training method to semi-supervised learning with a labeled English corpus and an unlabeled Chinese corpus for Chinese sentiment classification. Sindhwani and Melville (2008) proposed a semi-supervised sentiment prediction algorithm that utilizes lexical prior knowledge in conjunction with unlabeled examples based on a bipartite graph representation of the labeled and unlabeled. Zhang and He (2013) applied a variant of a self-training algorithm on two partitions split from a test dataset and combined the classification results into a pseudo-labeled training set and an unlabeled test set; then they trained an initial classifier on the pseudo-labeled training set and adopted a standard self-learning cycle to obtain the overall classification results. Although this method requires just a small number of labeled samples to save on manpower and material resources (He and Zhou 2011; Li et al. 2010), its generalization ability for unknown datasets is reduced due to the training corpus' expansion through different methods.
    3. Unsupervised: This method automatically performs sentiment classification through model construction or mutual information calculation of a seed word, which does not require labeled instances to derive a sentiment classifier (Smailović et al. 2014; Fang, Dutta and Datta 2014). Turney (2002) pioneered predicting the sentiment orientation of a text via the average sentiment orientation of the extracted phrases that contain adjectives or adverbs. The sentiment orientation of a phrase is estimated using point-wise mutual information. Kennedy and Inkpen (2006) used an enhanced term-counting method to determine the sentiment orientation of a customer review by counting positive and negative terms and taking into account contextual valence shifters. Zagibalov and Carroll (2008, 1073) first proposed the seed word of automatic selection and statistical method to classify the sentiment of Chinese texts. The experimental results demonstrate the accuracy is close to a supervised classification method. Zhai, Xu and Jia (2010) explored an unsupervised sentiment classification method in the case of Chinese sentiments, using unlabeled data to identify and remove noise words to improve accuracy. Wan

(2008) identified sentiment polarity of Chinese reviews by making full use of bilingual knowledge in an unsupervised way. The unsupervised method proposed by Dai et al. (2012) involves manually categorizing texts as positive or negative and combining it with a semi-supervised method to construct the sentiment classifier. Lin, He and Everson (2010) presented a comparative study of the latent sentiment model (LSM), the joint sentiment topic (JST) model, and the reverse-JST model. The results suggest the JST model is more appropriate than reverse-JST for sentiment classification. Although this method does not need manual tagging (Dasgupta and Ng 2009), one disadvantage is that the selection process of the seed word is difficult, so this may negatively affect performance. Another drawback is its high time complexity, which is not ideal for real-time information acquisition.

The aforementioned methods are summarized in Table 1.

Generally, supervised sentiment classification may have the highest classification accuracy, but it requires manually labeling data, which may increase time cost and decrease generalization ability. Although semi-supervised sentiment classification utilizes less manually labeled data to train the classification model, it also has the same problems as a supervised model. On the other hand, unsupervised sentiment classification solves the problem of manual labeling and saves on time cost, which is especially suitable for big data analysis. Moreover, it may provide more accurate and effective analyses to support the decisions of governments, companies, and consumers. However, the current research still has some limitations: 1) weak generalization (Lin, He and Everson 2010); 2) simple categorization—the current research mainly focuses on positive and negative sentiment classification; thus, multi-class sentiment classification is necessary (Liu, Bi and Fan 2017a and 2017b); and, 3) high time complexity.

Current research has categorized sentiment classification methodology as supervised, semi-supervised, and unsupervised from a labeled dataset perspective and compared their performance but have not analyzed their differences. In terms of methodology, generative, and discriminative are regarded as different perspectives. With a generative approach, we model the joint distribution between variable x and label y as $P(x, y) = P(x|y)P(y)$. This can be done by learning the class prior probabilities $P(y)$ and the class-conditional densities $P(x|y)$ separately. In contrast, within a discriminative approach, a parametric model for the posterior probabilities is constructed as $P(x, y) = P(y|x)$, and the values of the parameters from a set of labeled training dataset are inferred. The generative approach focuses on the different probability distributions of each class, while the discriminative approach focuses on how to separate data from different classes.

Although the discriminative approach results in high classification performance, it lacks flexible modeling tools and adding prior knowledge is difficult. Therefore, it cannot reflect the characteristics of the training data itself and loses information contained in the sample distribution p(x). The relationship is not as clear as in the generative approach. The problem-solving process is like a "black box." When labeled training dataset is plentiful, the discriminative methodology results in excellent generalization. However, manual labeling increases time cost, and, especially in the context of big data analysis, it would be rather expensive.

Consequently, there is increasing interest in the generative approach as it can take advantage of an unlabeled dataset (Suzuki, Fujino and Isozaki 2007; Bernardo et al.

Sentiment Classification				
Method	**Semantic Orientation**	**Machine Learning**		
		Supervised	**Semi-supervised**	**Unsupervised**
Label	**No**	**Yes**	**A little**	**No**
Feature	(1) Simple, easy to realize (2) Poor applicability	(1) High accuracy (2) Time-consuming	(1) Saves tagging time (2) Weak generalization ability, etc.	(1) No manual tagging (2) Difficult seed word selection, high complexity etc.
Typical literature	Jiang et al. (2011) Wang et al. (2012)	Pang and Lee (2004) Xu, Ding and Wang (2007) Li and Huang (2010)	**Zhang and He (2013)** Li et al. (2011) Wan (2011) Sindhwani and Melville (2008)	**Wan (2008)** **Dai et al. (2012)** **Lin, He and Everson (2010)** Zhai, Xu and Jia (2010) Turney (2002) Kennedy and Inkpen (2006) Zagibalov and Carroll (2008)

Table 1. Summary of sentiment classification methods (references in bold show the selected comparison methodology).

2007). When dealing with multi-classification problems, the training efficiency of the discriminative approach is often lower than that of the generative approach. Many successful discriminative classification models have been originally proposed for two types of classification problems. When dealing with multiple types, it is usually necessary to convert one-class C problem into a two-class C problem and then merge the results (Rifkin and Klautau 2004).

Because the generative approach can use an unlabeled dataset and classify the sentiment of user-generated content into multiple types, it suits multi-class sentiment classification of user-generated content. GMM, as a generative approach, involves K Gaussian distribution. Each Gaussian distribution is called a Gaussian process, which represent different sentiment classifiers and the probability density function of GMM. We use GMM to construct the multi-class sentiment classification model GMSC.

Considering that GMSC does not involve tagging samples and a sentiment lexicon is only used to obtain the feature matrix, we selected semi-supervised and unsupervised methods to compare them to GMSC. We did not choose a supervised sentiment classification methodology, because manual labeling has a high time cost, especially in the context of big data, and the time cost is difficult to measure. Zhang and He (2013) found that the proposed methodology obtained better results compared to other semi-supervised methods. They proposed a typical semi-supervised methodology. Wan (2008) put forward the unsupervised sentiment classification method without tagging samples; the methodology involves matching words with the sentiment lexicon. Dai et al. (2012) improved the accuracy of the unsupervised methodology from multi-angle. Lin, He and Everson (2010) constructed a simultaneous consideration of the topic and sentiment for the unsupervised method, which has broader scope of application.

3.0 Overview and concepts

This section introduces the framework and principle of the proposed approach and several important concepts.

3.1 Overview

Figure 1 presents the framework of the proposed method and its two stages: 1) extracting features and dimension reduction using PCA methodology; and, 2) identifying multi-class sentiment of user-generated content based on a DPC unsupervised clustering algorithm and GMM methodology.

3.2 Concepts definition

This section provides brief overviews of relevant concepts. D, regarded as a known object database that stores a large number of user-generated content, is represented by $\mathrm{D} = \{D_1, D_2, \ldots D_i \ldots D_n\}$. The number of user-generated content is defined as $|\mathrm{D}|$.

The sentiment of user-generated content is defined as follows:

Definition 1 (feature calculation of sentiment): m represents the number of multi-class sentiments in the lexicon; user-generated content D_i contains p sentiment of the lexicon. According to the characteristics of the lexicon, the intensity and polarity are combined to compute a feature value of each sentiment; EV_{nm} equals the sum of intensity and polarity of sentiment, which denotes the feature value of sentiment m of user-generated content D_n in the equation as follows:

$$\begin{cases} EV_{nm} = \sum_{i'}^{w} \alpha \times intensity_{i'} + \beta \times polarity_{i'} \\ \alpha + \beta = 1 \quad , \quad \alpha、\beta \in [0,1] \end{cases} \tag{1}$$

Where α、β denote adjustment coefficients, different values reflect the influence degree of intensity and polarity on EV_{nm}; $intensity_{i'}$ denotes intensity value of sentiment word i', and $polarity_{i'}$ denotes polarity value of sentiment word i' in user-generated content. The way to calculate the feature EV_{nm} in equation (1) is based on the lexicon value of intensity and polarity.

Definition 2 (sentiment vector representation): The feature matrix of sentiment obtained by definition one can be derived as follows:

$$EV_{n\,m} = \begin{bmatrix} EV_{1\,1} & EV_{1\,2} & \ldots & EV_{1\,m-1} & EV_{1\,m} \\ EV_{2\,1} & EV_{2\,2} & \ldots & EV_{2\,m-1} & EV_{2\,m} \\ \vdots & & & & \vdots \\ EV_{i\,1} & EV_{i\,2} & \ldots & EV_{i\,m-1} & EV_{i\,m} \\ \vdots & & & & \vdots \\ EV_{n\,1} & EV_{n\,2} & \ldots & EV_{n\,m-1} & EV_{n\,m} \end{bmatrix} \tag{2}$$

Definition 3 (dimension reduction of sentiment feature matrix): We use a dimension reduction algorithm to reduce the dimension of EV_{nm} from m to t. The low-dimensional sentiment feature matrix EV_{nt} is as follows:

$$EV_{n\,m} = \begin{bmatrix} EV_{1\,1} & EV_{1\,2} & \ldots & EV_{1\,m-1} & EV_{1\,m} \\ EV_{2\,1} & EV_{2\,2} & \ldots & EV_{2\,m-1} & EV_{2\,m} \\ \vdots & & & & \vdots \\ EV_{i\,1} & EV_{i\,2} & \ldots & EV_{i\,m-1} & EV_{i\,m} \\ \vdots & & & & \vdots \\ EV_{n\,1} & EV_{n\,2} & \ldots & EV_{n\,m-1} & EV_{n\,m} \end{bmatrix} \Rightarrow$$

$$\boldsymbol{EV_{nt}} = \begin{bmatrix} EV_{1\,1} & EV_{1\,2} & \ldots & EV_{1\,t} \\ EV_{2\,1} & EV_{2\,2} & \ldots & EV_{2\,t} \\ \vdots & & & \vdots \\ EV_{i\,1} & EV_{i\,2} & \ldots & EV_{i\,t} \\ \vdots & & & \vdots \\ EV_{n\,1} & EV_{n\,2} & \ldots & EV_{n\,t} \end{bmatrix} \tag{3}$$

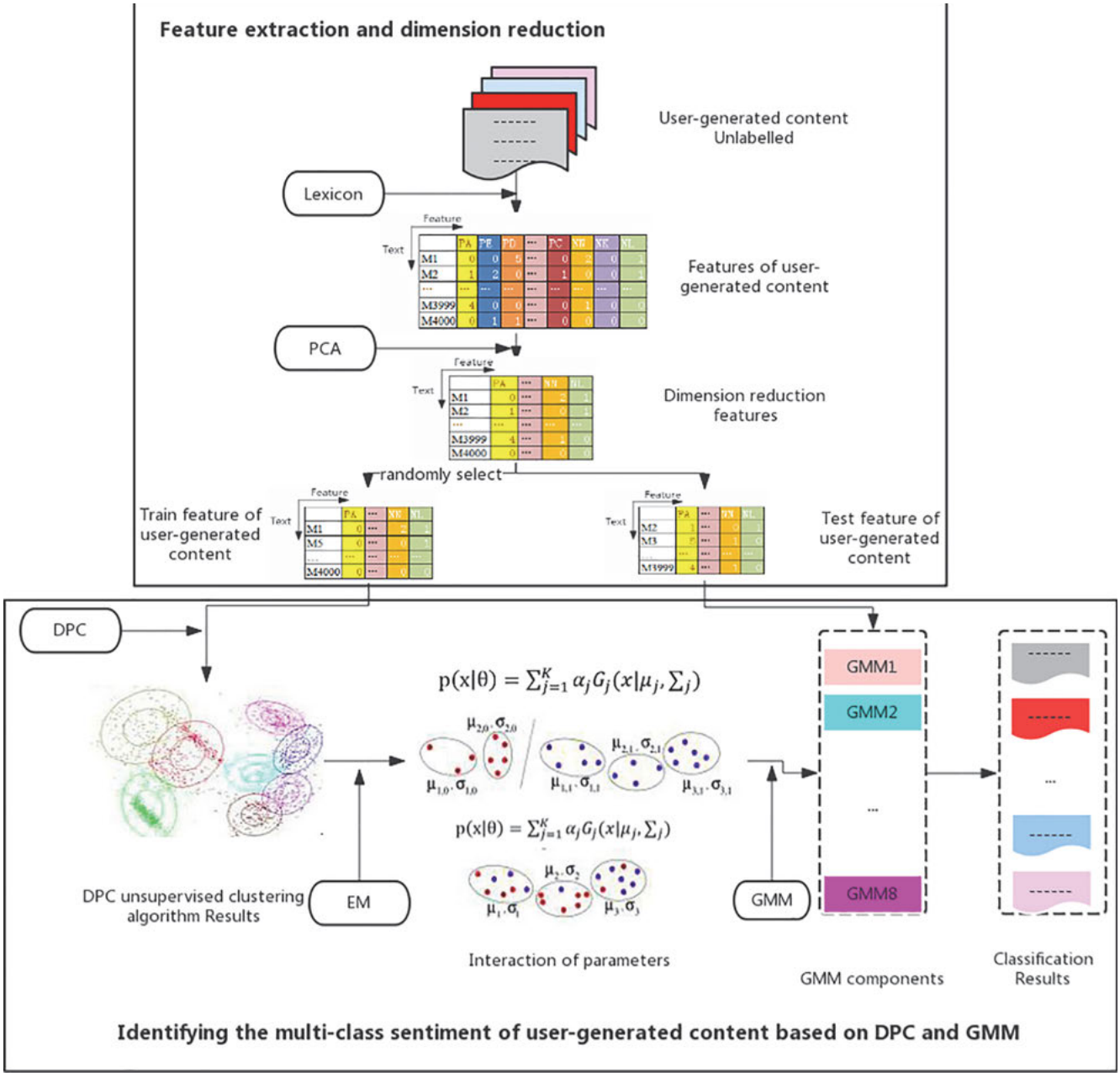

Figure1. Overall process of the proposed approach.

Definition 4 (Gaussian process regression): For training dataset $\mathrm{D} = \{(x_i, y_i)\}_{i=1}^{N} = (X, Y)$, where $x_i \in R^d$, $\mathrm{X} = [x_1, x_2, \cdots x_n]$ denotes the feature matrix of $\mathrm{d} \times \mathrm{n}$ dimension, and $y_i \in R$ is the output result. A given set X can form a set of random variables $\{\mathrm{f}(x_1), \mathrm{f}(x_2), \cdots \mathrm{f}(x_N)\}$, which satisfies a joint Gaussian distribution (Sung 2004), so the mean $\mathrm{m}(\mathrm{x})$ and covariance function $\mathrm{k}(\mathrm{x}, x')$ can denote the Gaussian process as follows:

$$\begin{cases} \mathrm{f}(\mathrm{x}) \sim \mathrm{G}(\mathrm{m}(\mathrm{x}), \mathrm{k}(\mathrm{x}, x')) \\ m(x) = E[f(x)] \\ k(x, x') = E((f(x) - m(x)) - (f(x') - m(x'))) \end{cases} \quad (4)$$

We choose a standard exponential covariance function as in the following:

$$\begin{cases} \mathrm{cov}(\mathrm{f(x)}, \mathrm{f}(x')) = \mathrm{k(x}, x') = \theta_0^2 \exp\left(-\frac{(x-x')^2}{2\theta_1^2}\right) + \sigma^2\delta_{ij} \\ \mathrm{i} = \mathrm{j}, \delta_{ij} = 1 \\ i \neq j, \delta_{ij} = 0 \end{cases} \quad (5)$$

where θ_0 , $\theta_1$denote weight and size respectively; δ_{ij} is the Dirac function.

Definition 5 (Gaussian mixture model of sentiment): Each sentiment can be denoted as a function of GMM, that is, the linear combination of multiple GMMs compose the dataset. In Eq. (4) the dataset x is the linear combination of the independent GMM of all components $\mathrm{G} = \{G_1, G_2, \cdots, G_K\}$, and $\alpha_1, \alpha_2, \cdots, \alpha_k$ refer to the weight of all variables. Therefore, the probability density function of x can be represented as:

$$\begin{cases} \mathrm{p(x}|\theta) = \sum_{j=1}^{K} \alpha_j G_j(x|\mu_j, \sum_j) \\ \sum_{j=1}^{K} \alpha_j = 1 \\ \theta = \{\alpha_j, \mu_j, \sum_j\}, \quad j = 1,2, \cdots K \end{cases} \quad (6)$$

where $G_j(\cdot)$ denotes the probability density function of Gaussian mixture model; α_j is the weight for the Gaussian process j.

4.0 The proposed approach

The GMM algorithm precisely quantifies things via the Gaussian probability density function (normal distribution curve). Gaussian process regression is used to describe the distribution of probability space. That is, each category is represented by a Gaussian process, and the weighted combination of all Gaussian processes constitute GMM. The GMM algorithm is primarily used in object detection, trajectory prediction, speaker recognition, image recognition, text classification, and other contexts to efficiently and automatically classify an object.

Given that low standardization, a prevalence of new words, and dynamically changing expressions characterize language used in user-generated content, this leads to data processing challenges. The GMM model is dependent on each sample for sentiment classification, and the gradual iterative estimation maximum (EM) algorithm is introduced to obtain optimal parameters, which increases the generalization ability and reduces time complexity. Thus, in this paper, we consider the GMM algorithm as the basis of an unsupervised multi-class sentiment classification method. Meanwhile, as to the rigorous initial value selection problem of GMM, the DPC algorithm is introduced. In general, the method consists of three components: 1) GMM: the core of the method to perform unsupervised multi-class sentiment classification; 2) EM: this algorithm solves the selection problem of the initial value of GMM and obtains optimal parameters; and, 3) DPC: this algorithm solves the initial value of the selection problem of EM, regarding it as an initial algorithm to obtain clustering numbers of the EM algorithm.

4.1 GMM algorithm

4.1.1 GMM definition

GMM is commonly used in describing the model of mixture density function distribution. Several weighted Gaussian probability density functions are added to describe the distribution of the feature vector in probability space. That is, GMM is composed of K Gaussian distribution, and each Gaussian distribution is called a Gaussian process, which represents different sentiment classifiers and the probability density function of GMM. The latter can be denoted as follows:

$$\begin{cases} \mathrm{p(x_n}|\theta) = \sum_{j=1}^{K} \alpha_j G_j(x_n|\mu_j, \sum_j) \\ \sum_{j=1}^{K} \alpha_j = 1 \\ \theta = \{\alpha_j, \mu_j, \sum_j\}, \quad j = 1,2, \cdots K \end{cases} \quad (7)$$

where, $G_j((x_n|\mu_j, \sum_j)$ denotes the Gaussian probability density function of sentiment vector x_n to sentiment j. K is the numbers of sentiment, α_j denotes the weight of the sentiment j, and $\mu_j, \sum_j$ refer to the mean and covariance of sentiment j, respectively.

The Gaussian probability density function is defined by the following:

$$\begin{cases} G_j(x_i|\mu, \sum) = \frac{1}{\sqrt{(2\pi)^d|\sum|}} \exp\{-\frac{1}{2}(x_i - \mu)^T \sum^{-1}(x_i - \mu)\} \\ \mu = [E(x_1), E(x_2), \cdots E(x_d)^T \\ \sum = (C_{ij})_{d\times d}, C_{ij} = COV(x_i, x_j) \end{cases} \quad (8)$$

The GMM corresponding likelihood function can be represented as:

$$\mathrm{P(X}|\theta) = \prod_{i=1}^{N} p(x_i|\theta) \quad (9)$$

Therefore, the key step in the construction process is to obtain a $\boldsymbol{\theta}$ value of the model. To do so, we maximize the probability of the sentiment training model by using Eq. (9) to obtain an optimal value of $\boldsymbol{\theta}$ for the analysis of sentiment classification. The detailed process solution is described in the appendix.

The graphical model of GMM is shown in Figure 2, wherein each component of the Gaussian process is represented by the parameters of mean and covariance. The classification process is an ellipsoid distribution of mean-

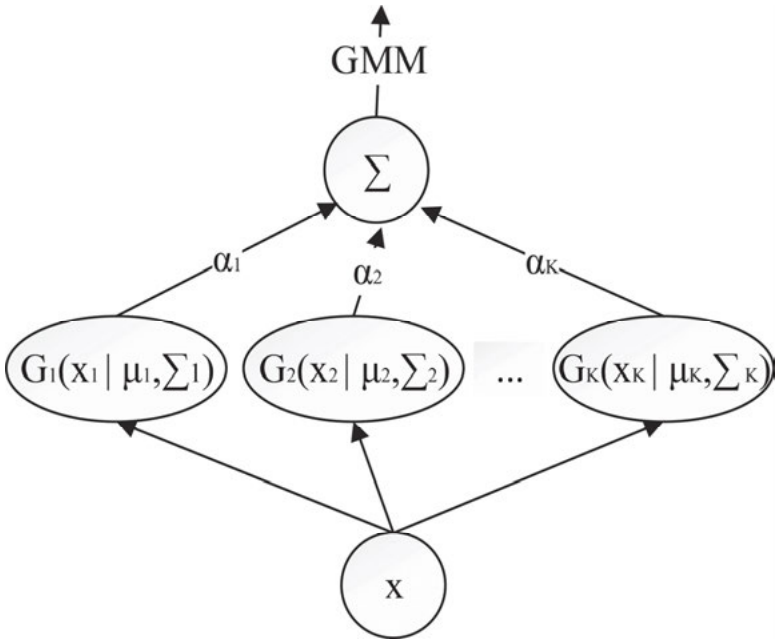

Figure 2. The graphical model of GMM.

centered, and geometric properties is determined by covariance.

4.1.2 GMM principle

The basic idea of GMM involves three steps: first, construct the model using the probability density function; second, use the EM algorithm to obtain the optimal solution of corresponding parameters, according to the normal distribution of the conditional distribution to obtain the K Gaussian process function; lastly, calculate the sentiment category value of the test sample.

The GMM classification model is based on the sentiment sample to compute the probability distribution of GMM, using a model to obtain the K Gaussian process and then combine definition four of the Gaussian process to classify testing samples. The process of unsupervised multi-class sentiment classification based on GMM can be derived as follows:

$D_{train} = (X, Y), D_{test} = (X', Y')$ represent the training and testing dataset respectively, where X denotes the sentiment feature matrix of training dataset, Y is the sentiment classification result of training dataset, and X', Y' are similar to X, Y. The joint probability density function of $[\mathrm{Y}, Y'^{T}]$ complies with the following GMM model:

$$\begin{cases} \mathrm{p}_{YY'}(Y, Y') = \sum_{i=1}^{K} \alpha_i G(Y, Y' | \mu_i, \sum_i) \\ \mu_i = [\mu_{iY}, \mu_{iY'}] \\ \sum_i = \begin{bmatrix} \sum_{iY} & \sum_{iYY'} \\ \sum_{iY'Y} & \sum_{iY'} \end{bmatrix} \end{cases} \tag{10}$$

where $\sum_{i=1}^{K} \alpha_i = 1$, the joint probability density function is:

$$p_{YY'}(Y, Y') = \sum_{i=1}^{k} \alpha_i G(Y' | Y, f_i\hat{}(Y), \sigma_i^2) \tag{11}$$

and:

$$\begin{cases} f_i\hat{}(Y) = E[Y' | Y] = \mu_{iY'} + \sum_{iY'} \sum_{iYY}^{-1} (y - \mu_{iY}) \\ \sigma_i^2 = Var[Y' | Y] = \sum_{iY'} - \sum_{iY'Y} \sum_{iYY}^{-1} \sum_{iYY'} \end{cases} \tag{12}$$

The marginal density function of Y can be formulated as:

$$p_y(y) = \int_{p_{YY'}} (Y, Y')\, dy = \sum_{i=1}^{K} \alpha_i G(Y, \mu_{iY}, \sum_{iY}) \tag{13}$$

The conditional density function as:

$$p_{Y'|Y} = \sum_{i=1}^{K} \phi_i(Y) G(Y, f_i\hat{}(Y), \sigma_i^2) \tag{14}$$

where weight can be defined as the following:

$$\phi_i(Y) = \frac{\alpha_i G(Y, \mu_{iY}, \sum_{iY})}{\sum_{i=1}^{K} \alpha_i G(Y, \mu_{iY}, \sum_{iY})} \tag{15}$$

The sentiment category value of Y' is obtained as follows :

$$\overline{Y'} = \mathrm{f}\hat{}(Y) = \mathrm{E}[Y' | \mathrm{X}, \mathrm{Y}, X'] = \sum_{i=1}^{K} \phi_i(Y) f\hat{}(Y) \tag{16}$$

4.2 GMM initialization process

An efficient classification model construction of GMM depends on the accurate initial value of θ. The EM algorithm is commonly used in parameter estimation through gradual iteration to improve the value of parameter θ. As iteration process increases, the match rate between the estimation parameter θ and training sentiment sample x_i until it fulfils the formula $\mathrm{P}(\mathrm{X}|\theta^{k+1}) > P(X|\theta^{k})$ each time, where k denotes the number of the iteration.

Although the log-likelihood of the observed data keeps increasing by combining the EM algorithm of the GMSC approach at each iteration, the main drawback is slow convergence. Therefore, to overcome this issue and reduce the influence on the final result of the GMSC approach, we consider how to select good initial parameters as follows: assuming that the parameter **θ** of the EM algorithm in Section 4.2 is already known, then the initialization for the EM algorithm is necessary. At present, the commonly used method combines K-means (Jing et al. 2007; Roy and Sharma 2010) and the EM algorithm, using the center value calculated by the K-means algorithm (Jing et al. 2005) as an initial input parameter of the EM mean to determine a rough classification of the initial sample; however, the K-means algorithm needs the cluster number.

The density peaks clustering (DPC) algorithm is a density-based clustering algorithm and an unsupervised clustering algorithm without the input of cluster number (Rodriguez and Alessandro 2014). Its main purpose is to determine the center of dense clusters and the number of clusters based on the central decision graph, where K is determined by the first K high density points. Then, the sample points of the non-cluster center are divided into the cluster where the nearest peak density sample is located. Finally, the clustering of the sample data is completed. Therefore, the number of clustering can be regarded as the initialization parameter for the EM algorithm, and the optimal iteration parameters of the EM algorithm can be regarded as initialization parameters for the GMSC method. Two algorithms are brought together to increase the convergence speed of the EM algorithm, thereby improving the accuracy of the GMSC method and reduce the time complexity of our model.

4.3 GMSC method

The GMSC method can be expressed by the following pseudo-code in Figure 3:

5.0 Experimental evaluation

5.1 Dataset description and evaluation metrics

We collected user-generated content from Sina Microblog, a Chinese social media site that is like a hybrid of Twitter and Facebook. Characterized by weak information (lack of in-

Algorithm1. GMSC method
Input: Training sentiment samples $\boldsymbol{D_{train}} = \{\boldsymbol{T_1}, \boldsymbol{T_2}, \cdots \boldsymbol{T_n}\}$
Testing sentiment samples $\boldsymbol{D_{test}} = \{\boldsymbol{T_1^*}, \boldsymbol{T_2^*}, \cdots \boldsymbol{T_m^*}\}$
Output: Classification results of fine-grained sentiment $\boldsymbol{R_{test}} = \{\boldsymbol{R_1^*}, \boldsymbol{R_2^*}, \cdots \boldsymbol{R_m^*}\}$
1.$T^* = \{EV_1, EV_2, \cdots, EV_N\}$ //Sentiment sequence
2. $D'_{train} = \mathrm{PCA}(D_{train})$ //Dimension reduction
3.$D'_{test} = PCA(D_{test})$
2.begin initialize $\theta^0, Th, t \leftarrow 0$ //Initialization
3.Model_inital = $\mathrm{DPC}(D'_{train})$ //DPC initialize model
4.do t ⟵ t + 1
5.$\mathrm{J}(\theta, \theta') = \theta(D'_{train}, \mathrm{Model\_inital})$ // E-step of EM algorithm
6.$\theta^{t+1} \leftarrow \arg max \mathrm{J}(\theta, \theta')$ // M-step of EM algorithm
7.until $\mathrm{J}(\theta^{t+1}, \theta') - \mathrm{J}(\theta^{t}, \theta^{t-1}) \leq Th$ //Iteration ends
8.return $\hat{\theta} \leftarrow \theta^{t+1}$ //$\theta = \{\alpha_i, \mu_i, \sum_i\}$
9.end
10.Model = $\mathrm{GMM\_train}(D'_{train}, \theta)$ // GMM construction
11.for i = 1 to m //Testing samples
12.R = $\mathrm{Classify}(D'_{test}, \mathrm{Model})$ // Classify
13.Output(R_{test}) //Output results
14.end for

Figure 3. Pseudo-code of GMSC methodology.

formation) and strong sentiment, it has become an important channel for the public to express their emotions, attitudes, and ideas in the context of web 2.0 as a We Media (A carrier people use to post about events). Widely used in China since August 2009, Sina Microblog grew to 411 million active users a month as of March 2018 according to the company's Q1 earning report in 2018. It can provide a variety of data for multi-class sentiment classification.

As the study requires a detailed classification of sentiment, it is necessary to choose a dataset with a variety of sentiments, so we made use of all of the published Sina Microblog content concerning the "Japan 311 earthquake" (more than forty thousand samples in total). We removed invalid comments that would interfere with the data analysis, such as retweeted comments. In other words, we use non-repeatable microblogs as research samples. Afterward, 4,000 samples of eight categories of sentiment were selected randomly, including 2,000 training samples and 2,000 testing samples. Table 2 offers a brief summary of the sentiment dataset, while Table 3 includes examples of microblogs.

A supervised sentiment classification methodology (such as naïve Bayes) needs labeled samples, and the number of labeled training samples decides the methodology's classification efficiency, which may decrease its generalization ability. Therefore, we choose semi-supervised and unsupervised methodologies to compare to the proposed GMSC methodology. Other studies (Zhang and He 2013; Dai et al. 2012; Wan 2008; Lin, He and Everson 2010) have used different datasets to compare the methodologies. We use the same training and test samples from Sina Microblog, the biggest social network in China, to compare the experimental results.

To compare the performance of the proposed methodology, manually labeling is needed, we, therefore, asked five college students to label multi-class sentiment types of Sina Microblog samples, and the labeled results are based on the majority judgment. We trained them to identify multi-class sentiment beforehand.

Prior research (Inoue and Narihisa 2000; Ishibuchi and Nii 1998; Leng and Wang 2008) used accuracy as an index to measure the generalization ability of method. We used recall, precision, F-measure, and specific sample discrimination conditions to measure generalization ability comprehensively. These indexes are defined as in the general sentiment classification to evaluate performance. The evaluation metrics are described in reference (Patil and Sherekar 2013); we, therefore, omitted here.

5.2 Experimental settings

Data normalization processing for the obtained dataset includes: first, extracting comments from dataset samples (we collected data using a web crawler); second, NLPIR segmentation and denoising. After the data normalization process, standardized samples are stored as .txt files. We selected Lin Hongfei's (Xu, Lin and Zhao 2008) sentiment ontology library as the experimental lexicon to calculate the feature value of a sentiment. The sentiment lexicon was used to calculate the semantic orientation values of the microblogs; then the vector space representation model and feature calculation formula are used to construct the feature matrix of each sample in the form of a vector. The lexicon divides sentiment into seven categories and twenty-one subcategories as shown in Table 4.

We should note here that the recall of Lin Hongfei's ontology, popular in China and proposed by Dalian University of Technology (Xu, Kin and Zhao 2008), is about 64.3%. It identifies several major sentiment types: happy, good, anger, sad, fear, disgust, and surprise, which is also consistent with the Chinese tradition of seven emotions and six sensory pleasures. As many published academic pa-

Category	No	Happy	Good	Angry	Sad	Fear	Disgust	Surprise
Quantity	100	300	250	280	320	300	310	140

Table 2. The summary of dataset (the type of multi-class sentiment is based on Lin Hongfei's ontology lexicon).

Time	Microblog Text
2011-04-28 08:56	其实我知道会有日本地震、其实我知道今年 8 月富士山会喷发岩浆。其实我知道的、所以我才感到害怕。

Table 3. Information of microblog text (translation: "Actually, I know there will be an earthquake in Japan. In fact, I know Mt. Fuji will erupt in August this year. Actually, I know, so I am scared.")

Num	Category	Subcategory	NumMm	Category	Subcategory	Num	Category	Subcategory
1	Happy	(PA)	8	Angry	(NA)	15		(NG)
2		(PE)	9	Sad	(NB)	16	Disgust	(NE)
3	Good	(PD)	10		(NJ)	17		(ND)
4		(PH)	11		(NH)	18		(NN)
5		(PG)	12		(PF)	19		(NK)
6		(PB)	13	Fear	(NI)	20		(NL)
7		(PK)	14		(NC)	21	Surprise	(PC)

Table 4. Sentiment category of Lin Hongfei (Xu, Lin and Zhao 2008) sentiment ontology library.

pers have used Lin Hongfei's ontology lexicon in China, it holds a certain authority in the Chinese academic field.

The lexicon and Eqs. (1) and (2) are used to obtain the feature value and construct the feature matrix of a sentiment. The coefficient setting in the experiment is $\alpha = 0.9, \beta = 0.1$. The result is the 4000×21 multi-class emotion feature matrix and the 3900×21 polarity emotion feature matrix. In this paper, the feature matrix of sentiment $EV_{n\,m}$ is a sparse matrix; thus, we chose the dimension reduction method of principal component analysis (PCA) to reduce the error of a sparse matrix on experimental accuracy. It involves using less comprehensive variables to replace the original variables. The reduction results largely determine the performance of the subsequent algorithm, which is the prerequisite to construct a classification model (Hogenboom et al. 2015; Zhou, He and Wang 2008). The new feature matrix is formed to construct an unsupervised sentiment classification model.

We randomly selected half of the samples in each category as training samples and the remaining half became test samples. In the following experiment, we repeated the data selection process ten times. The following results are the average of ten classifications for different methodologies.

By using the DPC algorithm to obtain the clustering number for the EM algorithm, the GMM model is initialized by K to obtain an optimal parameter solution of $\alpha_k, \mu_k, \sum_k$ as a means to construct a model of the Sina Microblog training samples:

$$\begin{aligned} p(x|\theta) = {} & \alpha_1 G_1(x|\mu_1, \sum_1) + \alpha_2 G_2(x|\mu_2, \sum_2) + \\ & \alpha_3 G_3(x|\mu_3, \sum_3) + \alpha_4 G_4(x|\mu_4, \sum_4) + \\ & \alpha_5 G_5(x|\mu_5, \sum_5) + \alpha_6 G_6(x|\mu_6, \sum_6 + \\ & \alpha_7 G_7(x|\mu_7, \sum_7) + \alpha_8 G_8(x|\mu_8, \sum_8) \end{aligned} \quad (27)$$

$$p(x|\theta) = \alpha_1 G_1(x|\mu_1, \sum_1) + \alpha_2 G_2(x|\mu_2, \sum_2) \quad (28)$$

Formula (27) above is the GMSC model of multi-class sentiments, and formula (28) is the GMSC model of polarity sentiments. We used the training samples to obtain the initial parameters of $\alpha_k', \mu_k', \sum_k'$ using the above formula (27) and (28) without labeled samples. Because an unsupervised DPC clustering methodology can provide an initial multi-class sentiment classification of the training samples. After that, $\alpha_k', \mu_k', \sum_k'$ are considered to be the input to the test classification model. By bringing the testing feature matrix D_{test} into eight components of $\alpha_i G_i(x|\mu_i, \sum_i)$ respectively, microblog texts belong to the category with the highest value of $\alpha_i G_i(x|\mu_i, \sum_i)$. This process completes the unsupervised multi-class sentiment classification and outputs GMSC results.

5.3 Results and discussion

Experimental verification mainly includes the following four parts:

1) Effect of the cumulative contribution rate on the GMSC approach. For accurate results in terms of dimension reduction, different cumulative contribution rates are selected to obtain an F-measure of classification with a different number of samples, which is shown in Figure 4. $G(r)$ represents accordance with the original features. A larger G(r) value means more features can be selected with a higher fitting value. In Figure 4, with G(r) increasing to 99%, the F-measure achieves the highest point with 2,000 samples. In sum, the following experiments select G(r)=99% for dimension reduction of the feature matrix.

2) Gradual optimization process of the GMSC approach. To verify the gradual optimization of the GMSC approach is effective with multi-class classification, the precision and recall results before and after optimization are shown in Figure 5 and 6, respectively. Table 5 summarizes the overall results.

As seen in Figures 5 and 6, the recall of a non-sentiment class is generally high while precision is relatively low. The reason may be attributed to misclassifying microblog text as a non-sentiment class. In other words, the methodology is failing to identify a certain sentiment. The initialization method is the only difference between the GMSC (DPC initializes the EM parameter) and K-means+GMM (K-means initialize the EM parameter) methodology. We can see that the GMSC approach performs better than K-means+GMM.

Precision, recall, and F-measure average values of the GMSC methodology have been improved in comparison to the previous method as shown in Table 5. The K-means algorithm performs simple clustering by inputting certain K values without iteration processing to obtain an optimal parameter, which leads to the infective results, while the GMSC approach combines DPC and EM to obtain optimal parameters and construct a different model based on historical sentiment samples for higher precision. The re-

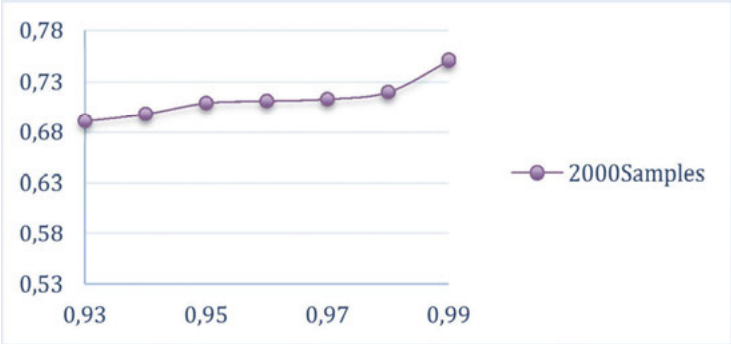

Figure 4. Effect of G(r) on F-measure.

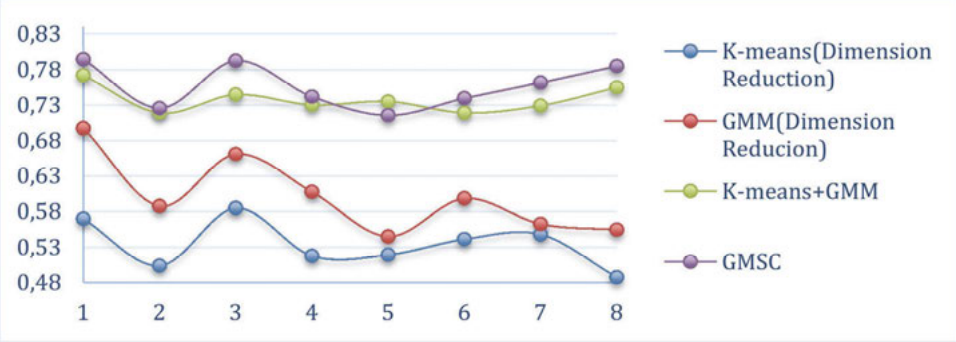

Figure 5. Recall of different methods (number represents sentiment type).

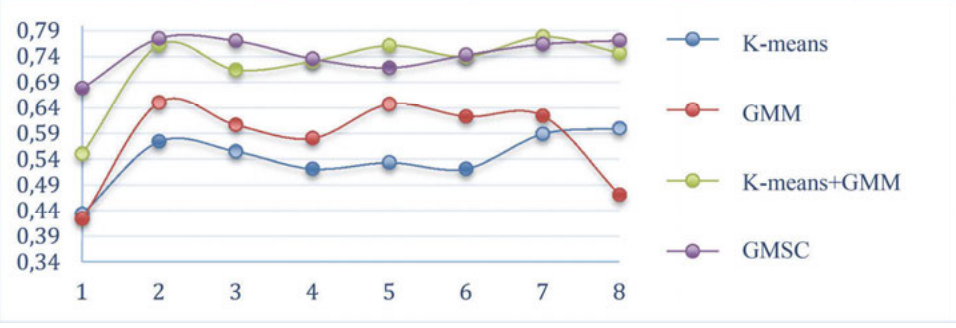

Figure 6. Precision of different methods (number represents sentiment type).

sults suggest the proposed methodology is better than the original methodology. Combining DPC and GMM resulted in higher precision, recall, and F-measure values.

3) Generalization ability of sentiment classification method. Figure 7 offers a conclusion and Table 6 shows the multi-class sentiment classification results for each categories as follows:

As shown in Figure 7, three evaluation indexes exhibit the same trend, except for a difference in Wan's (2008) methodology wherein recall is significantly higher than precision. Fitting three parameters show that the residual parameters have a coincide rate with each other. Therefore, the ability of the method to classify sentiments as positive and negative does not make a difference.

As seen in Table 6, multi-class sentiment classification results are extracted for successive comparison with one other, which can comprehensively prove the high generalization ability of the GMSC approach. The semi-supervised methodology of Zhang and He (2013) has a higher classification accuracy of non-emotion than other methodologies, which can be attributed to the partially labeled samples of semi-supervised methods. Generally, the semi-supervised methodology has a higher classification precision than unsupervised, except for GMSC. We think the fitting model of training data influenced the classification accuracy of the unsupervised methodologies of Dai et al. (2012), Wan (2008), and Lin, He and Everson (2010). Overall, we can conclude that our approach is effective. The generalization ability of the GMSC approach has been improved more so than other methods.

Method	Recall	Precision	F-measure
K-means(dimension reduction)	53.4%	54.1%	53.8%
GMM(dimension reduction)	60.2%	57.8%	59%
K-means+GMM	73.8%	72.3%	73%
GMSC	76%	76.4%	76.2%

Table 5. Comparison experiment results.
Note: Average results of ten times.

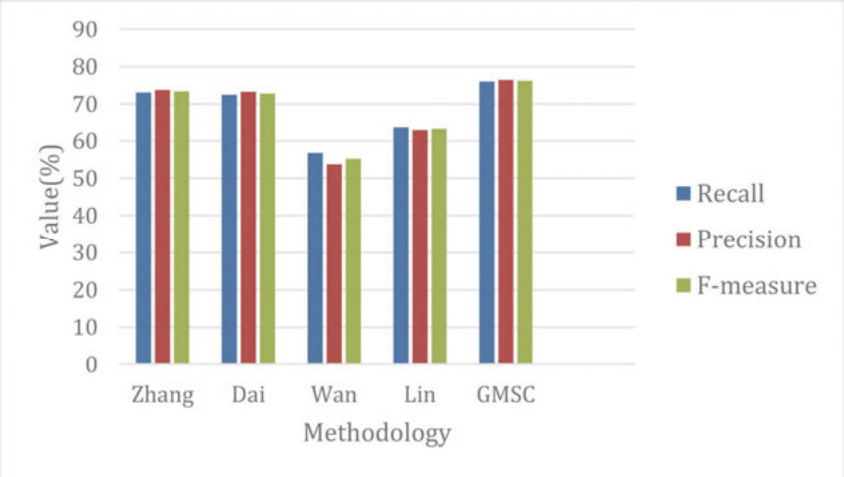

Figure 7. Summary of different methods.

	No	Happy	Good	Anger	Sad	Fear	Disgust	Surprise
Zhang	82.5%	71.4%	72.8%	74.3%	73.1%	71.0%	73.3%	72.0%
Dai	64.8%	72.1%	73.5%	74.3%	73.5%	75.1%	75.3%	76.3%
Wan	55.3%	49.8%	45.3%	57.2%	58.7%	53.9%	54.9%	54.5%
Lin	52.9%	67.2%	63.3%	65.4%	61.3%	65.7%	63.9%	63.5%
GMSC	68.2%	78.9%	78.4%	74.9%	74.9%	77.4%	78.6%	79.7%

Table 6. Precision comparison of different method in detail.

4) Time complexity of the comparison method. To verify whether the GMSC approach has a relatively low time complexity, Table 7 compares the advantages and disadvantages, time complexity, and other characteristics of each sentiment classification method.

Table 7 shows Zhang and He's (2013) method has the highest time complexity, and its precision is higher than GMSC. Wan's (2008) method has the lowest time complexity with the lowest accuracy. Dai et al. (2012) combines a maximum entropy algorithm and coordination training to enlarge the number of training samples; although the time complexity is not high, the generalization ability for unknown data is lower. Lin, He and Everson's (2010) method takes the number of topics m and sentiments n into account; it holds that when m and n are large enough, the time complexity is higher than that of the GMSC approach. Overall, the time complexity of GMSC is low when compared to an unsupervised method with high accuracy.

Evaluation index recall, precision, and F-measure values have improved when compared to semi-supervised (Zhang and He's method) and unsupervised methods (Wan's methods, Dai et al.'s method, Lin, He and Everson's method). The generalization ability of GMSC methodology is improved and time complexity is reduced.

By analyzing the results, we find that the feature matrix is extracted to solve the rough tagging of sentiments. The extracted feature matrix can fully reflect the sentiment and lay a foundation for a high accuracy model construction. The PCA method is used to solve the computation problem of a sparse matrix. Classification model construction based on each sentiment categories of microblogs can realize high fitting degree of samples. The GMSC method performs better with unknown microblog text and achieves high accuracy of multi-class sentiment classification, because each sentiment is represented by a Gaussian Process, which can distinguish the difference between sentiments more easily. The model parameters are initialized by an unsupervised clustering DPC algorithm and optimized by EM gradual iteration, thereby leading to a lower time complexity of the proposed GMSC methodology. Meanwhile, the generative approach focuses on the different probability distributions of each class; therefore, the GMSC methodology has higher extensibility and sufficiency of multi-class sentiment, especially in the context of big data research.

6.0 Conclusion

To deal with the limitations of current research, we propose a novel generative and unsupervised classification methodology of multi-class sentiment based on GMM called GMSC. The experiments on a sentiment dataset illustrate its effectiveness. By computing the probability distribution of different types of sentiment with GMM, user-generated content can be divided into distinct Gaussian components, which further realize effective and accurate multi-class sentiment classification. This methodology takes advantage of a generative approach that does not need any parameters set-up nor a labeled training dataset. The sentiment value can be obtained by the probability distribution characteristics of samples, which further addresses the limitations of the existing approaches and improves accuracy.

Our research has several theoretical implications. First, the DPC algorithm initializes the GMM algorithm, so the model parameters can be obtained according to different historical sentiment samples of user-generated content, which are divided into different Gaussian processes, thus enhancing the generalization ability of each category of sentiment. Second, the EM algorithm is combined with the GMM algorithm to obtain the optimal parameter for the GMSC methodology through progressive iteration, update, and self-adaptation, so the GMSC approach has low time complexity and high sentiment classification accuracy. And our research divides user-generated content into different Gaussian processes to obtain high multi-class sentiment classification accuracy, which addresses the limitation of having only positive and negative sentiment

	Zhang	**Dai**	**Wan**	**Lin**	**GMSC**
Tanging	Yes	No	No	No	No
Advantages	Precision High	Precision High	Simple Operation	Topic and Sentiment Classification	Generalization ability High Relative low time complexity
Disadvantages	Generalization ability low	Generalization ability Low	Precision Low	Time complexity High	Global optimal acquisition difficulty
Time Complexity	O (N^3)	O(NP|A|)	O(N)	O(log(m*n)*N)	O(N^2)

Table 7. Comprehensive comparison of methods (N denotes the number of samples, P is iteration, and A is the size of the event set).

classifications. Third, we analyze the proposed unsupervised sentiment classification methodology from a generative perspective, which can provide the mechanism to illustrate why the GMSC methodology is completely unsupervised and suits our research problem.

Although the GMSC approach solves the limitation of existing semi-supervised and unsupervised methods, it does have some drawbacks: The computation complexity of the model is increased as the dataset becomes larger, thereby resulting in convergence rate instability. As the proposed methodology only considers the characteristics of the microblog text without taking into account the features of the user who posted it.

Future studies have numerous avenues to explore. One would be to further optimize the EM algorithm to obtain the global optimal parameter solution and improve the stability of its convergence rate and then further improve the accuracy of the proposed method. Second, according to the characteristics of the sentiment samples, more research about relevant features with in-depth data, dataset analysis from different cultural backgrounds and sources, and mining relationships between users posting comments are needed, so as to construct a graph model to improve sentiment classification by utilizing the relationship between comments.

References

Agarwal, Apoorv, Boyi Xie, Vovsha Ilia, Owen Rambow, and Rebecca Passonneau. 2011. "Sentiment Analysis of Twitter data." In *Proceedings of the Workshop on Languages in Social Media, June 23, 2011*. Stroudsburg, PA*:* Association for Computational Linguistics, 30-8.

Albrechtsen, Hanne, and Annelise M. Pejtersen. 2003. "Cognitive Work Analysis and Work Centered Design of Classification Schemes." *Knowledge Organization* 30: 213-27.

Boiy, Erik and Marie-Francine Moens. 2009. "A Machine Learning Approach to Sentiment Analysis in Multilingual Web Texts." *Information Retrieval* 12: 526-58.

Bernardo, J. M., M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, 2007. "Generative or Discriminative? Getting the Best of Both Worlds." *Bayesian Statistics* 8: 3-24.

Café, Lígia Maria Arruda and Renato Rocha Souza, 2017. "Sentiment Analysis and Knowledge Organization: An Overview of the International Literature." *Knowledge Organization* 44: 199-214.

Cambria, Erik, Björn Schuller, Yunqing Xia, and Catherine Havasi. 2013. "New Avenues in Opinion Mining and Sentiment Analysis." *IEEE Intelligent Systems* 28: 15-21.

Dai, Daming, Zhongqing Wang, Shoushan Li, Peifeng Li, and Qiaoming Zhu. 2012. "Unsupervised Chinese Sentiment Classification with Emotion Words." *Journal of Chinese Information Processing* 26, no. 4: 103-9.

Dasgupta, Sajib, and Vincent Ng. 2009. "Topic-wise, Sentiment-wise, or Otherwise?: Identifying the Hidden Dimension for Unsupervised Text Classification." In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, August 6-7, 2009*. Stroudsburg, PA: Association for Computational Linguistics, 580-9.

Fang, Fang, Kaushik Dutta, and Anindya Datta. 2014. "Domain Adaptation for Sentiment Classification in Light of Multiple Sources." *INFORMS Journal on Computing* 26: 586-98.

Fernández-Gavilanes, Milagros, Jonathan Juncal-Martínez, Silvia García-Méndez, Enrique Costa-Montenegro, and Francisco J. González-Castaño. 2018. "Creating Emoji Lexica from Unsupervised Sentiment Analysis of Their Descriptions." *Expert Systems with Applications* 103: 74-91.

Ghose, Anindya and Panagiotis Ipeirotis. 2009. "The EconoMining Project at NYU: Studying the Economic Value of User-generated Content on the Internet." *Journal of Revenue and Pricing Management* 8: 241-6.

Goh, Khim-Yong, Cheng-Suang Heng, and Zhijie Lin. 2013. "Social Media Brand Community and Consumer Behavior: Quantifying the Relative Impact of User-and Marketer-generated Content." *Information Systems Research* 24: 88-107.

He, Yulan, and Deyu Zhou. 2011. "Self-training from Labelled Features for Sentiment Analysis." *Information Processing & Management* 47: 606-16.

Hogenboom, Alexander, Bas Heerschop, Flavius Frasincar, Uzay Kaymak, and Franciska d. Jong. 2014. "Multi-lingual Support for Lexicon-based Sentiment Analysis Guided by Semantics." *Decision Support Systems* 62: 43-53.

Hogenboom, Alexander, Flavius Frasincar, Franciska d. Jong, and Uzay Kaymak. 2015. "Polarity Classification Using Structure-based Vector Representations of Text." *Decision Support Systems* 74: 46-56.

Inoue, Hirotaka, and Hiroyuki Narihisa. 2000. "Improving Generalization Ability of Self-generating Neural Networks through Ensemble Averaging." In *Knowledge Discovery and Data Mining. Current Issues and New Applications. 4th Pacific-Asia Conference, PAKDD 2000 Kyoto, Japan, April 18–20, 2000 Proceedings, ed. Takao Terano, Huan Liu, and Arbee L. P. Chen*. Lecture Notes in Computer Science 1805. Berlin: Springer, 177-80. doi:10.1007/3-540-45571-X_22

Ishibuchi, H. and M. Nii. 1998. "Fuzzification of Input Vectors for Improving the Generalization Ability of Neural Networks." In *The 1998 IEEE International Conference on Fuzzy Systems Proceedings: IEEE World Congress on Computational Intelligence May 4-May 9,1998 Anchorage, Alaska, USA*. Piscataway, NJ: IEEE Service Center, 1153-8. doi:10.1109/FUZZY.1998.686281

Jiang, Long, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. "Target-dependent Twitter Sentiment Classification." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, June 19-24, 2011*. Stroudsburg, PA: Association for Computational Linguistics, 151-60.

Jing, Liping, Michael K. Ng, and Joshua Z. Huang. 2007. "An Entropy Weighting K-means Algorithm for Subspace Clustering of High-dimensional Sparse Data." *IEEE Transactions on Knowledge and Data Engineering* 19: 1026-41.

Jing, Liping, Michael K. Ng, Jun Xu, and Joshua Z. Huang. 2005. "Subspace Clustering of Text Documents with Feature Weighting K-means Algorithm." In *Advances in Knowledge Discovery and Data Mining: 9th Pacific-Asia Conference, PAKDD 2005, Hanoi, Vietnam, May 18-20, 2005 Proceedings*, ed. Tu Bao Ho, David Cheung, and Huan Liu. Lecture Notes in Computer Science 3518. Berlin: Springer, 802-12.

Kennedy, Alistair and Diana Inkpen. 2006. "Sentiment Classification of Movie Reviews Using Contextual Valence Shifters." *Computational Intelligence* 22: 110-25.

Kim, Hyunsoo, Peg Howland, and Haesun Park. 2005. "Dimension Reduction in Text Classification with Support Vector Machines." *Journal of Machine Learning Research* 6: 37-53.

Leng, XueMing and YiDing Wang. 2008. "Improving Generalization for Gender Classification." In *2008 15th IEEE International Conference on Image Processing, October 12-15, 2008. San Diego.* Piscataway, NJ: IEEE Service Center, 1656-9. doi:10.1109/ICIP.2008.4712090

Li, Shoushan and Churen Huang. 2010. "Chinese Sentiment Classification based on Stacking Combination Method." *Journal of Chinese Information Processing* 24: 56-61.

Li, Shoushan, Zhongqing Wang, Guodong Zhou, and d Sophia Y.M. Lee. 2011. "Semi-supervised Learning for Imbalanced Sentiment Classification." *Journal of the Royal Statistical Society* 22: 1826.

Li, Shoushan, ChuRen Huang, Guodong Zhou, and Sophia Y.M. Lee. 2010. "Employing Personal/Impersonal Views in Supervised and Semi-supervised Sentiment Classification." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, July 11- 16, 2010*. Stroudsburg, PA*:* Association for Computational Linguistics, 414-43.

Lin, Chenghua, Yulan He, and Richard Everson. 2010. "A Comparative Study of Bayesian Models for Unsupervised Sentiment Detection." In *Proceedings of the Fourteenth Conference on Computational Natural Language Learning, July 15-16, 2009*. Stroudsburg, PA: Association for Computational Linguistics, 144-52.

Liu, Yang, Jianwu Bi, and Zhiping Fan. 2017a. "Multi-class Sentiment Classification: The Experimental Comparisons of Feature Selection and Machine Learning Algorithms." *Expert Systems with Applications* 80: 323-39.

Liu, Yang, Jianwu Bi, and Zhiping Fan. 2017b. "A Method for Multi-class Sentiment Classification based on An Improved One-vs-one (OVO) Strategy and the Support Vector Machine (SVM) Algorithm." *Information Sciences* 394: 38-52.

Mai, Jens-Erik. 2004a. "Classification in Context: Relativity, Reality, and Representation." *Knowledge Organization* 31: 39-48.

Mai, Jens-Erik. 2004b. "Classification of the Web: Challenges and Inquiries." *Knowledge Organization* 31: 92-7.

Melville, Prem, Wojciech Gryc, and Richard D. Lawrence. 2009. "Sentiment Analysis of Blogs by Combining Lexical Knowledge with Text Classification." In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, June 28 - July 1, 2009*. New York: ACM, 1275-84.

Ouyang, Chunping, Xiaohua Yang, Longyan Lei, Qiang Xu, Ying Yu, and Zhiming Liu. 2014. "Multi-strategy Approach for Fine-grained Sentiment Analysis of Chinese Microblog." *Acta Scientiarum Naturalium Universitatis Pekinensis* 1: 67-72.

Pang, Bo and Lillian Lee. 2008. "Opinion Mining and Sentiment Analysis." *Foundations and Trends in Information Retrieval* 2, no. 1/2: 1-135. doi:10.1561/1500000011

Pang, Bo and Lillian Lee. 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, July 21-26, 2004*. Stroudsburg, PA: Association for Computational Linguistics, 271.

Peng, Ting-Chun and Chia-Chun Shih. 2010. "An Unsupervised Snippet-Based Sentiment Classification Method for Chinese Unknown Phrases without Using Reference Word Pairs." In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, August 31 - September 3, 2008, Toronto*. Piscataway, NJ: IEEE Service Center, 243-8. doi10.1109/WI-IAT.2010.229

Patil, Tina R. and Mrs. S. S. Sherekar. 2013. "Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification." *International Journal of Computer Science and Applications* 6: 256-61.

Qiao, Shaojie, Kun Jin, Nan Han, Changjie Tang, and Duoji Gesang 2015. "Trajectory Prediction Algorithm based on Gaussian Mixture Model." *Journal of Software* 26: 1048-63.

Rifkin, Ryan and Aldebaro Klautau. 2004 "In Defense of One-vs-all Classification." *Journal of Machine Learning Research* 5: 101-41.

Rodriguez, Alex, and Alessandro Laio. 2014. "Clustering by Fast Search and Find of Density Peaks." *Science* 344: 1492-6.

Roy, K. Dharmendra and Lokesh K. Sharma. 2010. "Genetic K-Means Clustering Algorithm for Mixed Numeric and Categorical Data Sets." *International Journal of Artificial Intelligence & Applications* 1: 23-8.

Sarvabhotla, Kiran, Prasad Pingali, and Vasudeva Varma. 2011. "Sentiment Classification: A Lexical Similarity based Approach for Extracting Subjectivity in Documents." *Information Retrieval* 14: 337-53.

Serrano-Guerrero, Jesus, Jose A. Olivas, Francisco P. Romero, and Enrique Herrera-Viedma. 2015. "Sentiment Analysis: A Review and Comparative Analysis of Web Services." *Information Sciences* 311: 18-38.

Silvaa, Nádia Félix Felipe da, Luiz F.S. Colettaa, Eduardo R. Hruschkaa, and Estevam R. Hruschka Jr. 2016. "Using Unsupervised Information to Improve Semi-supervised Tweet Sentiment Classification." *Information Sciences* 355: 348-65.

Sindhwani, Vikas and Prem Melville. 2008. "Document-word Co-regularization for Semi-supervised Sentiment Analysis." In *Proceedings Eighth IEEE International Conference on Data Mining: ICDM 2008; 15-19 December 2008 Pisa, Italy. Piscataway, NJ: IEEE Service Center,* 1025-30. doi:10.1109/ICDM.2008.113

Smailović, Jasmina, Miha Grčar, Nada Lavrač, and Martin Žnidaršič. 2014. "Stream-based Active Learning for Sentiment Analysis in the Financial Domain." *Information Sciences* 285: 181-203.

Sung, H., Guang. 2004. "Gaussian mixture regression and classification." PhD diss., Rice University.

Suzuki, Jun, Akinori Fujino, and Hideki Isozaki. 2007. "Semi-supervised Structured Output Learning based on A Hybrid Generative and Discriminative Approach." In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, June 28–30, 2007 Prague, Czech Republic*. Stroudsburg, PA: Association for Computational Linguistics, 791-800.

Taboada, Maite, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. "Lexicon-based Methods for Sentiment Analysis." *Computational Linguistics* 37: 267-307.

Tan, Songbo, Yuefen Wang, and Xueqi Cheng. 2008. "Combining Learn-based and Lexicon-based Techniques for Sentiment Detection without Using Labelled Examples." In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 20-24, 2008*. New York: ACM, 743-44.

Tang, Huifeng, Songbo Tan, and Xueqi Cheng. 2007. "Research on Sentiment Classification of Chinese Reviews based on Supervised Machine Learning Techniques." *Journal of Chinese Information Processing* 21: 88-94.

Turney, D., Peter. 2002. "Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews." *In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, July 7-12, 2002*. Stroudsburg, PA*:* Association for Computational Linguistics, 417-44.

Wan, Xiaojun. 2011. "Bilingual Co-training for Sentiment Classification of Chinese Product Reviews." *Computational Linguistics* 37: 587-616.

Wan, Xiaojun. 2008. "Using Bilingual Knowledge and Ensemble Techniques for Unsupervised Chinese Sentiment Analysis." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, October 25-27, 2008*. Stroudsburg, PA*:* Association for Computational Linguistics, 553-61.

Wang, Gang, Jianshan Sun, Jian Ma, Kaiquan Xu, and Jibao Gu. 2014. "Sentiment Classification: The Contribution of Ensemble Learning." *Decision Support Systems* 57: 77-93.

Wang, Hao, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. "A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle." In *Proceedings of the ACL 2012 System Demonstrations, July 10, 2012*. Stroudsburg, PA*:* Association for Computational Linguistics, 115-120.

Xu, Jun, Yuxin Ding, and Xiaolong Wang. 2007. "Sentiment Classification for Chinese News Using Machine Learning Methods." *Journal of Chinese Information Processing* 21: 95-100.

Xu, Linhong, Hongfei Lin, Jing Zhao. 2008. "Construction and Analysis of Emotional Corpus." *Journal of Chinese Information Processing* 22: 116-22.

Zagibalov, Taras, and John Carroll. 2008. "Automatic Seed Word Selection for Unsupervised Sentiment Classification of Chinese Text." In *Proceedings of the 22nd International Conference on Computational Linguistics, August 18-22, 2011*. Stroudsburg, PA*:* Association for Computational Linguistics, 1073-80.

Zhai, Zhongwu, Hua Xu, and Peifa Jia. 2010. "An Empirical Study of Unsupervised Sentiment Classification of Chinese Reviews." *Tsinghua Science & Technology* 15: 702-8.

Zhang, Pu, and He Zhongshi. 2013. "A Weakly Supervised Approach to Chinese Sentiment Classification Using Partitioned Self-training." *Journal of Information Science* 39: 815-31.

Zhang, Wenhao, Hua Xu, and Wei Wan. 2012. "Weakness Finder: Find Product Weakness from Chinese Reviews by Using Aspects based Sentiment Analysis." *Expert Systems with Applications* 39: 10283-91.

Zhou, Lizhu, Yukai He, and Jianyong Wang. 2008. "Survey on Research of Sentiment Analysis." *Journal of Computer Applications* 28: 2725-8.

Zhou, Shusen, Qingcai Chen, and Xiaolong Wang. 2010. "Active Deep Networks for Semi-Supervised Sentiment Classification." *In Proceedings of the 23rd International Conference on Computational Linguistics, August 23-27, 2010.* Stroudsburg, PA*:* Association for Computational Linguistics, 1515-23.

Appendix

The process of EM algorithm:

The optimal parameter can be obtained through gradual iteration; furthermore, the maximum $\mathrm{P(X|\theta)}$ can be derived as the following:

$$\theta^* = \arg\max P(X|\theta) \tag{17}$$

To facilitate solving θ^*, $\arg\max \log \mathrm{P(X|\theta)}$ is used to substitute $\mathrm{P(X|\theta)}$. As to difficulty in directly solving $\mathrm{P(X|\theta)}$, the indirect parameter solution after deformation of Eq. (7) can be derived as:

$$\mathrm{J}(\theta, \theta') = \sum_{i=1}^{K} p(x, i|\theta) \log p(x, i|\theta') \tag{18}$$

Where $\theta' = \{\alpha_i', \ \mu_i', \ \sum_i'\}$ represents another set of parameters in the model construction process, $p(x, i|\theta)$ denotes the probability density of sentiment sample x belonging to sentiment category i on the condition of parameter θ. The following is the calculation result using Eq. (16):

$$\mathrm{J}(\theta, \theta') - \mathrm{J}(\theta, \theta) = \sum_{i=1}^{k} p(x, i|\theta)\{\log p(x, i|\theta') - \log p(x, i|\theta)\} = \sum_{i=1}^{k} p(x, i|\theta) log \frac{p(x, i|\theta')}{p(x, i|\theta)} \tag{19}$$

Function $\mathrm{f(x)} = \log \mathrm{x}$, tangent equation is $\varphi(\mathrm{x}) = \mathrm{x} - 1$ at point $(\mathrm{x}, \mathrm{f(x)})|_{x=1}$, and $\mathrm{f(x)} \le \varphi(\mathrm{x})$, according to the formula above, we can further obtain the following equation:

$$\mathrm{J}(\theta, \theta') - \mathrm{J}(\theta, \theta) \le \sum_{i=1}^{k} p(x, i|\theta) \left[\frac{p(x, i|\theta')}{p(x, i|\theta)} - 1\right] = \sum_{i=1}^{k} \{p(x, i|\theta') - p(x, i|\theta)\} \tag{20}$$

That is:

$$\mathrm{J}(\theta, \theta') - \mathrm{J}(\theta, \theta) \le \mathrm{p}(\mathrm{x}, \theta') - \mathrm{p}(\mathrm{x}, \theta) \tag{21}$$

By analyzing Eq. (19), we can see that $\mathrm{J}(\theta, \theta')$ and $\mathrm{p}(\mathrm{x}, \theta')$ have the same monotonicity; therefore, the differential of $\mathrm{p}(\mathrm{x}, \theta)$ at θ is as follows:

$$\nabla_\theta \mathrm{p}(\mathrm{x}, \theta) = \nabla_\theta \sum_{i=1}^{K} \left(\nabla_\theta p(x, i|\theta)\right) = \sum_{i=1}^{k} p(x, i|\theta)(\nabla_\theta \log p(x, \theta)) \tag{22}$$

Eqs. (11) and (15) are combined:

$$\nabla_\theta p(x, \theta) = \nabla_\theta J(\theta, \theta')|_{\theta=\theta'} \tag{23}$$

When $\theta = \theta'$, $\mathrm{J}(\theta, \theta')$ and $\mathrm{p}(\mathrm{x}, \theta)$ will reach to the extreme value at θ simultaneously. Combined with Eq. (19), the two parameters not only have the same monotonicity but also the same extreme point. So we can obtain the new value of θ' by maximum $\mathrm{J}(\theta, \theta')$ through iterative convergence. Eq. (9) has been taken into Eq. (18) to derive the formula:

$$\begin{cases} \mathrm{J}(\theta, \theta') = \sum_{n=1}^{N} \sum_{i=1}^{K} \emptyset_n(i) log \alpha_i' G'(x_n|\mu_{x,i}', \sum_{x,i}') \\ \emptyset_n(i) = p(x_n, i|\theta) = p(x_n|\theta) p(i|x_n, \theta) \end{cases} \tag{24}$$

As to Eq. (22), when the partial derivative equals zero, the corresponding value of estimation parameter θ' of parameter solution process can be divided into two steps: $\mathrm{E-step}$ (Expectation Calculation) and $\mathrm{M-step}$ (Maximum Calculation) as follows:

1) $\mathrm{E-step}$: Probability of sentiment sample x_i belonging to multi-class sentiment i can be calculated as:

$$\mathrm{p(i}|x_n, \theta) = \frac{\alpha_i p(x_n|i, \theta)}{p(x_n|\theta)} = \frac{\alpha_i G(x_n|\mu_{x,i}, \sum_{x,i})}{\sum_{k=1}^{K} \alpha_k G(x_n|\mu_{x,k}, \sum_{x,k})} \tag{25}$$

2) $\mathrm{M-step}$: Iteration formula of GMM parameters is obtained with the expectation maximization algorithm:

$$\begin{cases} \alpha_i' = \frac{1}{T} \sum_{n=1}^{N} p(i|x_n, \theta) \\ \mu_i' = \frac{\sum_{n=1}^{N} p(i|x_n, \theta) x_n}{\sum_{n=1}^{N} p(i|x_n, \theta)} \\ \sum_i' = \frac{\sum_{n=1}^{N} p(i|x_n, \theta) x_n^2}{\sum_{n=1}^{N} p(i|x_n, \theta)} - \mu_i'^2 \end{cases} \tag{26}$$

The weight is estimated for GMM when the parameters under the unknown condition in $\mathrm{E-step}$. $\mathrm{M-step}$ optimizes and determines the parameters of GMM based on the estimated weight by $\mathrm{E-step}$. The two steps will be terminated until there are small fluctuations and an approximate extreme value.