Sensors and Materials, Vol. 34, No. 6 (2022) 2263–2280
MYU Tokyo

S & M 2971

Emotional Feature Extraction from Texts by Support Vector Machine with Local Multiple Kernel Learning

Kai-Xu Han,[1,2†] Shu-Fang Yuan,[1†] Wei Chien,[3*] and Cheng-Fu Yang[4,5**]

[1]School of Electric and Information Engineering, Beibu Gulf University, Qinzhou 535011, China
[2]School of Computer, Electronics and Information, Guangxi University, Nanning 530004, China
[3]School of Information Management and Master's Program, Lunghwa University of Science and Technology, Taoyuan City 333, Taiwan
[4]Department of Chemical and Materials Engineering, National University of Kaohsiung, Kaohsiung 811, Taiwan
[5]Department of Aeronautical Engineering, Chaoyang University of Technology, Taichung 413, Taiwan

(Received December 30, 2021; accepted April 6, 2022)

Keywords: multiple kernel learning (MKL), local multiple kernel learning (LMKL), support vector machine (SVM), emotional text

Emotional analysis in texts is one of the difficult problems in text feature extraction. Semantic information is not unique for a large number of text features, which increases the difference in feature weight. Previous studies had dimensional disasters, loss of feature information, and a weak generalization ability in text feature extraction. To solve these problems, we first analyzed the advantages of support vector machines (SVMs) by multiple kernel learning (MKL). Then, an algorithm with local multiple kernel learning (LMKL) was proposed for a threshold model to select the locally optimal kernel function. It helped understand which text feature distinguishes emotions more effectively. Next, we analyzed the features of the local multiple kernel learning algorithm and discussed its generalization ability. The effectiveness of the method in this study was verified through comparison with other methods. The method reduced the feature dimensionality of the sample data set. Since the features with a weak classification ability were reduced, the accuracy of the classification was improved with increased efficiency.

1. Introduction

The widespread use of voice sensors on mobile terminal devices has brought about the popularization of voice recognition functions. The information carrier after voice recognition is used as the main text, which contains a large amount of emotional information. Such emotional information considerably enhances the application value of text emotional analysis research. Emotional analysis has significantly attracted the interest of researchers to grasp the state of society and the dynamics of events. The analysis also helps decision makers of governments, enterprises, and institutions.(1–3) The emotions on the Internet are expressed mainly with texts. Thus, the analysis of emotions in texts is the key factor in the research on emotional analysis on the Internet.

*Corresponding author: e-mail: air180@seed.net.tw
**Corresponding author: e-mail: cfyang@nuk.edu.tw
†These authors are co-first authors of this article.
https://doi.org/10.18494/SAM3833

The emotional analysis in texts is one way of text classification that needs a support vector machine (SVM) and an artificial neural network (ANN). A deep neural network (DNN), as an artificial neural network, has shown significant improvement with improved computing power. However, DNN has drawbacks such as high requirements for calculating the power of computers and the ambiguous effect of small-scale data sets. This makes SVM applicable for small data sets with low computing power.[(4)] Thus, SVM becomes effective for text classification with multiple kernel learning (MKL) that effectively improves the accuracy of text classification and identification.

It is difficult to extract emotional features from texts. A large number of texts hinder semantic information from being found and create a large difference in weight for important texts. Previous research studies on text classification had dimensionality disasters, loss of feature information, and a weak generalization ability. These problems led to reduced classification efficiency and increased calculation complexity, reducing the accuracy of the emotional analysis of texts.

In this study, we propose a new feature selection method based on local multiple kernel learning (LMKL), focusing on solving the problems of classifiers and text extraction. We applied the method of emotional analysis in texts to improve the text extraction for emotional analysis more effectively. The method is expected to increase the efficiency of text classification, which removes the dimensional disasters in texts and improves the efficiency.

1.1 SVM in text classification

SVM is a learning system that uses a hypothesis space with linear functions in a high-dimensional space. SVM has been widely used for image processing and pattern identification. It has many advantages in solving problems of small numbers of samples and nonlinear and high-dimensional pattern identification. Its application extended to other machine learning problems such as function fitting. An SVM method combines dimensionality reduction and classification. An SVM algorithm is based on the principle of structural risk minimization. It compresses original data sets to support vector sets and then uses the subset to learn new knowledge. At the same time, it also gives the rules determined by these support vectors and the upper boundary of the probability of learning error.

Shen and Yu[(5)] and Ding *et al.*[(6)] studied the active learning strategy of SVM and applied it to text classification. Wang *et al.* proposed to train SVM with a progressive iterative algorithm.[(7)] Mayy studied the effects of various text models on SVM text classification.[(8)] Cristianini *et al.* proposed an SVM text classifier based on the semantic core.[(9)] They implemented a latent semantic indexing in the feature space defined by the core. Uğuz[(10)] and Li *et al.*[(11)] studied how to add the prior knowledge of texts in the learning process of SVM. In addition, in-depth research studies were also conducted on various deficiencies of SVM in text classification.[(12–14)]

In brief, the advantages of SVM are as follows. The classification is performed by using the support vector in a hyperplane space. The features of the previous classification are changed according to the statistics of all data, which considerably reduces the number of calculations for the classification.

(1) SVM is not affected by the dimensionality of the data.
(2) SVM is efficient in solving nonlinear and high-dimensional pattern identification for a small number of samples.
(3) The accuracy and recall rate of the classification are very high in the text classification.

On the other hand, its disadvantages are as follows.

(1) The number of calculations in the classification process increases because of too many inputs, increasing training time.
(2) It is difficult to determine the dimension of the feature vector with SVM.

1.2 MKL SVM

SVM with MKL has advantages over single-core SVM.[15] Lanckriet *et al.* proposed MKL.[16] They provided a powerful asymptotic algorithm for direct reduction for multiple kernel learning models to avoid locally optimal solutions. However, the algorithm was only suitable for problems with small numbers of samples. Subsequently, Sonnenburg and Ratsch proposed a new dual form to solve medium-scale problems.[17] In recent years, many research results suggested alternative optimization methods to solve problems with large numbers of samples.[18] The results revealed that multiple kernel functions have higher accuracy and recognition of classification than single functions. The simplest way for higher accuracy is using the sum of kernel functions without weights. However, it is only ideal when the unweighted sum provides the same preference for all the kernel functions. A better strategy is obtaining the sum of equal weights of convex combinations, which makes it possible to extract the information from the weight of the kernel function. Rakotomamonjy *et al.* transformed the problem of multi-core learning into a semi-definite planning one to find the combined weights and support vector coefficients at the same time.[19] For this, Xu *et al.* proposed an efficient algorithm using sequential minimal optimization (SMO).[20]

On the basis of the above research results, we propose an LMKL algorithm that uses a threshold model to select the locally optimal kernel function. This algorithm determines the emotional texts to be effectively distinguished. In this study based on MKL SVM, we suggest a multiple kernel framework for the emotional text selection. Then, the method of emotional analysis in texts is proposed to select the emotional words to verify the effectiveness of the analysis.

2. Feature Selection Method

2.1 MKL

Kernel-based methods such as SVM are used by many researchers owing to their efficiency in processing data. The basic idea of the classification is training SVM with the data from the input space to a linearly separable feature space (usually a space with a higher dimension than the input space). The classification function of SVM after training is defined as

$$f(x) = w^T \Phi(x) + b, \tag{1}$$

where w is the weight coefficient, b is the critical value, and $\Phi(x)$ is the mapping function to the relevant feature space. Here, the mapping function does not need to be clearly defined. With w substituted, the following discriminant is obtained as

$$f(x)=\sum_{i=1}^{n}\alpha_i y_i \Phi(x)^T \Phi(x_i)+b\,. \tag{2}$$

In Eq. (2), $K(x,x_i)=\Phi(x)^T\Phi(x_i)$ is the related core. Each $\Phi(x)$ has its own features and is related to a different kernel function, such that a different discriminant function is generated in the original space. It is important for SVM training to choose a kernel or mapping function, and SVM training usually requires cross-validation. Although the kernel function has the weighted sum, the discriminant functions in different feature spaces are regarded as the unweighted sum of discriminant values.

$$f(x)=\sum_{i=1}^{n} w_m^T \Phi_m(x)+b\,, \tag{3}$$

Here, m is the index of the kernel function, w_m is the weight coefficient of the feature mapping function, $\Phi_m(x)$ is the mapping function of the first feature space, and p is the number of kernel functions. Substituting w_m in Eq. (3) leads to the following equation:

$$f(x)=\sum_{m=1}^{p}\eta_m\sum_{i=1}^{n}\alpha_i y_i \Phi(x)^T \Phi(x_i)+b. \tag{4}$$

The weight of the kernel function meets $\eta_m \geq 0$ and $\sum_{m=1}^{p}\eta_m = 1$. Here, to combine linear, Gaussian, and polynomial kernel functions with different hyperparameters together, different data representations or feature subsets are also combined together.

For this, a fixed joint rule (without weight) assigns the same weight to the kernel function in the entire input space. Different weights on kernel functions in different regions of the input space enable better classification. If the data has potential locations, then the appropriate kernel function that matches the complexity of the data distribution in each local area is given a higher weight. The localized form of the MKL problem modifies the discriminant function in the MKL framework and optimizes the parameters by a two-step optimization method. Then, the method introduces the key attributes of the proposed algorithm. Finally, the LMKL method is applied to select the features for the emotional analysis of texts.

2.2 Feature selection based on LMKL

The probabilistic latent semantic analysis (PLSA) model is a statistical model based on the aspect model, which is a latent variable model serving the data of the co-occurrence matrix. Each hidden variable, called z_k, is associated with each observation. Here, each observation refers to the appearance of a term in a document. We can derive and solve the PLSA model to obtain the topical features of the text.

To combine the kernels, the classifier function is defined as

$$f(x) = \sum_{m=1}^{p} \eta_m(x) w_m^T \Phi_m(x) + b \,. \tag{5}$$

Here, $\eta_m(x)$ is for the selection of the feature space m as the threshold function of input x and defines a set of parameters from the data. By the new discriminant function that modifies the original discriminant function of the SVM, the following constrained optimization problem is obtained:

$$\begin{aligned}
&\min \frac{1}{2}\sum_{m=1}^{p}\left\|w_m\right\| + C^2\sum_{i=1}^{n}\varepsilon_i, \\
&\text{w.r.t. } w_m, b, \varepsilon, \eta_m(x), \\
&\text{s.t. } y_i\left(\sum_{m=1}^{p}\eta_m(x_i) w_m^T \Phi_m(x_i) + b\right) \ge 1 - \varepsilon_i \forall i, \\
&\varepsilon_i \ge 0 \forall i.
\end{aligned} \tag{6}$$

Here, C is a normalized parameter and ε is a slack variable. Owing to the introduction of interval restriction in Eq. (6), the optimization problem is nonconvex. It is very difficult to directly solve Eq. (6). Thus, we use a two-step optimization algorithm to find the parameters and $\eta_m(x)$ discriminant function. The first step considers w_m and b while fixing $\eta_m(x)$. The second step uses a gradient descent method with the objective function of Eq. (6). A fixed objective value from $\eta_m(x)$ is the upper bound of Eq. (6), and the parameters of $\eta_m(x)$ are updated according to the current solution. The use of the descent gradient limits the target value from the next iteration not to be greater than the current value and the target value of Eq. (6) not to be increased. This does not guarantee the convergence to the global optimum, and the initialization parameters affect the quality of the solution. With fixed $\eta_m(x)$, the Lagrangian operator in Eq. (6) is obtained as

$$L_D = \frac{1}{2}\sum_{m=1}^{p}\left\|w_m\right\|^2 + \sum_{i=1}^{n}(C - \alpha_i - \beta_i)\varepsilon_i + \sum_{i=1}^{n}\alpha_i - \sum_{i=1}^{n}\alpha_i y_i (\sum_{m=1}^{p}\eta_m(x_i) w_m^T \Phi_m(x_i) + b) \,. \tag{7}$$

With the initialized variables, the partial derivative of L_D is defined as

$$\begin{aligned}
&\frac{dL_D}{dw_m} \Rightarrow w_m = \sum_{i=1}^{n}\alpha_i y_i \eta_m(x_i) \Phi_m(x_i) \forall m, \\
&\frac{dL_D}{db} \Rightarrow \sum_{i=1}^{n}\alpha_i\, y_i = 0, \\
&\frac{dL_D}{d\varepsilon_i} \Rightarrow C = \alpha_i + \beta_i \forall i.
\end{aligned} \tag{8}$$

According to Eqs. (6)–(8), the dual planning problems are obtained as

$$\begin{aligned}&\max\sum_{i=1}^{n}\alpha_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j K_\eta\left(x_i,x_j\right),\\&\text{w.r.t.}\alpha,\\&\text{s.t.}\sum_{i=1}^{n}\alpha_i y_i=0,\\&C\geq\alpha_i\geq 0\forall i.\end{aligned}\tag{9}$$

Here, the locally combined kernel matrix is defined as

$$K_\eta\left(x_i,x_j\right)=\sum_{m=1}^{p}\eta_m\left(x_i\right)\Phi_m\left(x_i\right),\Phi_m\left(x_j\right)\eta_m\left(x_j\right)\cdot\tag{10}$$

Equation (10) is equivalent to a positive semi-definite kernel matrix $K_\eta(x_i, x_j)$ that solves the dual problem of SVM. According to a quasi-positive transformation, a semi-positive and definite kernel matrix is obtained by the multiplication of the outputs from the nonnegative function by the kernel function. Therefore, the local joint kernel matrix is obtained from the quasi-conformal transformation of each kernel function. The summation of the matrices constructs a joint kernel matrix. Here, the only restriction is to use nonnegative $\eta_m(x)$ to obtain a positive semi-definite kernel matrix. One of the possible kernel functions is regarded as a classification problem with the assumption that the region using the kernel is linearly separable. In this case, the threshold model is expressed as

$$\eta_m\left(x\right)=\frac{\exp\left(v_m^T x+v_{m0}\right)}{\sum_{k=1}^{p}\exp\left(v_k^T x+v_{k0}\right)},\tag{11}$$

where v_m and v_{m0} are nonnegative. In the space defined by the basic functions instead of the original input space, the parameters of the threshold model are nonnegative. At the same time, more complex threshold models or equivalent thresholds can be used. Then, the threshold model can be a combination of several different $\Phi_m(x)$ values. If a continuous threshold function is used instead of the function of x, the algorithm can be used to find a fixed combination in the entire input space and is similar to the original MKL.

The method in this study does not extract the subsets from the training set, but trains the classifier of each subset, and then combines the trained classifiers. The difference from the previous method is that LMKL uses the combination of subset selection and local classifier in a combinatorial optimization problem in this study. LMKL in this study is similar to an expert model. However, it is a combination of a threshold model and an expert model based on a kernel function and joint learning. In addition, each expert becomes a classifier in an expert model, whereas all the kernel functions in LMKL share a discriminant in this study.

With a given $\eta_m(x)$, the strong duality makes the target value of Eq. (8) equal that of Eq. (6). Equation (8) is used for a function $J(\eta)$ to calculate the original objective with the parameters of $\eta_m(x)$. To train the threshold model, the following gradient descent is used:

$$\frac{\partial J(\eta)}{\partial v_{m0}} = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{p}\alpha_i\alpha_j y_i y_j \eta_k(x_i) K_k(x_i, x_j)\eta_k(x_j)(\delta_m^k - \eta_m(x_i) + \delta_m^k - \eta_m(x_j)), \tag{12}$$

$$\frac{\partial J(\eta)}{\partial Vm} = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{p}\alpha_i\alpha_j y_i y_j \eta_k(x_i) K_k(x_i, x_j)\eta_k(x_j)\left(x_i\left[\delta_m^k - \eta_m(x_i)\right] + x_j\left[\delta_m^k - \eta_m(x_j)\right]\right). \tag{13}$$

Here, if $m = k$ or 0, then $\delta_m^k = 1$. After updating the parameters of $\eta_m(x)$, it is necessary to use $K_\eta(x_i, x_j)$ to solve a single-core SVM problem.

The convergence of the algorithm of complete LMKL is obtained by observing the changes in α, the parameter of $\eta_m(x)$. The algorithm for the linear threshold model is summarized as follows.

(1) Initializing v_m and v_{m0} to small random values, $m = 1, 2, ..., p$.
(2) Repeating the following steps.
(3) Calculating $K_\eta(x_i, x_j)$ with a threshold model.
(4) Solving standard SVM problems using $K_\eta(x_i, x_j)$.
(5) $v_{m0}^{(t+1)} \Leftarrow v_{m0}^{(t)} - \mu(t)\dfrac{\partial J(\eta)}{\partial v_{m0}}$.
(6) $v_m^{(t+1)} \Leftarrow v_m^{(t)} - \mu(t)\dfrac{\partial J(\eta)}{\partial v_m}$.
(7) Repeating the above steps until convergence stops.

The final discriminant function from the SVM solution is

$$f(x) = \sum_{i=1}^{n}\sum_{m=1}^{p}\alpha_i y_i \eta_m(x) K_m(x, x_i)\eta_m(x_i) + b. \tag{14}$$

2.3 Discussion of generalization

For machine learning, the generalization ability always needs to be considered and is defined as the ability to classify the outliers of the training set correctly.

In each iteration, the joint kernel by the current threshold model is used to solve the standard SVM problem and calculate the gradient of $J(\eta)$. Compared with the SVM solution process, the calculation of gradient descent has negligible time complexity. The step size $u(t)$ of each iteration should be determined by a linear search that requires additional SVM optimization for better convergence. The complexity of the LMKL algorithm is mainly determined by how complex the standard calculation process for SVM is in the main loop. The complexity is reduced by α as an input. The number of iterations for the convergence depends on the selection of training data and step size. The time complexity of the test is also reduced owing to the localization of the algorithm. Here, $K_m(x, x_i)$ only needs to be estimated if $\eta_m(x)$ and $\eta_m(x_i)$ are nonzero at the same time.

LMKL is applied to kernel-based algorithms apart from the binary classification of SVMs, such as the regression and single-class classifications of SVM. Two basic changes are required here, namely, (1) the optimization problem and (2) calculation of the target value.

If the kernel function is estimated in different feature subsets or data representatives, then the important kernel function will have a higher combination weight. An LMKL framework derives similar information from different input spaces. Owing to the nonlinearity by the threshold model, the method proposed in this study allows the multiple use of the same kernel function to obtain a local discriminant. For example, a linear kernel model and a threshold model are combined to obtain a segment-wise linear boundary.

LMKL essentially localizes SVM. According to the data distribution, as a feature space is divided into multiple regions, SVM classifies the features in different regions. In this case, LMKL has two iteration processes: (1) dividing the feature space and (2) training the SVM in the divided region using the features in the regions. Overall, the interface of LMKL becomes more complex and classifies highly complex nonlinear data. The following describes the generalization ability of LMKL:

(1) When the kernel functions selected by LMKL and SVM are the same in each divided region, LMKL and SVM have the same generalization ability and upper limit of the generalization error rate.
(2) When LMKL has one divided region and the same selected kernel functions of SVM, LMKL and SVM are completely the same.
(3) A large number of regions divided by LMKL reduce the generalization ability of LMKL. The generalization error rate of the classifier is inversely proportional to the size of training samples and the numbers of functions. That is, when the number of training samples increases, the generalization error rate decreases. When LMKL has many regions, the size of training samples of each classifier decreases, increasing the generalization error rate. On the other hand, the generalization ability of the classifier is inversely proportional to the difference between the empirical and real distributions of the sample. The greater the difference between the empirical and real distributions, the worse the generalization ability of the classifier becomes. Many regions of LMKL cause the number of training samples of each classifier to decrease. The estimated empirical distribution of the samples is different from the real distribution, so the generalization ability becomes worse. The amount of "classification margin" that is increased by the divided regions of LMKL needs to be balanced with the amount of sample reduction caused by the regions.

2.4 Method of feature selection of emotional words

The emotional analysis in this study is based on emotional words. However, emotional classification by the emotional words is not obvious. The emotional words require the complexity of the calculation for the emotional analysis. The complexity impacts the accuracy and efficiency of the emotional analysis. The feature selection of LMKL in this study is proposed to solve this problem. In the emotional analysis of texts, the features of LMKL are emotional words. Selecting emotional words with larger feature thresholds improves the speed of the analysis and reduces the number of features with a weak classification ability. This also leads to the improvement of classification accuracy.

In this paper, the improved LMKL SVM for feature word selection is abbreviated as WS. The following explains the process of the LMKL method proposed or the feature selection. The feature selection method uses the weight of the linear classifier that is obtained with a simple linear decision function as Eq. (15).

$$y = \text{sign}(w^T x + b) \tag{15}$$

Here, w and x have the same dimensions. $|w_d|$ measures the importance of x_d for discrimination. That is, the larger the $|w_d|$, the more important the x_d. x is for the occurrence probability vector of each word, so $x_d = 0.1$ means that the occurrence probability of the dth word in a document is 0.1. Therefore, a larger $|w_d|$ means that the dth word has a large contribution to emotional analysis, whereas $|w_d| = 0$ means that the dth word has no contribution. As long as a threshold is determined, important words are selected from the lexicon. The discriminant function of the mth core with LMKL as a linear kernel function is defined as

$$f(x) = \sum_{m=1}^{p} \eta_m(x) w_m^T x + b, \tag{16}$$

Combining all the cores yields $\sum_m |\eta_m(x) w_{md}|$.

The selected feature words are sent to the classifier for emotional analysis. The specific process is shown in Fig. 1. The screening and filtering of feature words were performed through repeated tests to determine the expressive power and accuracy of text emotions.

3. Results

The emotional analysis is not completed even after WS. The analysis needs the classifier to be completed. The feature WS needs to be combined with the text classifier to finish the emotional analysis.

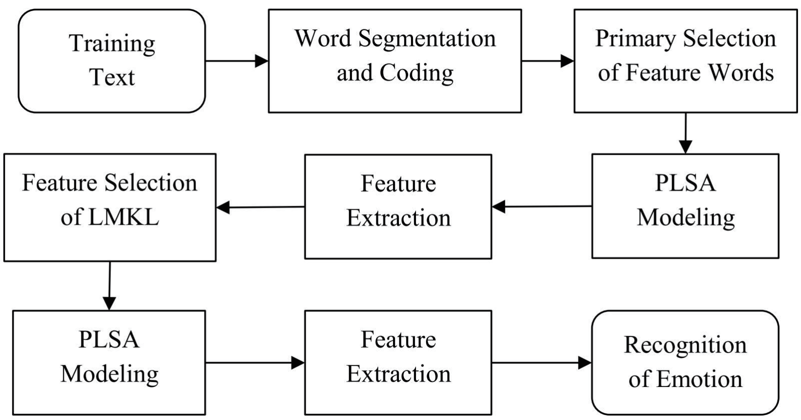


Fig. 1. Flow chart of text emotional analysis based on WS method.

Theoretically, the reduction in the number of weak classification features improves the classification accuracy of the classifier. The emotional analysis with WS is based on two considerations: the rigor of the experiment and the selected comparison method. We took three different methods of text emotional analysis based on SVM and combined them with WS to compare the effects of the combination on the analysis. The three methods are HIST-SVM, PLSA-SVM, and FK-SVM.[4]

(1) HIST-SVM uses the number of occurrences/frequency of words (statistical histograms) as text features in text recognition and SVM for the classification and recognition of the words.
(2) PLSA-SVM calculates the text topic vector based on PLSA as a feature of the text and uses SVM for classification and recognition.
(3) FK-SVM derives the Fisher kernel similarity function based on PLSA as the kernel function of SVM for classification and recognition.

Combining the feature selection method with the above three methods, we obtained new methods such as (WS)HIST-SVM, (WS)PLSA-SVM, and (WS)FK-SVM. The effectiveness of the WS was verified by the effects of the new methods on emotional analysis.

3.1 Experimental data set and result evaluation criteria

The evaluation criteria of the experimental data set and results are the same as those in the literature as follows.[4]

(1) The corpus used here is extracted from Stanford University's "Sentiment140" (The data set can be downloaded from the following link: http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip).[21]
(2) The two indexes of "recall rate" and "accuracy" are involved in all the retrieval and selection processes involving large-scale data collection. Owing to the two mutually constrained indicators, it is usually necessary to select a suitable degree for the "search strategy" according to needs. Moreover, it should not be very strict or lenient, and a balance point should be achieved between the recall rate and the accuracy. This balance point is determined by specific needs.

Assumption: The documents can be divided into the following four groups when retrieving documents from a large data set:

1. documents that are relevant and retrieved by the system (TP);
2. documents that are irrelevant and retrieved by the system (FP);
3. documents that are relevant but not retrieved by the system (FN);
4. documents that are irrelevant but not retrieved by the system (TN).

The negative samples in this case do not mean the wrong classification, but the type of sample with the category "negative". Thus, it can be concluded that FN and TN are used to calculate the classifier error rate. Then,

Recall rate R: The number of related documents retrieved is used as the numerator, with the total number of all related documents as the denominator, namely, $Recall = TP / (TP + FN)$.

Accuracy P: The number of related documents retrieved is used as the numerator, with the total number of all retrieved documents as the denominator, namely, $True\ positive\ rate = TP / (TP + FP)$.

3.2 Data preprocessing and feature selection (WS)

Data preprocessing in the experiment is processing the emotional feature words. We used the database of "Twitter" in English that was standardized in the comparison of the methods.

In the beginning, there were tens of thousands of words in the word database, that is, tens of thousands of dimensions. In the preprocess of the experiment, stop words and vocabulary were removed as they are not used to express emotions. There were also a large number of nouns, verbs, and others that have nothing to do with emotions. Many of the words for emotional expression are not used frequently, and some have ambiguous emotional meanings. WS calculated the degree of contribution of each emotional word to emotional judgment. A threshold reduced the dimensions of the features.

3.3 Experiment design

The experiment procedures of the new methods are similar to those of the three methods for comparison. The changes in the three new methods include the feature word filtering link based on LMKL and the filtered features to represent the documents for training. The experiment design of (WS)FK-SVM is shows as below to illustrate the combination process of WS.

Figure 2 shows the process of (WS)FK-SVM.

(1) Preparing the training data set
(2) Data preprocessing
(3) Filtering feature words: WS with LMKL extracts feature words with a greater effect on the emotion classification from the data set.

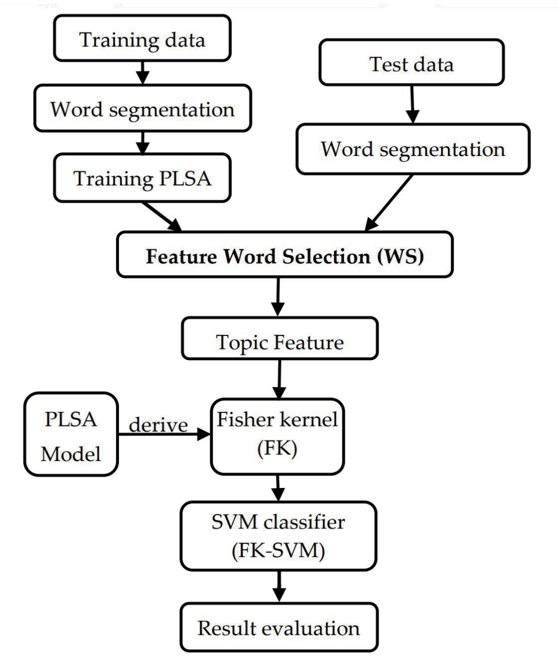


Fig. 2. Experimental process of (WS)FK-SVM emotional analysis method.

(4) Training the text model: The selected feature word set expresses the emotion of the text. The model training is carried out according to the emotion classification that has been manually marked by the corpus. We defined two types of emotional themes as “positive” and “negative” in this study. Since FK-SVM is based on a PLSA model, the emotion classification of the known corpus and the feature words in the training corpus obtain the “vocabulary-document probability”. Then, the emotion vectors of the document were obtained for the “vocabulary-emotional topic probability” and “emotional topic-document probability”.
(5) Training and classifying (WS)FK-SVM classifiers. The emotion vector in the previous step is delivered to (WS)FK-SVM for binary classification.

The experiment designs of (WS)HIST-SVM and (WS)PLSA-SVM are the same as the design idea of (WS)FK-SVM. It only needs to add the WS link with LMKL to the three methods for comparison. Documents were represented by filtered features and trained with the model.

3.4 Experimental results

Experiment 1:

The twitter data set was trained with five epochs of cross-validation and tested with (WS) HIST-SVM, (WS)PLSA-SVM, and (WS)FK-SVM. Then, the recognition accuracy, the recall rate, and the corresponding average value of each round were compared. The accuracy and recall rate were used as the evaluation criteria to verify the effect of the methods with WS.

Table 1 and Fig. 3 present the experimental results. The horizontal axis of each subgraph in Fig. 3 represents the round of the experiment. The vertical axis in Fig. 3(a) represents the accuracy of the experiment, and the vertical axis in Fig. 3(b) represents the recall rate.

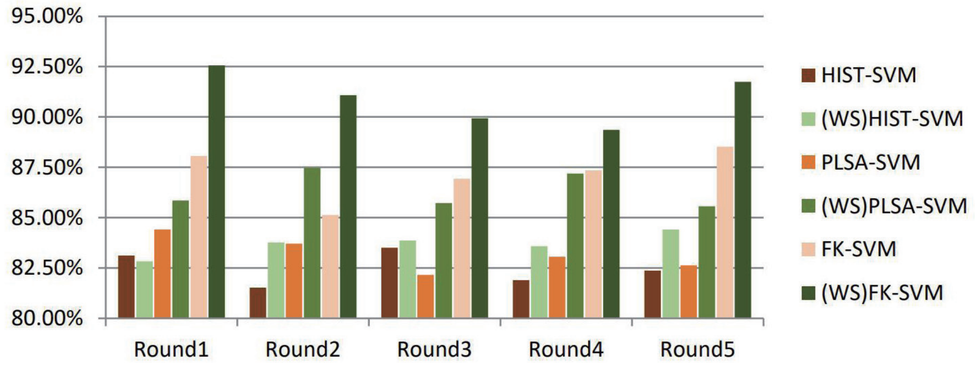
Table 1
Multi-round comparison of experimental results of HIST-SVM, (WS)HIST-SVM, PLSA-SVM, (WS)PLSA-SVM, FK-SVM, and (WS)FK-SVM. (a) Accuracy. (b) Recall rate.

(a) Accuracy

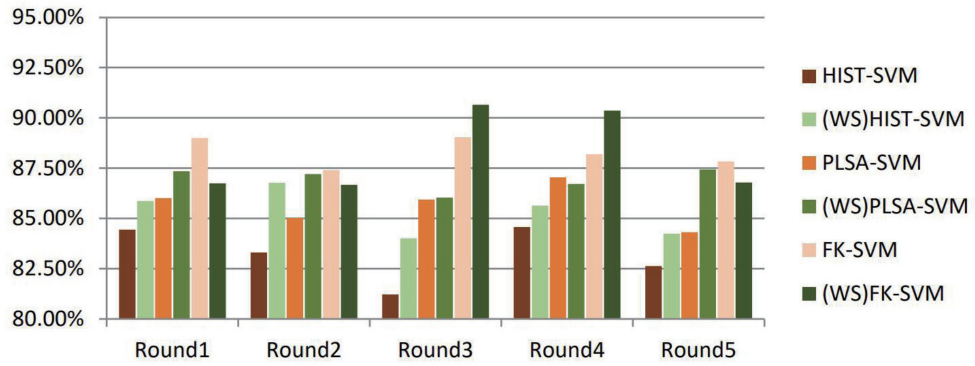
Experiment Round	HIST-SVM (%)	(WS)HIST-SVM (%)	PLSA-SVM (%)	(WS)PLSA-SVM (%)	FK-SVM (%)	(WS)FK-SVM (%)
Round 1	83.12	82.83	84.42	85.85	88.07	92.57
Round 2	81.53	83.77	83.71	87.48	85.13	91.08
Round 3	83.51	83.87	82.16	85.72	86.93	89.93
Round 4	81.89	83.58	83.07	87.19	87.35	89.36
Round 5	82.37	84.41	82.64	85.57	88.52	91.74
Average	82.49	83.69	83.20	86.36	87.20	90.94

(b) Recall rate

Experiment Round	HIST-SVM (%)	(WS)HIST-SVM (%)	PLSA-SVM (%)	(WS)PLSA-SVM (%)	FK-SVM (%)	(WS)FK-SVM (%)
Round 1	84.45	85.87	86.01	87.35	89.00	86.74
Round 2	83.31	86.77	85.03	87.21	87.41	86.67
Round 3	81.22	84.01	85.94	86.04	89.05	90.65
Round 4	84.58	85.64	87.05	86.72	88.19	90.37
Round 5	82.63	84.25	84.31	87.43	87.83	86.79
Average	83.24	85.31	85.67	86.95	88.30	88.24



(a)



(b)

Fig. 3. (Color online) Comparison of experimental results of HIST-SVM, (WS)HIST-SVM, PLSA-SVM, (WS) PLSA-SVM, FK-SVM, and (WS)FK-SVM. (a) Accuracy. (b) Recall rate.

Experiment 2:

With the same twitter data set, five rounds of training and testing for different percentages of sample data in the data set for each experiment were tested for (WS)HIST-SVM, (WS)PLSA-SVM, and (WS)FK-SVM.

The accuracy and recall rate from each round were compared. The result showed that the training sample ratio affected the experimental results. Table 2 and Fig. 4 show the experimental results in the form of tables and bar graphs, respectively. The horizontal axis of each subgraph in Fig. 4 represents the percentage of training samples in each experiment. The vertical axis of Fig. 4(a) represents the accuracy of the experiment, and the vertical axis of Fig. 4(b) represents the recall rate.

4. Discussion

4.1 Analysis of Experiment 1

Experiments using the method proposed in Sect. 3 show the classification performance in Table 1 and Fig. 3. (WS)FK-SVM has the highest recognition accuracy and recall rate, which are

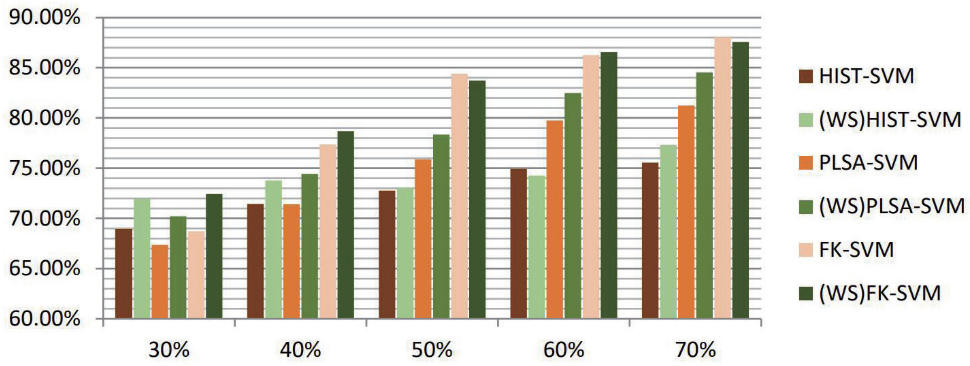
Table 2
Percentage sample comparison of experimental results of HIST-SVM, (WS)HIST-SVM, PLSA-SVM, (WS)PLSA-SVM, FK-SVM, and (WS)FK-SVM. (a) Accuracy. (b) Recall rate.

(a) Accuracy

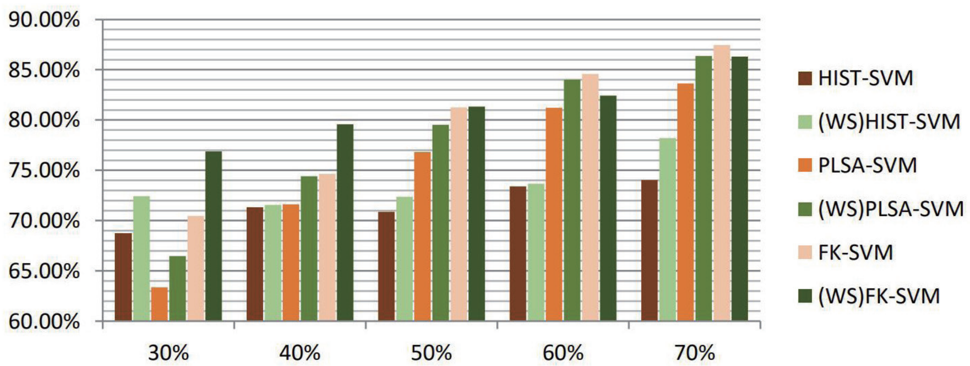
Experiment Round	HIST-SVM (%)	(WS)HIST-SVM(%)	PLSA-SVM (%)	(WS)PLSA-SVM (%)	FK-SVM (%)	(WS)FK-SVM (%)
30%	68.93	71.93	67.37	70.22	68.71	72.42
40%	71.46	73.78	71.43	74.45	77.37	78.68
50%	72.77	73.02	75.86	78.35	84.42	83.73
60%	74.92	74.27	79.75	82.47	86.25	86.56
70%	75.56	77.31	81.24	84.53	88.04	87.58
Average	72.73	74.06	75.13	78.00	80.96	81.79

(b) Recall rate

Experiment Round	HIST-SVM (%)	(WS)HIST-SVM (%)	PLSA-SVM (%)	(WS)PLSA-SVM (%)	FK-SVM (%)	(WS)FK-SVM (%)
30%	68.72	72.43	63.38	66.47	70.47	76.87
40%	71.33	71.57	71.61	74.41	74.63	79.59
50%	70.86	72.36	76.83	79.53	81.25	81.33
60%	73.39	73.67	81.22	84.02	84.59	82.43
70%	74.03	78.21	83.63	86.37	87.45	86.32
Average	71.67	73.65	75.33	78.16	79.68	81.31



(a)



(b)

Fig. 4. (Color online) Comparison of experimental results of (WS)HIST-SVM, HIST-SVM, (WS)PLSA-SVM, PLSA-SVM, (WS)FK-SVM, and FK-SVM for different percentages of training samples. (a) Accuracy. (b) Recall rate.

92.57 and 90.65%, respectively. The average accuracy and recall rate of (WS)FK-SVM are 90.94 and 88.24%, respectively.

The experimental result shows that the accuracy and recall rate of the three new methods, namely, (WS)HIST-SVM, (WS)PLSA-SVM, and (WS)FK-SVM, are not significantly different. Thus, the overall results are relatively stable, indicating that the three methods have good results in text emotional analysis.

The comparisons between (WS)HIST-SVM and HIST-SVM, (WS)PLSA-SVM and PLSA-SVM, and (WS)FK-SVM and FK-SVM show that the new methods have higher accuracies and recall rates even with the reduced number of features than the original methods. The difference between the new and original methods is the set of feature words used to express emotions. The difference is not only in quantity but also in the quality of sentiment words. Thus, the method of WS has a decisive effect on the experiments. The overall result of the new methods is more stable, because the new methods use the LMKL framework of SVM. According to the threshold in the discriminant function, the extracted feature words distinguish the samples more effectively. That is, the new methods reduce the interference of classification with a weak distinguishing ability and the amount of calculation by the participation of inefficient feature words.

4.2 Analysis of Experiment 2

Experiments using the method proposed in Sect. 3 show the classification performance in Table 2 and Fig. 4. The most effective method is (WS)FK-SVM. The highest recognition accuracy and recall rate are 87.58 and 86.32%, respectively, as shown in the experimental results in Table 2. The average accuracy and recall rate of (WS)FK-SVM are 81.79 and 81.31%, respectively.

It can be seen from Experiment 2 that the test results of the three new methods, namely, (WS) HIST-SVM, (WS)PLSA-SVM, and (WS)FK-SVM, are basically stable when facing different percentages of training samples. Moreover, as the percentage of training samples increases, the test results also increase slightly. The reason is that there are fewer training samples. The probability distribution of vocabulary, documents, and emotional topics in the sample is greatly affected by noise. As the sample proportion increases, the corresponding probability distribution gradually approaches the real distribution. We make the thematic mining more accurate, so the effect on the test set will be improved. The frequency and count of words in the sample will be relatively less affected.

From the comparisons of (WS)HIST-SVM and HIST-SVM, (WS)PLSA-SVM and PLSA-SVM, and (WS)FK-SVM and FK-SVM, the three new methods based on the feature selection method proposed in this paper did not reduce the accuracy and recall rate while reducing the number of features, and the accuracy and recall rate in the three new methods were even higher than in the corresponding original method. Since the main difference between the new method and the original method lies in the set of feature words used to express emotions, it can be concluded that the method of selecting feature words has a decisive effect on Experiment 2.

On the basis of the results of the three new methods with WS, while reducing the number of features, the average accuracy and recall rate are slightly higher than those of the original method, and the overall result is more stable. This is the method of selecting feature words, which has a decisive effect on this set of experiments. The feature selection method can use the local multiple kernel learning framework of SVM to understand which feature of the sample can distinguish the sample more effectively according to the threshold in the discriminant function. On the basis of this, the feature words that distinguish the samples more effectively are extracted, thereby reducing the classification interference caused by the feature words with a weak distinguishing ability.

5. Conclusions

In this study, we proposed an algorithm that adopts an LMKL framework for the kernel function learning algorithm, localizes the multiple kernel learning method, and uses the threshold model for selecting the local optimal kernel function. The new methods are used to understand which features are distinguished more effectively in a sample and to analyze the features of the LMKL algorithm. We applied the new method for emotional analysis in texts and verified the effectiveness of the methods through experimental comparison.

From the perspective of feature fusion and dimensionality reduction, the new methods effectively reduce the dimensions of the features in a sample data set. To ensure the expressive ability of the feature words, proper thresholds need to be selected according to the specific conditions to filter the feature words in the actual application process. Since the features with a weak classification ability are removed, the accuracy and efficiency of the classification are improved.

In the next step, we will use voice data in real environments to conduct applied research studies on emotional analysis. As text emotional analysis technology becomes more accurate and efficient, the technology will also become popular in terminal applications, for example, voice sensors.

Acknowledgments

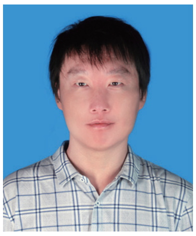
This research was supported by the project of improving the basic scientific research ability of young and middle-aged college teachers in Guangxi "Text sentiment classification based on Fisher kernel function and local multicore learning" (Grant No. 2021KY0434), "Research on vision tracking in complex scene based on deep learning" (Grant No. 2020KY10019), the Research Initiation Project of Introducing High-level Talents from Beibu Gulf University of China (Grant No. 2018KYQD35), and the Guangxi Higher Vocational Education Reform Project (Grant No. GXGZJG2020A001).

References

1 T. K. Tran and T. T. Phan: Appl. Sci. **9** (2019) 2760. https://doi.org/10.3390/app9132760
2 M. Coşkun and M. Ozturan: Information **9** (2018) 102. https://doi.org/10.3390/info9050102

3 Y. L. Wang and H. Y. Youn: Appl. Sci. **9** (2019) 92. https://doi.org/10.3390/app9010092
4 K. X. Han, C. Wei, C. C. Chiu, and Y. T. Cheng: Appl. Sci. **10** (2020) 1125. https://doi.org/10.3390/app10031125
5 X. L. Shen and L. Yu: EURASIP J. Image Video Process. **2013** (2013) 4. https://doi.org/10.1186/1687-5281-2013-4
6 X. Ding, G. Liu, and K. Meng: Comput. Sci. **40** (2013) 86. https://doi.org/10.3969/j.issn.1002-137X.2013.z2.023
7 L. Wang, J. Li, and Q. Yue: Comput. Eng. Appl. **49** (2013) 144. https://doi.org/10.3778/j.issn.1002-8331.1301-0195
8 M. Mayy: J. Intell. Learn. Syst. Appl. **05** (2013) 84. https://doi.org/10.4236/jilsa.2013.52009
9 N. Cristianini, J. Shawe-Taylor, and H. Lodhi: Proc. Int. Conf. Machine Learning (ICML 2001) 66–73.
10 H. Uğuz: Knowledge Based Syst. **24** (2011) 1024. https://doi.org/10.1016/j.knosys.2011.04.014
11 C. H. Li, L. T. Yang, and M. Lin: Int. J. Parallel Program. **42** (2014) 505. https://doi.org/10.1007/s10766-013-0245-x
12 L. Cao and F. E. H. Tay: IEEE Trans. Neural Networks **14** (2003) 1506. https://doi.org/10.1109/TNN.2003.820556
13 F. Ding and L. Sun: Sci. Technol. Eng. **14** (2014) 81. https://doi.org/10.3969/j.issn.1671-1815.2014.03.017
14 M. Gan and X. Fan: Mod. Comput. **6** (2013) 9. https://doi.org/10.3969/j.issn.1007-1423(z).2013.06.002
15 T. Takeshi: J. Global Optim. **54** (2012) 831. https://doi.org/10.1007/s10898-011-9797-8
16 G. R. G. Lanckriet, T. D. Bie, N. Cristianini, and M. I. Jordan: Bioinformatics **20** (2004) 2626. https://doi.org/10.1093/bioinformatics/bth294
17 S. Sonnenburg and G. Ratsch: J. Mach. Learn. Res. **7** (2006) 1531. https://jmlr.csail.mit.edu/papers/volume7/sonnenburg06a/sonnenburg06a.pdf
18 G. Chen, Y. Tang, and Y. Pan: J. Comput. **6** (2011) 2325. https://doi.org/10.4304/jcp.6.11.2325-2334
19 A. Rakotomamonjy, F. Bach, and S. Canu: J. Mach. Learn. Res. **9** (2008) 2491. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.157.95&rep=rep1&type=pdf
20 Z. Xu, R. Jin, H. Yang, I. King, and M. R. Lyu: Proc. Int. Conf. Machine Learning (2010) 1175–1182.
21 Stanford Computer Science: http://cs.stanford.edu/people/alecmgo/trainingandtestdata.zip (accessed April 2022).

About the Authors



Kai-Xu Han received his Ph.D. degree from Northeast Petroleum University, China, in 2015. Since 2017, he has been an assistant professor in the School of Electronic and Information Engineering, Beibu Gulf University. His research area is machine learning. (H.Kaixu@outlook.com)

Shu-Fang Yuan received her M.S. degree from Northeast Petroleum University, China, in 2014. Since 2017, she has been an assistant researcher in Beibu Gulf University. Her research area is machine learning. (frog0696@163.com)

Wei Chien received his B.S.C.E. degree from Ta Tung University, Taipei, Taiwan, in 1997 and his M.S.E.E. degree from Tamkang University, Taipei, in 1999. From 1999 to 2001, he served in the ROC Army Force and was an assistant in the Department of Electrical Engineering. From 2007 to 2014, he was an assistant professor in the Department of Electronic Engineering, De Lin Institute of Technology. From 2014 to 2021, he was a professor in the Department of Computer Engineering, Ningde Normal University.
(air180@seed.net.tw)

Cheng-Fu Yang earned his B.S., M.S., and Ph.D. degrees in 1986, 1988, and 1993, respectively, from the Department of Electrical Engineering of Cheng Kung University. In 2004, he became a professor of chemical and materials engineering at the National University of Kaohsiung (NUK). He received "The Outstanding Contribution Award of the Chinese Ceramic Society" in 2009. In 2010, he was the first (and only) person to become a distinguished professor of NUK. He became a fellow of the Taiwanese Institute of Knowledge Innovation (TIKI) in 2014 and a fellow of the Institution of Engineering and Technology (IET) in 2015. He became a Mingjiang scholar and invited chair professor of Jimei University, Xiamen, Fujian, China, and an honorary chair professor of Chaoyang University of Technology in 2020.
(cfyang@nuk.edu.tw)