Design of Employee Comment Sentiment Analysis Platform Based on AE-SVM Algorithm

Yongxiong Zhang¹ and Liangming Wang²
¹Department of Economics and Trade, GuangZhou College of Technology and Business, Guangzhou, China
²School of Software Engineering, South China University of Technology, Guangzhou, China
Email: csyxzhang@qq.com

Abstract. In order to solve the high-dimensional data set of employees' sentiment analysis on enterprise comment, in this paper, we presented three model algorithms: SVM algorithm, PCA-SVM algorithm, and AE-SVM algorithm. And we made a comparative analysis of three index parameters such as the precision, recall and F1 in three different algorithm model scenarios. According to the experimental results, the AE-SVM algorithm is superior to other two algorithms in this kind of Chinese text sentiment analysis. Based on the AE-SVM Algorithm, we also design a solution of sentiment analysis platform of employee comment. This solution has the characteristics of high availability and high concurrency. It can also collect and analyse the sentiment text information of employees' comment in time.

1. Introduction
In the competition of enterprises, talent is an important strategic resource. The managers of enterprise are under a great pressure of internal and external change [1]. As the changeable and individualized demands of the employees, it will be a new challenge for the human resource management of the enterprise. The competition of enterprises will be more intense under the weak economy, which will cause enterprises to pay more and more attention to the retention and competition of talents. The old "interest-centric" incentive mechanism of the enterprise is no longer applicable. At present, they are gradually transforming to the dual management purpose of "people-oriented and high satisfaction"[2]. Due to the changes in the individual needs of employees, previous economic means is no longer the only way to motivate employees. Factors such as promotion channels, development prospects, and personal growth maybe become considerations for them when choosing a career. Therefore, it is very important for the management of enterprise, if the employee’s comments can be analysed in time and fed back to the managers. How to understand the needs of employees timely and effectively, improve satisfaction and engagement of employee, so as to improve enterprise performance, has become an urgent problem to be solved.

With the development of text sentiment analysis technology in recent years, data mining and analysis can be carried out through the text information of employee’s comment. So that the employees' attitudes towards the company can be found in time. Now there are two kind of text sentiment analysis: dictionary-based methods and machine learning-based method [3, 4].The dictionary-based method is mainly to make a series of emotional dictionaries and rules, then to split words and analyse the text, and finally to judge the emotional tendency of the text according to the emotional value. The precision of this method is high, but the recall is low and the cost of word formation is high [5]. The machine learning-based method is to convert the problem of text sentiment
analysis into a classification problem. Generally, there are two steps: text pre-processing and selection of classification algorithm [6]. The analysis of sentiment word involves high-dimensional data sets. And the SVM algorithm has better processing power and generalization ability in high-dimensional data sets. So in this paper, after analysing and comparing the SVM algorithm, the PCA-SVM algorithm and the AutoEncoder-SVM (AE-SVM for short) algorithm, we chose the best algorithm as the Chinese text classification model to design an employee comment sentiment analysis platform.

The rest of this paper is organized as follows. Section II introduces the concepts and the construction of model for the SVM algorithm, the PCA-SVM algorithm and the AE-SVM algorithm. In the third section, we analyse and compare the performance of these three algorithms. Then, the fourth section introduces architecture design of the platform. Finally, the conclusion is presented and future work is discussed.

2. The Construction of AE-SVM Model

2.1 SVM
SVM (Support Vector Machine) is a kind of supervised learning algorithm, which is used to solve the problem of data classification. The idea of SVM algorithm is to solve the separation hyperplane with the largest geometric interval when the training data set is correctly divided [7, 8].

Separated hyperplane formula:

\[ \omega^T x + \alpha = 0 \]  

(1)

Formula: \( \omega \) is expressed as the variable weight vector, \( \alpha \) is expressed as the offset value, which is the offset of the hyperplane from the origin, \( x \) is expressed as the input vector, which is the vector in the sample set.

The formula of classification interval:

\[ d = \frac{\| \omega^T x + \alpha \|}{\| \omega \|} \]  

(2)

If \( y \) is the classification value, the positive comment is expressed as 1, and the negative comment is expressed as -1. The sample set can be expressed as following:

\[ \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\} \]  

(3)

Formula: \( (x_i, y_i) \) represents a sample point, \( i = 1, 2, \ldots, m \).

The classification discriminant function is expressed as following:

\[ y(x) = \text{sign}(\omega^T x + \alpha) \]  

(4)

In this paper, RBF kernel function is used for high-dimensional projection. The Gaussian radial basis function formula:

\[ k(x_i, x_j) = \exp \left(-\frac{\| x_i - x_j \|^2}{2\sigma^2}\right) \]  

(5)

Formula: \( \sigma \) is expressed as the width of Gaussian kernel and \( \sigma > 0 \). \( (x_i, x_j) \) are expressed as the vector of feature word.

2.2 PCA-SVM
PCA is known as Principal Component Analysis. Its idea is to map m-dimensional features to k-dimensional space, where k<m. This k-dimensional feature is a new orthogonal feature. Since
sentiment analysis is a high-dimensional data set, the technology of PCA is to reduce the high-dimensional data to a low-latITUDE hyperplane through projection and it is expressed by linear change as following [9]:

$$Q = A^T X$$

(6)

Formula: Q represents the low latitude matrix, X represents the high latitude matrix, and A is marked the mapping relationship between them. The implementation steps are as follows:

**Step 1** Decentralize all the X samples.

**Step 2** Calculate the covariance matrix $XX^T$ of the sample.

**Step 3** Calculate the eigenvalue of the covariance matrix $XX^T$.

**Step 4** Sort the eigenvalue, and get the max eigenvector of k eigenvalues to make the matrix A.

**Step 5** Using the matrix multiplication equation as in equation (6), the reduced-dimensional data set Q is obtained.

In the process of sentiment analysis, PCA algorithm is used to pre-process the data, and then SVM is used to analyse. This algorithm can remove the noise in the data and make SVM learning more accurate.

2.3 **AE-SVM**

AutoEncoder known as an automatic encoder, is a neural network that reproduces the input signal as much as possible. AutoEncoder can be regarded as a non-linear patch enhanced version of the PCA algorithm. Its goal is to learn an approximate identity function. The formula as following:

$$h(x) \approx x$$

(7)

AutoEncoder is more flexible than PCA. It can represent both linear and non-linear transformations. PCA can only perform linear transformation. By setting appropriate dimensions and sparse constraints, AutoEncoder is more accurate than PCA [10].

The analysis model of AE-SVM is divided into three stages. The first stage is the preparation of sentiment analysis word sample data; the second stage is the deep learning of Autoencoder features; in the third stage, SVM is used to identify and classify the feature vectors extracted by AutoEncoder. The model of AE-SVM is shown as figure 1.

![Figure 1. AE-SVM model](image-url)

3. **Experimental Results and Analysis**

In this study, some leather enterprises with certain production and a large number of employees, from Shiling leather goods professional towns in Guangdong Province, were selected as the survey objects. The data is from the anonymous reviews by the employees of these enterprises. It is including the text reviews of enterprises for two years from 2018 to 2019. About 20000 pieces of data about these text views were taken as the initial data set of the experiment. Firstly, NLPIR word segmentation tool is used to do Chinese word segmentation. Then the stop words are removed. And the processed data is divided into training set and test set by random partition method, with 80% of training set and 20% of test set.

After the useless information is removed, the value is set: the positive emotion value is 1, and the negative emotion value is -1. In the evaluation sentiment analysis model, the evaluation index is
The precision is expressed as the classification of sample data. The recall is used to identify the sample data. And the F1 value is used to comprehensively evaluate the precision and recall. In this paper, SVM algorithm, PCA-SVM algorithm and AE-SVM algorithm are used to compare and analyse the sample data. The result is shown as Table 1.

| Model algorithm | Precision | Recall | F1  |
|-----------------|-----------|--------|-----|
| SVM             | 0.832     | 0.814  | 0.823 |
| PCA-SVM         | 0.853     | 0.861  | 0.857 |
| AE-SVM          | 0.859     | 0.867  | 0.863 |

After comparing, the precision, recall and F1 value of AE-SVM algorithm are higher than the other two.

4. Design of the System

After determining the AE-SVM algorithm model, in order to solve the problem of automatic analysis, a system platform needs to be designed for data analysis and processing. The employees may focus on a on a certain period of time when making reviews, so high availability and concurrency must be considered in the design of the system platform. In addition, the system platform adopts modular design, so that the algorithm model can be switched according to the needs, reducing the coupling degree of the system. The architecture of system platform is shown in Figure 2.

Figure 2. Architecture of system platform

4.1. Interaction-Tier

Interaction-tier is an interaction that users, including enterprise employees and managers, operate the system platform by smart devices, such as smart phones, computers, etc. The system platform uses corresponding interaction strategies to generate response content based on user identity, emotional text content, etc. And it returns to the intelligent terminal. After receiving the response content, the terminal performs corresponding interactive operations according to the definition of the response. For example, the employees fill in the enterprise comment anonymously, and the manager checks the analysis result.

4.2. Service-Tier

Service-tier needs to consider the business requirements of high concurrency and high availability, as well as the mechanism of interaction with other functional modules, realizing the call to external services. Service-tier uses Keepalived to realize the high availability of the system platform. By the configuration of the master and slave servers, whenever there is a problem with the master server, Keepalived will switch to the slave server to ensure that the service is not interrupted. Service-tier
balances the load of each request by Nginx, and distributes the request to different application servers. The load balance server puts forward the request by the strategy of load statistics distribution. Because it is necessary to consider the better switching performance of the sentiment analysis algorithm, the algorithm analysis module is designed as a service for the application server to call. In a certain period of time, the algorithm analysis service may be called by a large number of application services at the same time. Therefore, the algorithm analysis server is designed as a cluster to improve the execution efficiency of the system.

4.3. Data-Tier
Data-tier keeps structured data and unstructured data. The operation of structured data selects the open source MySQL relational database. And the operation of unstructured data selects the free open source Redis memory database. Redis is a high-performance K-V value database with a rich data structure. It supports data persistence and data synchronization hot backup.

In order to reduce frequent operations on the relational database and improve the efficiency of the system platform, the less changed data in the relational database is pre-loaded into the Redis cache server cluster. If the application server doesn’t obtain the corresponding data in calling Redis cache server cluster, it will query the relational database.

5. Conclusion
In this paper, we use SVM algorithm, PCA-SVM algorithm and AE-SVM algorithm to compare and analyse the enterprise comment information of employees, and conclude that AE-SVM algorithm model is a better model method in this kind of emotional analysis. Meanwhile, based on the AE-SVE algorithm model, a system platform is designed, which has the characteristics of high availability and high concurrency, so that the enterprise can collect the changeable personalized demand information of employees in time. Then the managers can realize effective incentive mechanism by the accurate data.

In next stage, we will complete the test and deployment of the system. And the system platform will be freely used by leather enterprises in Shiling Town, Guangdong Province. Then, the algorithm will be improved when the comment information attach to a certain amount. At the same time, the microservice architecture of the system is reconstructed to make the system more robust.

6. Acknowledgments
This research was financially supported by the Characteristic Innovation Project (Natural Science) of Department of Education of Guangdong Province in 2017 (Subject No.: 2017KTSCX202); Project of philosophy and Social Sciences Planning of Guangzhou in 2020 (Subject No.: 2020GZGJ295).

7. References
[1] Chen C, Peng J and Shi W 2017 Insight of Huaxia cornerstone e, Future has come -- Dialogue between Chen C, Peng J and Shi W J. Chinese and foreign corporate culture 10 30-43.
[2] Lu J, Shi K and Yang F 2001 Evaluation structure and method of job satisfaction J. Human resource development in China 1 15-17.
[3] Feng X, Zhang Z and Shi J 2018 Text sentiment analysis based on convolutional neural networks and attention model J. Application Research of Computers 35 (5) 1434-36.
[4] Liu M, Wang X, Li H and Zhang B 2018 Emotional Analysis of Online Commodity Reviews Based on Text Mining J. Journal of Liaoning University of Technology 38 (5) 1-5.
[5] Wang H, Song Y, Du Z, Zheng L, Hua J and Zhang Y 2017 Evaluation of Service Quality for Express Industry Through Sentiment Analysis of Online Reviews J. Journal of Beijing university of technology 43 (3) 402-12.
[6] Medhat W, Hassan A and Korashy H 2014 Sentiment Analysis Algorithms and Applications: A Survey J. Ain Shams Engineering Journal 5 (4) 1093-113.
[7] Burges C J C 1998 A Tutorial on Support Vector Machines for Pattern Recognition J. Data mining and knowledge discovery 2 121-67.
[8] Zhou L, Lai K K, and Yu L 2010 Least squares support vector machines ensemble models for credit scoring J. Expert Systems with Applications 37 127-33.

[9] Liang Y, Balcan M F, Kanchanapally V and Woodruff D 2014 Improved distributed principal component analysis Proc. Int. Conf. on Neural Information Processing System vol 2 (MA, United State) pp 3113-21.

[10] Qin S and Lu Z 2013 Research on text categorization based on sparse autoencoder algorithm J. Science Technology and Engineering 13 9422-26.