Architecture Design of Emotion Recognition Early Warning System Based on AE-TAN Algorithm

Yongxiong Zhang¹, Xiangzheng Huang¹ 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 improve the analysis effect of human emotion text data, this paper proposed NB algorithm, TAN algorithm and AE-TAN algorithm model method. In addition, three index parameters, precision, recall and F1 value, are compared and analyzed under three different algorithm model scenarios. The experimental results show that the AE-TAN algorithm is superior to the other two algorithms in Chinese text emotion analysis. Based on the AE-TAN algorithm, the architecture design scheme of emotion recognition early warning system was proposed in this paper, which has the characteristics of high concurrency and high availability, thus laying a technical foundation for the system that can quickly and accurately warn people of psychological problems.

1. Introduction

With the rapid development of China's economy, people's mental health problems become more and more prominent with the rapid development of the pace of life. According to the literature [1], about 15% of the respondents have mental illness, and most of them feel that they have no place to complain about their psychological problems. This shows that the supply and demand of mental health is not in direct proportion, and the society cannot provide people with adequate mental health services and meet people's mental health needs. When people comment on something, they often express their true feelings. If people with psychological problems can be quickly analyzed from the comments, the negative social impact caused by mental illness can be reduced.

With the development of artificial intelligence, text emotion analysis technology has attracted the attention of many scholars. Generally, they use dictionary-based method and machine-learning method [2,3]. Due to the low recall rate and high word formation cost of dictionary-based method [4], most text emotion analysis technologies currently adopt machine learning methods. The classification methods of machine learning generally include k-nearest neighbor, neural network, support vector machine and naive Bayes. Among the above classification methods, naive Bayes has the characteristics of stable classification efficiency, less sensitive to missing data, simple algorithm and commonly used in text classification. Naive Bayes classifier (NB) is established based on the condition that each feature is independent from each other, which obviously does not conform to the actual situation of text analysis. Scholars at home and abroad have made many achievements in the study of Naive Bayes classifier in text analysis. For example, in order to weaken the limitation of feature independence in Naive Bayes algorithm, Friedman et al. proposed tree-enhanced Naive Bayes (TAN). Rao Lili et al. proposed the calculation of anti-document frequency based on the features of word frequency and document frequency between classes [5], and Wu Jianjun et al. made weighted
processing in category mutual information [6]. Escalante et al. describes a novel approach to learning term-weighting schemes [7]. The efficiency of deep learning model in extracting text features is usually better than naive Bayes classifier [8]. Since emotional text analysis involves high latitude data sets, AutoEncoder (AE) technology in deep learning can achieve better dimensionality reduction of data. Compared with NB algorithm, the classification effect of TAN algorithm is better. Therefore, this paper proposes a classifier algorithm combining AE and TAN for text emotion analysis, and designs the architecture of emotion recognition early warning system based on this algorithm, which provides a feasible technical scheme for the subsequent implementation of the system.

The rest of this paper is organized as follows. Section II introduces the concept and model construction of NB algorithm, TAN algorithm and AE-TAN algorithm. In section III, the performance of the three algorithms is compared and analyzed. Then, in section IV, the system architecture is designed. Finally, the conclusion is presented and future work is discussed.

2. Building AE-TAN Classification Model

2.1. NB

NB (Navie Bayes) classifier is a classification algorithm based on Bayesian decision theory. It is to implement the decision under the probability framework. In the ideal case where all relevant probabilities are known, the optimal category is selected according to the probability and misjudgment loss [9].

Suppose there are m categories, which are recorded as \( C = \{C_1, C_2, \ldots, C_m\} \). \( \lambda_{ij} \) is the loss value of a sample that truly belongs to class \( C_j \) and is wrongly marked as class \( C_i \); \( P(C_i|X) \) represents the probability that the sample belongs to class \( C_i \) under the sample condition \( X \); therefore, the expected loss generated by a sample classified as class \( C_i \) is as follows:

\[
EL(C_i|X) = \sum_{j=1}^{m} \lambda_{ij} P(C_j|X)
\]

(1)

According to formula (1), the smaller the value of \( EL \) is the better. The Bayesian optimal classifier can be expressed as:

\[
BOC(X) = \arg \min_{C_i \in C} (EL(C_i|X))
\]

(2)

According to formula (2), the classification error rate needs to be minimized, so the loss value \( \lambda_{ij} \) can be denoted as:

\[
\lambda_{ij} = \begin{cases} 
0, & \text{if } i = j \\
1, & \text{otherwise}
\end{cases}
\]

(3)

Then the expected loss is:

\[
EL(C_i|X) = 1 - P(C_i|X)
\]

(4)

It can be obtained from formula (2) and formula (4):

\[
BOC(X) = \arg \max_{C_i \in C} (P(C_i|X))
\]

(5)

According to Bayes theorem, it can get
Naive Bayes classifiers are based on independent features. Suppose the sample $X$ has $k$ characteristics. $X_j$ is the value of $X$ in the $j$-th feature and if all categories are the same, the expected loss is as follows:

$$BOC(X) = \arg \max_{C_i \in C} (\prod_{j=1}^{k} P(X_j | C_i))$$

(7)

This is the expression of NB classifier. If the training set sample is not sufficient and the probability estimation is zero, then Laplace correction can be used to improve it.

2.2. TAN

TAN (Tree Augmented Naive Bayes) is a classification scheme proposed by Friedman et. al [10]. In order to reduce the difficulty of assuming that all features of NB classifier are completely independent, the basic idea is to add some interdependent information among attributes, and the dependence of stronger attributes cannot be ignored.

Suppose $I(X_i, X_j | C_k)$ represents the conditional mutual information between feature $X_i$ and feature $X_j$.

Under the condition of class variable $C_k$, it is defined as:

$$I(X_i, X_j | C_k) = \sum_{X_i, X_j, C_k \in C} P(X_i, X_j | C_k) \log \frac{P(X_i, X_j | C_k)}{P(X_i | C_k)P(X_j | C_k)}$$

(8)

According to $I(X_i, X_j | C_k)$ as the weight, a spanning tree is constructed, which keeps the strong dependence of attributes.

According to formula (7) and the structure of Tan, the expected loss is as follows:

$$BOC(X) = \arg \max_{C_i \in C} (\prod_{j=1}^{k} P(X_j | \Pi_j, C_i, G))$$

(9)

Where, $G$ is the maximum weight span tree of the node $C_i$, and $\Pi_j$ is the characteristic parent node in the maximum weight span tree $X_j$, which can be evaluated as 0 or 1.

2.3. AE-TAN

AE(AutoEncoder) can be regarded as a nonlinear patch enhanced version of PCA algorithm. It is more flexible than PCA. It can represent both linear and nonlinear transformations. However, PCA can only perform linear transformations. In terms of data dimensionality reduction, setting appropriate dimensions and sparse constraints, AutoEncoder is more accurate than PCA algorithm [11].

The principle of AutoEncoder can be summarized as follows: Firstly, the data set is coded to reduce the dimension of data. Then, the data after dimensionality reduction is decoded, and the data before encoding is restored to minimize the loss of data before encoding and after decoding. Suppose $X$ is the input data, the encoding function is $F(X)$, the decoding function is $G(X)$, and $X'$ is the output data after decoding. The relationship can be expressed as $X \xrightarrow{F(X)} H \xrightarrow{G(X)} X'$, it is as follows:

$$X' = G(F(X))$$

(10)

As $X' \approx X$, the mean square error is used to measure the similarity between the data before encoding and the data after decoding, which is expressed as
Since the mean square error may become very small, resulting in over fitting, L1 regularization is used to suppress over fitting, and the mean square error is modified to

$$\text{MSE}(X, G(F(X))) = E \| X - G(F(X)) \|^2 + \alpha \sum_i \omega_i$$  \hspace{1cm} (12)

Where, $\alpha$ is the regularization coefficient, and $\omega_i$ is the excitation value of the $i$-th neuron.

AE-TAN is a text analysis model that combines deep learning AutoEncoder technology with Tan algorithm. The analysis model is divided into three stages: the first stage is the preparation of text sample data; The second stage is the feature deep learning of AutoEncoder, which realizes the dimensionality reduction of the data set and obtains the critical sensitive data. In the third stage, TAN algorithm is used to identify and classify the sensitive data extracted by AutoEncoder. The AE-TAN model is shown as figure 1.

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

According to the evaluation indicators of mental health, the results of TAN classifier analysis were comprehensively evaluated, so as to determine the emotional recognition warning level and realize the function of rapid and accurate warning of people's psychological problems.

3. Experimental Results and Analysis

The experimental data of this paper came from Professor Tan's review of non-equilibrium hotels after weight removal [12]. There were 1172 negative emotion data and 5358 positive emotion data. First, NLPIR word segmentation tool was used for Chinese word segmentation, and then stop words were removed. Then, the processed data was divided into training set and test set by random partition method, with a proportion of 70% training set and 30% test set.

Remove the stop words first, and then set the positive emotion value to 1 and the negative emotion value to -1. The Precision, Recall and F1 values were selected for analysis. Precision is used to represent the prediction accuracy of positive sample results, and Recall is used to represent the recognition degree of positive sample data. Because Precision and Recall are contradictory, it is impossible to guarantee the high Precision of both. For example, if high Precision is needed, Recall will be lowered, or if high Recall is needed, Precision will be lowered. F1 value is the highest precision value in the balance state of precision and recall. In this paper, NB algorithm, Tan algorithm and AE-TAN algorithm are used to analyse the sample data. The result is shown as table 1.

| Algorithm | Precision | Recall | F1   |
|-----------|-----------|--------|------|
| NB        | 0.815     | 0.767  | 0.790|
| TAN       | 0.822     | 0.774  | 0.797|
| AE-TAN    | 0.841     | 0.796  | 0.818|
Through contrastive Analysis in Table 1, AE-TAN algorithm model is higher than other two algorithm models in Precision, Recall and F1.

4. System Architecture Design
Because AE-TAN algorithm is superior to NB algorithm and Tan algorithm in Chinese text sentiment analysis, the text analysis module uses it to realize its analysis function. The early warning system is mainly composed of terminal, load balancing module, application service module, early warning module, text analysis module and data management module. The system architecture is shown in figure 2.

![System Architecture Diagram](image)

Figure 2. System architecture

The terminals include smart devices such as smartphones, computers and ipads. Based on the content of user identity and emotional text data, the system uses the corresponding interaction strategy to generate the corresponding content and return it to the terminal intelligent device. After receiving the affected content, the device performs the corresponding interaction operations, such as submitting the text data, viewing the analysis results, etc.

Load balancing module used the technology of keepalived and nginx. Keepalived is configured by master-slave server. When there is some problem with the master server, it can immediately switch to the slave server for execution to ensure that the service is not interrupted, thus realizing high availability. Nginx is responsible for distributing the request to the corresponding application server for processing, and uses the default allocation strategy to forward the request, so as to achieve high concurrency.

Data management module includes structured data and unstructured data. The structured data uses mysql database. The unstructured data uses Redis database. The Redis database stores the data that does not change frequently, which is used for application service module, early warning module and text analysis module to improve the system data access efficiency. Text analysis module uses AE-TAN algorithm to better classify the emotional text data. Early warning module comprehensively evaluates the early warning level based on the results classified by text analysis module. The application service module is responsible for processing terminal requests, such as verifying the validity of user identity, calling out analysis history data, analysing new text data and returning alert level, etc.

5. Conclusion
This paper uses the emotional data of non-equilibrium hotel reviews after weight removal. NB algorithm, Tan algorithm and AE-TAN algorithm are compared and analysed. And it is concluded that the AE-TAN algorithm is a better one in this kind of emotion analysis. Meanwhile, based on the AE-TAN algorithm model, an architecture scheme of emotion recognition early warning system is designed. This scheme has the characteristics of high availability and high concurrency. It can enable users to get the psychological status and early warning degree of the analysed object, so that people who have the tendency of psychological problems can be intervened in time, and effectively prevent the adverse impact on society.

In the next process, we will, complete the test and deployment of the system, and promote the system to be used free of charge by leather enterprises in Shiling Town of Guangdong Province and
some middle schools in Guangzhou City. We will continue to improve the algorithm of Bayesian classifier, improve its classification effect and efficiency by collecting data. At the same time, in order to ensure the high concurrency and high availability of the system, we will reconstruct the microservice architecture to make the system have better performance.

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] Xiaolan F, Kan Z, Xuefeng C, Zhiyan C 2019 Report on National Mental Health Development in China (2017-2018) (Beijing: Social Sciences Academic Press)
[2] Feng X, Zhang Z and Shi J 2018 Text sentiment analysis based on convolutional neural networks and attention model Application Research of Computers 35 (5) 1434-1436
[3] Liu M, Wang X, Li H and Zhang B 2018 Emotional Analysis of Online Commodity Reviews Based on Text Mining Journal of Liaoning University of Technology 38 (5) 1-5
[4] 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 Journal of Beijing university of technology 43 (3) 402-412
[5] Lili R, Xiong Hui L and Dongzhan Z 2012 An Improved Weighted Naive Bayes Classification Algorithm Using Feature Correlation Journal of Xiamen University (Natural Science) 51 (4) 682-685
[6] Jianjun W, Chang Bing L 2017 Mutual Information-Based Weighted Naive Bayes Text Classification Algorithm Computer Systems & Applications 26 (7) 178-182
[7] Escalante H, Garcia-Limon M A, et al 2015 Term-weighting learning via genetic programming for text classification Knowledge-Based Systems 83 (1) 176-189
[8] WenKai W, Liming W and Yumei C 2019 Sentiment analysis of micro-blog based on CNN and Tree-LSTM Application Research of Computers 36 (5) 1371-1375
[9] Lungan Z, Liangxiao J, Chaoqun L and Ganggang K 2016 Two feature weighting approaches for naive Bayes text classifiers Knowledge-Based Systems 100 (C) 137-144
[10] Friedman N, Geiger D, Goldszmidt M 1997 Bayesian Network Classifiers Machine Learning 29 (1) 131-163
[11] Shengjun Q and Zhiping L 2013 Research on text categorization based on sparse AutoEncoder algorithm Science Technology and Engineering 13 (31) 9422-9426
[12] "Emotional data" [Online]. Available: https://download.csdn.net/download/qq_4208267/10904313