Sentiment Mining Analysis of Fuzzy Evaluation Comment Data Based on the Structure of Similar Deep Learning

Haifu Jiang1,*

1School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang, China

* Corresponding author e-mail: jhfmail@foxmail.com

Abstract. With the rapid development of the Internet, it is increasingly necessary to mine and analyze the sentiment of comment data. The purpose of this paper is to use the deep learning feature construction method to develop an effective mechanism to solve the learning problem under the imbalanced distribution of sentiment analysis of comment data, build an efficient model that can classify comments, and help the government correctly guide the people's sentimental inclinations. This paper adopts the method of combining keyword extraction and incremental discretization through this model, extracts keywords from the comment corpus, sets up a fuzzy matrix according to the weight of the keywords, and uses incremental K- in the discretization of the fuzzy matrix. The means algorithm extracts some additional structural features as auxiliary information to increase the prediction accuracy of the network. Through experiments, it is found that the comprehensive evaluation results obtained by this method on the comment data set are only 0.02% different from the comprehensive scores of other papers, indicating that the results obtained by the model are extremely accurate. Moreover, the model is adjusted with hyperparameters, and it is compared with the text sentiment analysis model proposed in some articles in recent years. Experiments show that the review data sentiment analysis model constructed with deep learning features has a good classification effect and a wide range of applicability.

Keywords: Feature Construction; Deep Learning; Comment Data; Sentiment Analysis; Data Mining

1. Introduction
As an effective feature learning method, deep learning has been developed rapidly in recent years. It can transform complex problems into linearly separable problems through nonlinear models. It can be used in computer vision, natural language processing, unmanned driving and other fields. Plays a key role [1]. In the context of big data, time and manpower are relatively expensive resources. In the feature extraction stage, it takes time and effort to extract features from data manually. In the feature learning stage, traditional sentiment analysis relies on shallow learning algorithms to learn the sentiment information in features [2-3]. In machine learning, algorithms other than deep learning are
called shallow learning algorithms. The difference between shallow learning algorithms and deep learning algorithms is that shallow learning algorithms have low computational complexity and strong text learning capabilities for logic. Insufficient, and deep learning algorithms can learn more logical text information and improve the accuracy of sentiment analysis [4].

Text sentiment analysis is a relatively new research field, attracting a large number of scholars to conduct research. Scholars generally use relevant algorithms in natural language processing, machine learning, deep learning and other fields for research. Scholars at home and abroad generally use machine learning algorithms for sentiment analysis, that is, as a classification problem processing [5]. The usual approach is to first vectorize the document, and then train the machine learning algorithm, and finally the machine learning algorithm gives the emotional polarity of the sample. In the research on sentiment classification of Chines Weibo, Ombabi A H et al. used SVM, Naive Bayes and N-ary model to classify the same Weibo text data, and found that SVM obtains the best results [6]. Yuan H, and Xu W, use the maximum entropy model to identify the opinion holder and evaluate the object, and extract the evaluation word [7]. Deep learning is based on neural networks, which is a branch of machine learning, and it has developed tremendously in recent years. With the deepening of research, deep learning has also achieved satisfactory results in text processing.

Knowledge-based methods in review data were initially widely used to develop natural language processing technology. Researchers are committed to manually writing complex rules to give computers the ability to understand human language [8]. Deep learning is dedicated to exploring how to select appropriate features and expressions from the original data to solve various complex tasks, and has stronger expressive ability compared with traditional machine learning algorithms [9]. For natural language processing, the method based on deep learning is a completely different research mode. This kind of method has obvious advantages in modeling, interpretation, expression ability and optimization. Using deep learning models to mine deeper language information contained in texts from large-scale corpora has become a popular research direction in natural language processing [10].

This paper introduces a neural network to solve the problem of incremental learning in the case of uneven distribution. First, clean the data, then extract the keywords of the comment data, calculate the comment factor set and the weight of the factor set, and finally through the sentiment analysis of the comment, The evaluation score corresponding to each factor is also obtained, and the variance of each attribute distribution is obtained through the incremental discretization method. Then, in the discretization space, the incremental K-means clustering algorithm is used to describe the structure of the data distribution seen in the data stream, and a part of the additional structured features are extracted as auxiliary information for the clustering, which is further integrated into the original features In the vector, to improve the predictive ability of the classification model. A comprehensive evaluation result can be obtained through the constructed review data tendency matrix.

2. Proposed Method

2.1. Types of Deep Learning Feature Construction Methods

Some high-level hidden distribution features can be used through clustering techniques. Similarly, we believe that these hidden features may help to strengthen the representation of the data, and further strengthen the difference between the two different categories of data. Extract these new The process of structured features can also be called feature learning, which is a term widely used in deep learning.

We mainly focus on two new features, one is to reflect the global distribution, and the other is to represent the local information of the corresponding data. The former, which is the global distribution information, constructs a feature map from the k centers of the K-means clustering structure To f. Different from the 0-1 coding scheme adopted by the others, the Euclidean distance between the data xi and each cluster center is used to form a new feature sequence \( \left\{ g_1, g_2, ..., g_k \right\} \), where the calculation formula of \( g_j \) is as follows:
\[ g_i^f = \| x_i - \omega_j \|_2 \]  

(1)

The feature sequence represents the relative position of the data \( x_i \) in the entire data space. For local distribution information, this article is the idea of class vector generation. This means that we should first find the cluster \( c_j \) to which the data \( x_i \) belongs, and then calculate the difference in the cluster The percentage of training data of the category to form a new feature sequence \( \{ l_1, l_2, ..., l_q \} \), where \( q \) represents the number of categories. It should be noted that in the incremental learning process, a \( k \times q \) matrix \( M \) needs to be retained, and the matrix \( M \) records each Temporary percentages of different types of training data in the cluster, and are constantly updated.

2.2. Emotional Mining Model of Evaluation and Comment Data

With the improvement of people's quality of life, people's evaluation of things no longer stays in one dimension. Only a comprehensive evaluation result can tell whether a thing is good or bad. In view of this, this paper combines the fuzzy matrix and proposes a deep learning-like feature construction fuzzy evaluation review data sentiment mining model.

(1) Calculation of keyword probability and weight

The factor set of the model is a set of keywords extracted from hundreds of thousands of comment data. Assuming that each comment sentence extracts \( n \) keywords, there are a total of \( K \) comment data, let \( A \) denote the sum of keywords that can be extracted from all comments, namely:

\[ A = n \times K \]  

(2)

Arrange all the acquired keywords and merge the repeated keywords to obtain a set of keywords KEY.

\[ \text{KEY} = \{ \text{key}_1, \text{key}_2, ..., \text{key}_n \} \]  

(3)

Among them, \( \text{key}_n \) represents the \( n \) keyword, and the calculation formula for the probability of the occurrence of the \( q \) \( (q \leq n) \) keyword in the KEY \( P(\text{key}_q) \) is as follows:

\[ p(\text{key}_q) = \frac{N(\text{key}_q)}{A} \]  

(4)

\( N(\text{key}_q) \) represents the number of occurrences of the \( q \) keyword. This paper selects keywords with relatively high probability and which can reflect the characteristics of the factor set as the evaluation index.

\[ a_j \text{ (where } j \leq n \text{ ) in formula (5) can be calculated by } P(\text{key}_j). \]

\[ a_j = \frac{p(\text{key}_j)}{\sum_1^n p(\text{key}_n)} \]  

(5)

Formula (5) is to normalize the probability of \( n \) evaluation factors.

(2) Construction of fuzzy matrix

According to the above process, the factor set \( U \) can be obtained. The evaluation process is to match each comment with keywords in the factor set. If the match is successful, the sentiment value of the corresponding comment sentence is recorded, and the corresponding comment can be obtained according to the size of the sentiment value. Counting the number of comment results obtained, the number matrix \( C \) of evaluation results can be obtained.

The fuzzy matrix \( \tilde{R} \) can be expressed as:

\[ \tilde{R} = \begin{bmatrix} r_{11} & r_{12} & ... & r_{1n} \\ r_{21} & r_{22} & ... & r_{2n} \\ ... & ... & ... \\ r_{m1} & r_{m2} & ... & r_{mn} \end{bmatrix} \]  

(6)
Among them, $r_i$ can be calculated by formula (7):

$$r_i = C_i / \sum_j C_j$$

The comprehensive evaluation result $\widetilde{B}$ is obtained by multiplying the weight vector $\widetilde{A}$ and the fuzzy matrix $R$.

$$\widetilde{B} = \widetilde{A} \cdot R = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

Where $b_n$ represents the value of the $n$ comment.

3. Experiments

3.1. Acquisition of Experimental Data Set

In order to ensure the smooth progress of the experiment, the web crawler technology is used to crawl the restaurant review data. The format of the original data obtained is shown in Table 1, and about 211010 reviews are selected as the corpus. According to its review star rating and manual proofreading, it is divided into two categories: positive and negative. In the end, there were 106,270 positive comments and 104,290 negative comments. Because the number of data sets is small, we randomly divide the data into training set and test set according to the ratio of 80%, 20%, and repeat the experiment many times, and the experimental results are averaged.

| Table 1. Example of original format of comment data |
|-----------------------------------------------------|
| Shop name          | Set meal          | Username | Price | Comment                      | Score |
| Spicy              | Boutique package for two with free WiFi | wozuiaich i | 46    | The group's two-person set meal, the portion is very large, the taste is not good! | 45     |
| Sanduo·Chengdu Cold Pot |

3.2. Experimental Environment

In order to ensure the smooth progress of the experiment, this experiment uses deep learning-like feature selection, and deep learning is required for sentiment analysis. This time, the operating system Ubuntu16.04 development language Python development platform Google Tensorflow deep learning framework memory 16G hard disk 1T is selected.

4. Discussion

4.1. Types of Deep Learning Feature Construction and Data Mining Analysis

(1)For the analysis of sentimental tendency of comments using deep learning and deep learning feature selection, negative_prob represents the probability of negative emotions, positive_prob represents the probability of positive emotions, and sentiment represents the emotional tendency of comments. According to the input and output specifications, the sentiment score of each comment is obtained, as shown in Table 2.

| Table 2. Emotion score table derived from different feature structures |
|---------------------------------------------------------------|
| Input              | Output                     | Negative | Positive | Confidence | Sentiment |
| Test               | The group set meal for two people, the portion is quite large, and the taste is good. | 0.1238   | 0.8953   | 0.742      | 2         |
| Control            | Portion Taste Set meal     | 0.2133   | 0.7281   | 0.7291     | 1         |

According to the probability of sentiment tendency obtained by sentiment analysis, the evaluation results are divided into five categories: "very positive", "positive", "moderate attitude", "negative" and
"very negative". According to the results of sentiment analysis, as shown in Figure 1.

(2) Through sentiment analysis, it is found that in fact, the use of deep learning feature construction does not cause a big gap in the experimental results. Moreover, in terms of experiment complexity and experiment time, similar deep learning can instead bring faster calculation speed. Next, we compare the effects of the two feature construction methods on the specific experimental sentiment tendency for analysis, as shown in Figure 2.

4.2. Comparative Analysis of Emotional Tendency
As shown in Figure 3, we analyze the weights. The two factors with larger weights are "taste" and "taste", which shows that most customers pay more attention to the taste of dishes. As a merchant, focus on the taste of your dishes and start with the taste of your products to improve customer satisfaction. The weight of this article is based on the analysis of comment data and only represents the views of most people. Customers can also customize the weight according to their own preferences, and then multiply the weight value with the fuzzy matrix to get a fuzzy comprehensive evaluation result of their own, so as to provide a reference for their choice of merchants.
5. Conclusions
Aiming at the shortcomings of comprehensive scoring of review data, this paper proposes a kind of deep learning feature construction fuzzy evaluation review data sentiment mining analysis model. The model extracts the keywords of the comments and constructs a fuzzy matrix according to the weight of the keywords. Experiments on the review dataset prove the effectiveness of the model. According to the analysis results of the model, not only the comprehensive evaluation results of customers on the merchants can be obtained, but also the different perspectives of store taste, environment, price and so on can be obtained, which meets the needs of different customers for different attention points and helps to a certain extent. Customers more accurately choose the merchants that suit them, and also provide more data references for merchants to adjust their business models and change business strategies.

References
[1] Chen L, Xu S, Zhu L, et al. A deep learning based method for extracting semantic information from patent documents[J]. Scientometrics, 2020; 1-24.
[2] Abdi A, Shamsuddin S M, Hasan S, et al. Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion[J]. Information Processing & Management, 2019, 56(4): 1245-1259.
[3] Ghorbani M, Bahaghight M, Xin Q, et al. ConvLSTMConv network: a deep learning approach for sentiment analysis in cloud computing[J]. Journal of Cloud Computing, 2020, 9(1): 1-12.
[4] Manogaran G, Varatharajan R, Priyan M K. Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system[J]. Multimedia tools and applications, 2018, 77(4): 4379-4399.
[5] Gopi A P, Jyothi R N S, Narayana V L, et al. Classification of tweets data based on polarity using improved RBF kernel of SVM[J]. International Journal of Information Technology, 2020: 1-16.
[6] Ombabi A H, Ouarda W, Alimi A M. Deep learning CNN–LSTM framework for Arabic sentiment analysis using textual information shared in social networks[J]. Social Network Analysis and Mining, 2020, 10(1): 1-13.
[7] Yuan H, Xu W, Li Q, et al. Topic sentiment mining for sales performance prediction in e-commerce[J]. Annals of Operations Research, 2018, 270(1-2): 553-576.
[8] Chen X, Xie H. A Structural Topic Modeling-Based Bibliometric Study of Sentiment Analysis Literature[J]. Cognitive Computation, 2020: 1-33.
[9] Wang W M, Wang J W, Li Z, et al. Multiple affective attribute classification of online customer product reviews: A heuristic deep learning method for supporting Kansei engineering[J]. Engineering Applications of Artificial Intelligence, 2019, 85: 33-45.
[10] Ali F, Kwak D, Khan P, et al. Fuzzy ontology-based sentiment analysis of transportation and city feature reviews for safe traveling[J]. Transportation Research Part C: Emerging
Technologies, 2017, 77: 33-48.