E-commerce evaluation text classification algorithm

Jingmiao Yang¹, Hongwei Yang¹, Jianping Zhao¹,* And Jinlong He

¹School of Changchun University of Science and Technology, Changchun, China

*Corresponding author e-mail: 360377488@qq.com

Abstract. Based on the research results of the false evaluation and identification of e-commerce shopping, this paper analyzes the characteristics of the evaluation content of the merchant's brush list. In order to solve the problem that the convolutional neural network (CNN) model is difficult to capture the feature information of the whole evaluation text in the false evaluation identification task, a convolutional belief network (CBN) model based on keyword weighting is proposed. Firstly, the TF-IDF algorithm is used to construct the keyword set, then the word vector is weighted by the keyword vector. Then use restricted boltzmann machine instead of the pooling layer of CNN model to reduce the dimension; finally, the weighted word vector is classified by this algorithm model to complete the recognition task of false evaluation text.

1. Introduction

Hu et al.[1] found that consumer decisions are largely influenced by product evaluation, which will affect the sales of goods. For greater profit margins, some merchants will buy false reviews. On the one hand, writing praise for their own goods to increase sales, thereby promoting the profitability of goods; on the other hand, writing bad reviews for competitors, suppressing competitors. This behavior seriously damages the environment and order of online shopping. In order to eliminate the adverse effects brought by such false evaluations and ensure the authenticity of e-shopping evaluation, an effective false evaluation and recognition algorithm is urgently needed.

Liang constructs a new multi-edge graph model in which each node represents an evaluator, each edge represents a relationship between a particular product reviewer, and an unsupervised iterative computational framework is used to identify false evaluations of commodities[2]. Based on the characteristics of the false evaluation, based on the content of the review and the behavior of the reviewer, Lin et al. proposed six time-sensitive features for the review of the content and the behavior of the reviewer, Liang's model and the unsupervised iterative computational framework is used to identify false evaluations of commodities[2]. Based on the characteristics of the false evaluation, based on the content of the review and the behavior of the reviewer, Lin et al. proposed six time-sensitive features for the review of the content and the behavior of the reviewer, Lin et al. proposed six time-sensitive features for the review of the content and the behavior of the reviewer, Lin et al. proposed six time-sensitive features for the review of the content and the behavior of the reviewer. Heydari et al. studied the detection methods of false evaluations, the individual false evaluation manufacturers and the false evaluation of groups, and gave the advantages and disadvantages of each of the three different classification tests[4]. Wang et al. designed a top-down computing framework, based on the behavior of the reviewer to establish a topological structure, based on the analysis score to determine the false evaluation[5]. The use of historical evaluation messages, activity, account review, etc. for reviewers to identify false evaluations has achieved quite good results. However, the water army platform is also aware of this approach, they are hired to be fragmented, have high reputation and have a certain number of years. The account number is divided into time segments to brush the evaluation, which makes it difficult to identify the false evaluation method according to the user. Together with the shopping platform, the function of anonymous evaluation is opened, which makes the above research have no effect on the detection of false evaluation.
Kalchbrenner et al. proposed a DCNN model to extract global features using dynamic k-max pooling without relying on syntactic parse trees[6]. Kim et al. used multi-channel convolutional neural network models for supervised learning, using word vectors as input features. Semantic synthesis operations can be performed in different consignment windows to complete the text classification task[7]. ST Hsu organically combines CNN with Recurrent Neural Network (RNN) to classify sentence from the semantic layer[8]. Yin proposed a convolutional neural network based on attention mechanism, and used the network in the sentence pair modeling task to prove the effectiveness of the attention mechanism and CNN combination[9]. The above method uses the CNN model to extract and classify the text, but the pooling operation will lose more text semantic information when performing feature extraction and dimension reduction, resulting in lower classification accuracy.

In view of the above two problems, this paper discards the attention of users, identify only the content of the review text, uses the RBM network algorithm instead of the feature extraction and dimensionality reduction of the pooling layer, and inputs the convolved features into the RBM through the connection. The mapping between neurons is performed to eliminate unnecessary feature noise and better preserve the original semantic structure of the sentence, thereby improving the accuracy of the classification model.

2. E-commerce text classification model

2.1. Improved IT-IDF algorithm

In the e-commerce shopping evaluation text data, the real evaluation and the false evaluation are different in the information content, and there are differences in the usage and usage of the evaluation words[10], and the false evaluation is more diverse than the real shopping evaluation words. To achieve the effect of beautifying or smearing goods, real users tend to evaluate fewer words. According to the characteristics of evaluation data, in order to enhance the accuracy of false evaluation, this paper proposes a TF-IDF algorithm based on keyword weighting. The keyword language is screened to form a word set. By assigning weights to keyword vectors above the threshold, the accuracy of the model classification model is improved.

Algorithm 1: Word Vector Feature Enhancement Algorithm Based on Improved TF-IDF

Input: E-commerce evaluation text set \( T = \{T_1, T_2, \ldots, T_n\} \)

Output: word vector weight \( W_i \)

1) count the word frequency \( f(T, x) \) of all words in text set \( T = \{T_1, T_2, \ldots, T_n\} \)
2) Words with word frequency above average value, form word set \( C = \{C_1, C_2, \ldots, C_k\} \)
3) Calculate the value of each word TF-IDF in \( C \)
4) retain words that are higher than mathematical expectations, composition keyword set \( X = \{X_1, X_2, \ldots, X_n\} \) and \( T, C \) respectively Word2Vec vectorized, corresponding to \( T' \) and \( C' \)
5) for each \( T_i' \) do
6) Take the i-th word vector from the text set \( t_i' \)
7) for each \( X_i' \) do
8) Take the i-th word vector from the word set \( X_i' \)
9) Calculate the value of \( \text{Sim} = \frac{C_i X_i'}{||C||_2 ||X||_2} \), join the collection \( \text{Sim} = \{\text{Sim}_1, \text{Sim}_2, \ldots, \text{Sim}_n\} \)
10) end for
11) Select \( \max \text{Sim}_i \) in the set \( \text{Sim} \) as the weight \( w_i \) of the word vector \( t_i \), add the weight set \( W = \{w_1, w_2, \ldots, w_n\} \)
12) end for
13) Normalized weight \( w_i = \frac{\exp(w_i)}{\sum_{j=1}^{n} \exp(w_j)} \)
14) return \( W_i \)
2.2. Design DBN model

Compared with other dataset samples such as news texts and scientific papers, the relative length of the e-commerce shopping evaluation is shorter, the language information is more complicated, and the CNN convolution and pooling operations are used to obtain text features [7][11]. The simplification of the operation causes the noise to appear in the result of feature extraction. At the same time, the pooling layer performs dimensionality reduction on the text, only pays attention to the maximum value of the Feature Map, ignoring the feature distribution state, resulting in loss of feature position information and affecting text classification accuracy. In view of the above problems, this paper designs a Convolutional Belief Network Model (CBN), as shown in Figure 1:

![Figure 1. CBN model](image)

After convolution, the resulting convolutional semantic feature \( C_i = f(\sum w_i x + b) \), \( f \) is the activation function relu, \( w \) is the sentence matrix, \( x \) is the current convolution matrix window, \( b \) is the offset value. In this paper, the Feature Maps obtained by convolving the same sentence matrix are concatenated. The connection rules are connected according to the order Feature Map obtained from the small to large convolution window, as follows.

\[
C = \{C_1, C_2, ..., C_n\}
\]

Where \( n \) is the current number of convolution kernels. The result \( C \) obtained by the above operation is expressed as the semantic feature of the current sentence, and then the dimension reduction effect is achieved through the setting of the number of neurons, and the probability of the neuron opening is calculated to achieve the effect of extracting features.

Such a hidden layer can extract the features of the input data of the display layer relatively accurately, and can also reliably reconstruct the display layer according to the features acquired by the hidden layer. In this paper, the Softmax function is used to classify the final output. The final output unit of the classifier needs the Softmax function for numerical processing.

2.3. Classification process design

In order to improve the accuracy of the model and reduce the computational cost of e-commerce shopping false evaluation texts, this paper first uses the algorithm to build a keyword set, weights the keywords in the evaluation text, and then uses the CBN model to depth the sentence matrix. Feature information mining, and finally complete the work of e-commerce shopping evaluation false classification. The overall structure is shown in Figure 2.

In the data processing stage, data cleaning and word segmentation are first performed on the evaluation data. On the one hand, the processed text data is mapped to a multi-dimensional continuous vector by word2Vec, and each sentence constitutes a sentence matrix; on the other hand, these are The text data is constructed by the TF-IDF algorithm. Using these keywords to add a different weighting tendency to the sentence matrix in the form of a vector product, the expression is: \( y = w_t(s_1, s_2, ..., s_n) \), \( w_t \) is the weighting tendency, \( s = \{s_1, s_2, ..., s_n\} \) is the current sentence matrix word vector set, and the obtained operation result \( y \) is taken as the next input.
The CBN model receives the weighted word vector of the previous step as input. To obtain the semantic feature, convolution operation is performed with the convolution window of different sizes and the original matrix, and the Feature Map of the same sentence matrix is connected to enter the neuron, which is visible. Feature mapping between layers and hidden layers, the features with discriminantness are preserved in the way of calculating the probability, and finally classified by Softmax function, the formula is as follows:

\[
p(y = i | Q) = \frac{\exp(\theta^T i Q)}{\sum_{j=1}^{d} \exp(\theta^T j Q)}
\]

Where \(Q\) is an evaluation sample, \(\theta^T i\) represents the weight matrix of \(i\)-th feature, \(p(y = i | Q)\) represents the probability that each sample belongs to the category, and \(d\) is the number of classification categories. This article needs to detect false evaluations, so we set the category \(d = 2\).

3. Experiment and result analysis

Experiment 1: Improved TF-IDF enhanced evaluation word feature performance verification

In order to verify the effectiveness of the improved TF-IDF algorithm to enhance the evaluation of word features, this paper uses the commodity evaluation keyword word set constructed earlier, weights the important word vectors according to the algorithm process, and trains the classifier with the obtained word vectors. Among them, the word vector selects three data sets in Table 1, and uses the word2vec tool to map words into 50 dimension word vectors. Through direct classification, subject word weighting, and the keyword weighted classification in this paper, the convolution neural network is selected as the classifier. The experimental results are shown in the following table 1:

| Weighting method          | Headphone evaluation data set | Clothing evaluation data set | Shoe evaluation data set |
|---------------------------|------------------------------|------------------------------|--------------------------|
| Unweighted                | 0.814                        | 0.821                        | 0.816                    |
| Subject weighting         | 0.836                        | 0.854                        | 0.845                    |
| Keyword weighting         | 0.876                        | 0.921                        | 0.869                    |

It can be seen from the results in Table 3 that the maximum precision of directly using the convolutional neural network and the subject word weighting is 82.1% and 85.4%, respectively, which
is much lower than the keyword-weighted classification accuracy, which indicates the traditional
classification method for the key words. The extraction ability is weak, and keywords cannot be used
as the center of semantic features. The TF-IDF algorithm based on keyword weighting proposed in the
text has achieved good classification results on all three data sets. The highest classification accuracy
is 92.1%. Compared with other algorithms, the stability of the algorithm is stable. Stronger, this shows
that the TF-IDF algorithm based on keyword weighting can make the model better complete the text
classification work, and verify the effectiveness of the method.

Experiment 2: CBN model training optimization algorithm

In order to find the most suitable training algorithm for CBN model, bgd, SGD and MBGD
algorithm are selected to design a contrast experiment to test the applicability of the algorithm. The
error change rate of the experiment contrast model during training is as shown in the figure 3:

![Figure 3. Model training stability comparison](image)

This proves that the model is convergent, and the convergence algorithm most suitable for this paper is BGD algorithm. As shown in the figure, with the iteration, the error reduction speed is fast,
and the model can converge to the global best. The model is feasible.

Experiment 3: Model classification performance comparison

In this paper, SVM algorithm, CNN model, DBN model and CBN model proposed in this paper
are used for classification and comparison experiments. In this experiment, only the earphone
evaluation data set was used, and the training set and test set data were randomly selected in a ratio of
7:3 for comparison experiments. The experimental results are shown in the figure 4:

![Figure 4. Model Classification performance](image)
It can be seen from the figure that the CBN and SVM, CBN, DBN algorithms in the accuracy, precision, recall and F1-Measure performance, the algorithm is stable on these four classification indicators, SVM and DBN algorithm The fluctuation is large. With the increase of training samples, the algorithm presents an overall upward trend, indicating that the classification performance is also increasing with the increase of training learning. At the same time, the performance of each algorithm in this paper is superior to other algorithms, and the performance superiority of the algorithm is obvious.

4. Conclusion
In this paper, a evaluation text classification algorithm based on keyword weighting is proposed. The experimental results show that (1) keyword weighting can give more attention to the key evaluation with high discrimination, form a more appropriate classification feature expression of the evaluation text, and better complete the false evaluation (2) The classification accuracy of the classification models under different commodity evaluation sets and different sentence vector dimensions shows that the CBN model can extract more e-commerce evaluation text semantic features, has higher classification performance, and has more advantages for the false evaluation.

References
[1] Hu N, Bose I, Gao Y, et al. Manipulation in digital word-of-mouth: A reality check for book reviews[J]. Decision Support Systems, 2011, 50(3): 627-635.
[2] Liang D, Liu X, Shen H. Detecting spam reviewers by combing reviewer feature and relationship[C]//Proceedings 2014 International Conference on Informative and Cybernetics for Computational Social Systems (ICCSS). IEEE, 2014: 102-107.
[3] Lin Y, Zhu T, Wu H, et al. Towards online anti-opinion spam: Spotting fake reviews from the review sequence[C]//2014 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014). IEEE, 2014: 261-264.
[4] Heydari A, ali Tavakoli M, Salim N, et al. Detection of review spam: A survey[J]. Expert Systems with Applications, 2015, 42(7): 3634-3642.
[5] Wang Z, Gu S, Zhao X, et al. Graph-based review spammer group detection[J]. Knowledge and Information Systems, 2018: 1-27.
[6] Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[J]. arXiv preprint arXiv:1404.2188, 2014.
[7] Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.
[8] Hsu S T, Moon C, Jones P, et al. A hybrid cnn-rnn alignment model for phrase-aware sentence classification[C]//Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers. 2017, 2: 443-449.
[9] Yin W, Schütze H, Xiang B, et al. Abenn: Attention-based convolutional neural network for modeling sentence pairs[J]. Transactions of the Association for Computational Linguistics, 2016, 4: 259-272.
[10] Li J, Ott M, Cardie C, et al. Towards a general rule for identifying deceptive opinion spam[C]//Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2014, 1: 1566-1576.
[11] Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1510.03820, 2015.