Sentiment Analysis For Product Reviews Based on Deep Learning

Shaozhang Xiao, Hexiang Wang, Zhi Ling, Lanfang Wang, Zhaoxia Tang
Faculty of Computer and Software Engineering, Huaiyin Institute of Technology, Huai'an, Jiangsu 223003, China
Author’s e-mail: xiaoshaozhang@hyit.edu.cn

Abstract. With the popularity of the Internet and e-commerce, the sentiment analysis of text can help users to quickly and accurately obtain effective information they are interested in from massive product reviews to purchase satisfactory products. In this paper, a sentiment analysis system for product reviews was designed based on deep learning, and the digital and electronic products on Jingdong Mall with at least 100,000 reviews were crawled as the training data set. After data pre-processing operations such as word segmentation and removal of stop words for product reviews to remove useless features, the feature vectors were constructed based on the bag of words model and word2vec method, and then three classification algorithms, namely LSTM, Naive Bayes and logistic regression were used to model reviews. The LSTM algorithm is significantly superior to Naive Bayes and logistic regression algorithm in the training stage, and provides a reliable reference for the analysis of product reviews.

1. Introduction
The rapid development of information technologies such as the Internet allows users to make valuable reviews on certain products on the Internet. These product reviews directly express the user's attitude and evaluation of a large number of emotional information contained in product performance, quality, appearance, etc., and have an important impact on people's cognition and emotion[1]. The data obtained by mining and analyzing these product reviews using certain processing methods can not only help consumers to understand the information such as the quality of the product, but also help manufacturers to improve product quality, and merchants can adjust sales strategies and enhance service quality based on product reviews, which are of great significance for e-commerce, manufacturers and potential consumers. However, the traditional statistical and classification methods are unable to process massive information on product reviews, making it necessary to adopt automatic and intelligent method to analyze review information[2].

At present, automatic analysis of review information is mainly the text sentiment analysis in the field of natural language processing. The current methods for text sentiment analysis mainly include sentiment lexicon-based methods and machine learning-based methods. Since the sentiment lexicon-based methods mainly rely on sentiment lexicon, while the sentiment lexicon with high quality requires a lot of manpower, and sentiment lexicon is often unable to cover a large number of words that appear in texts in different fields, particularly, new words on the Internet will obviously influence the performance of the method. In recent years, many methods based on machine learning have emerged as artificial intelligence is continuously developing, among which the commonly used classification algorithms are Naive Bayes, K-Nearest Neighbor algorithm, logistic regression, decision tree algorithm, neural network algorithm and support vector machine. The constant development of
deep learning technology has made it more and more applied to the field of natural language processing, and satisfying results have been achieved. In 2014, Kim [3] applied the model based on CNN algorithm for text classification task for the first time, and achieved better experimental results than using traditional machine learning methods. Recurrent Neural Network (RNN) has become the primary method to study serialized data such as text. In 2016, Zhang and Liu [4] proposed a tree LSTM neural network model structure by combining the characteristics of the grammatical structure of natural language, and achieved good classification results on text sentiment classification task. The sentiment analysis methods based on deep learning have their unique advantages compared to the existing mainstream sentiment lexicon-based and machine learning-based methods for text sentiment analysis, which is mainly due to the strong expressive ability of deep network structures, and the ability of these deep network models to automatically extract deep abstract features from text.

Therefore, this paper used the LSTM algorithm to process product reviews, and two algorithms: Naive Bayes and logistic regression, were selected for comparison and experiment. Naive Bayes is a classification algorithm based on class probability, and it has a hypothesis that all features are independent of each other and do not affect each other. Since the Naive Bayes classifier has a strict hypothesis on the data, its training effect is usually worse than that of complex models, but it can realize fast speed of training and prediction. Logistic regression is a classification algorithm which is commonly applied in binary classification. It mainly converts the input values into the predicted values through the Sigmoid function to predict classification. LSTM [5, 6] is a kind of optimized Recurrent Neural Network (RNN), which is specially designed to solve the long-term dependence of RNN. The main key components of LSTM are cell state and gate structure. This structure allows LSTM to have "memory", so that it can learn the relationship between the context in a text sequence. LSTM not only conducts sentiment analysis through keywords, but also analyzes the relationship between words, making the analysis results superior to those of Naive Bayes and logistic regression.

2. Process of sentiment analysis based on machine learning
Text sentiment analysis involves multiple fields such as natural language processing, text processing, data mining, artificial intelligence, and machine learning, and the text sentiment analysis method based on machine learning is an application of artificial intelligence in the field of natural language processing. With this method, the data is preprocessed and converted the word vectors at first, then the text is modeled and the features suitable for the sentiment analysis task are extracted, and classified with the relevant algorithms of machine learning. The steps of text sentiment analysis based on machine learning are shown in the Figure1.

![Figure 1 Flow chart of text sentiment analysis based on machine learning](image)

3. Data processing
Word segmentation refers to dividing a Chinese sentence into individual words. It is a process where consecutive character sequence is recombined into word sequence according to certain criteria. In this stage, jieba was mainly used for word segmentation. First, the jieba word segmentation tool and a custom word segmentation table were loaded, and then regular expression was adopted to remove characters that are not Chinese in the Chinese sentence to facilitate the word segmentation of jieba.
Finally, relevant methods of jieba were called to perform word segmentation, and the results obtained by word segmentation were processed.

Processing stop words refers to the removal of words such as auxiliary particles, adverbs, prepositions, and conjunctions that have no impact on sentence sentiment, so as to reduce useless features in the data. In this stage, the stop word list developed by Harbin Institute of Technology was mainly used. First, the words in the word list were read by loading, and then compared with the word list after word segmentation to remove useless words in the word list after word segmentation.

Through word segmentation and removal of stop words, the useless data features in the training data samples can be reduced to increase the accuracy of model training.

4. Word vector conversion

Word vector conversion is mainly to convert the training data into digital information that can be processed and recognized by computer. In this paper, two methods were used to convert feature vectors, the method based on the bag of words model and the method based on the word vector model [9].

In the method based on the bag of words model, the CountVectorizer method of the scikit-learn framework was adopted for conversion.

In the method based on the word vector model, the Chinese word vector model open sourced by researchers from Beijing Normal University and Renmin University of China was employed [10]. The steps of feature vector conversion based on this model are as follows:

**STEP1:** A pre-trained word vector model is loaded; the dimension of each word vector is 300;

**STEP2:** The training data set is loaded;

**STEP3:** Text preprocessing is conducted on the data set to convert the training data into the form of list \( W = \{ x_1, x_2, \ldots, x_n \} \), where \( W \) is each review;

**STEP4:** Each word \( W = \{ x_1, x_2, \ldots, x_n \} \) in the list is matched with the word vector model, and the word is converted to the corresponding index \( W_2 = \{ m_1, m_2, \ldots, m_n \} \), where \( W_2 \) is the review data in digital form;

**STEP5:** The converted vector matrix is standardized, that is, the data length of each sample is the same, redundant data are deleted, and 0 is used for padding in the case of insufficient data.

5. Model establishment

In this paper, Naive Bayes, logistic regression and LSTM were used for training, among which the LSTM model is set as follows:

**STEP1:** The word2vec model is used to convert the sample data into a vector matrix;

**STEP2:** The sample vector matrix is standardized into a matrix with a length of 80;

**STEP3:** An embedding layer is constructed, where a matrix of dimension of 80 is input, and a matrix vector of 30*60000 is output;

**STEP4:** A bidirectional LSTM neural network layer is added, where the number of neural units is set to units=64;

**STEP5:** A layer of LSTM network layer is added, where the number of neural units=16;

**STEP6:** An output layer Dense layer is added and the activation function is set to "sigmoid";

**STEP7:** "adam" is selected for the optimizer and the learning rate is set to 0.001;

**STEP8:** Set within 5 epochs, if the verification loss rate does not change, the model stops training, if the verification loss rate does not improve, the learning rate is reduced;

**STEP9:** The prepared training data set is loaded for training, epoch=20, batch_size=128.

In this experiment, an 80*300 vector matrix was used, and the structure of the constructed neural network is shown in Figure 2.
6. Experiment and analysis

6.1. Preparation of data set
The data sets of reviews used in this paper are as follows: (1) Training data set: product reviews crawled from the Jingdong Mall website. The data contains 20 categories of products such as office, computer, personal care cleaning, mother and baby, each category contains 30 products, and each product has about 200 reviews. After depletion of duplicated data and other operations, there are 40558 negative reviews and 51889 positive reviews. Then 10000 pieces of data were selected from the data set obtained on the Internet. The final training data set contains 100,000 reviews, with half negative and half positive reviews (labeled as 0 and 1).

6.2. Data pre-processing
The training data were pre-processed. After the Chinese word segmentation, the number of features that are converted into word vectors using the CountVectorizer class is 49224, which contains many useless features such as numbers. Next, the data were further processed. By adding stop words and removing numeric characters, the number of features retained is 16842, indicating that many useless features are reduced through data pre-processing, which can reduce the effect of useless features on model performance. The processing results are shown in Figures 3 and 4.
6.3. Training results

The 75,000 reviews in the training data set were trained. Naive Bayes used is polynomial Naive Bayes, and cross-validation shows the accuracy of the Naive Bayes model is 92.077%. The 25,000 reviews in the training set were used to test the model, and the accuracy of the Naive Bayes model is 92.068%. The training results are shown in Figure 5.

Cross-validation in the training data set shows the accuracy of the logistic regression model is 94.731%. The 25,000 reviews in the training set were used to test the model, and the accuracy of the logistic regression model is 94.66%. The training results are shown in Figure 5.

When the amount of training data reaches 60,000, the accuracy rate of Naive Bayes and logistic regression tends to be unchanged, and the training has reached a bottleneck.
The analysis of the verification curve of Naive Bayes in Figure 7 reveals that when the Laplace smoothing coefficient (alpha) of the parameter of Naive Bayes is 1, the verification accuracy rate is close to the training accuracy rate, and the accuracy rate is higher, thus the effect is the best.

The analysis of the verification curve of logistic regression in Figure 8 shows that the inverse regularization parameter C of logistic regression is the optimal when the parameter C is greater than 1. When the parameter C is larger than 1, there is a huge difference between the results of training accuracy and those of testing accuracy, so the parameter settings are not good.

After the LSTM is trained in the training set, the accuracy rate and loss rate curves are shown in Figure 9. After 17 epochs, the accuracy rate (training accuracy) is 97.85%, val_accuracy (verification accuracy rate) is 96.88%, loss (training loss rate) is 6.59%, and val_accuracy (verification loss rate) is 9.72%. There are 81,000 training data samples and 9000 verification samples for this training. The 10,000 reviews in the training set were used for test, and the accuracy rate is 96.44%. The training effect has not changed much when the epoch is about 14, and it is no longer meaningful to continue training at this time. It can be seen from the figure that the training curve and the verification curve are very close, indicating that the features of the data set are basically learned by the network model, and the training has reached a bottleneck.
Figure 9 Learning curve of LSTM

Table 1 Accuracy rate of three models in the data set

| Data set         | Model algorithm   | Accuracy rate |
|------------------|-------------------|---------------|
| online_shopping_10_cats | Naive Bayes      | 79.41%        |
|                  | Logistic Regression| 79.89%        |
|                  | LSTM              | 85.48%        |
| weibo_senti_100k     | Naive Bayes      | 67.50%        |
|                  | Logistic Regression| 61.64%        |
|                  | LSTM              | 69.66%        |
| all2             | Naive Bayes      | 85.07%        |
|                  | Logistic Regression| 86.05%        |
|                  | LSTM              | 89.85%        |

It can be found from the analysis of Table 1 that the overall accuracy of the LSTM algorithm is higher than that of Naive Bayes and logistic regression. The accuracy of the three algorithms in the data set related to product reviews reaches about 80%, indicating that the models have achieved good training effect in the data set related to product reviews. However, the accuracy rate in the data set related to Weibo reviews is only about 60%, suggesting that the accuracy of the models needs to be improved by training the data in this aspect.

7. Conclusion
This study focuses on the sentiment analysis of product reviews. After data pre-processing and word vector conversion, the accuracy of the three algorithms, namely, Naive Bayes, logistic regression and LSTM in the data set related to product reviews reaches about 80%. However, the accuracy rate of the models in the data set related to Weibo reviews is only about 60%, indicating that the fitting ability of the models in the unseen data set is not good, which is attributed to the lack of corresponding features of training samples. Since this system is mainly applied to product reviews, the training effect of the models has reached expectation. LSTM exhibits good performance in product analysis in the system. It also performs well in predicting product reviews, but it does not achieve satisfying effects in the prediction of some short texts or unseen data set, and needs further improvement and optimization. In addition, the data sets of this system are mostly data sets on product reviews. The model has almost learned all the features of the data set after being trained to a certain extent, thus the effect of continuing to train the model is not good, and can be improved by expanding the range of the data set.

References
[1] Pang B, Lee L. A Sentimental Education: Sentiment Analysis using Subjectivity Summarization
based on Minimum Cuts[C]/Proceedings of Annual Conference of the Association for Computational Linguistics,2004:271-278.

[2] Kim S M, Hovy E. Automatic identification of pro and con reasons in online reviews[C]/Coling/acl on Main Conference Poster Sessions. Association for Computational Linguistics, 2006:483-490.

[3] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv,2014.

[4] Zhang L, Liu B. Sentiment Analysis and Opinion Mining[J]. Synthesis Lectures on Human Language Technologies,2016,30(1):152-153.

[5] Ye Yingya, Li Shujun, Feng Haonan, Li Mingxuan, Chen Ke. An LSTM-CNN Deep Network Model for Sentiment Analysis[J]. Journal of Guangdong University of Petrochemical Technology,2019,29(06):53-56+62.

[6] Du Yongping, Zhao Xiaozheng, Pei Bingbing. Short Text Sentiment Classification Based on CNN-LSTM Model[J]. Journal of Beijing University of Technology, 2019, 45(07):662-670

[7] Zhou Juan. Research on Text Sentiment Classification Based on Deep Learning [D]. Yangtze University, 2018.

[8] Olivier HABIMANA, Yuhua LI, Ruixuan LI, Xiwu GU, Ge YU. Sentiment analysis using deep learning approaches: an overview[J]. Science China (Information Sciences), 2020, 63(01): 21-56.

[9] Maria Giatsoglou, Manolis G. Vozalis, Konstantinos Diamantaras, Athena Vakali, George Sarigiannidis, Konstantinos Ch. Chatzisavvas. Sentiment analysis leveraging emotions and word embeddings[J]. Expert Systems With Applications,2017,69.

[10] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, Xiaoyong Du, Analogical Reasoning on Chinese Morphological and Semantic Relations, ACL,2018