Comment information extraction based on LSTM and Neural Networks

Qingliang Zhang, Binning Ma, Xier Zhong, Liangcai Mei* and Youyu Zhou
Department of Mathematics and Civil Engineering, Zhuhai Campus, Beijing Institute of Technology, Zhuhai, Guangdong, 519088, China
*Corresponding author’s e-mail: 13201@bitzh.edu.cn

Abstract. With the advent of the era of big data, the amount of data has also increased geometrically. People's ability to obtain effective information has gradually declined. At present, most e-commerce platforms only focus on the sentiment analysis of positive and negative reviews. It is difficult for users and businesses to extract user opinions and views from the massive review data. For the product review data of a certain hard disk, use the LSTM model to train the sentiment classification model. Finally, the neural network is used to find the keywords of the comment data and the word cloud diagram is used to display the analysis results. Through the research, it can be found that LSTM emotion classifier can classify comments with high accuracy and words closely related to comment emotion tendency can be found according to the weight of neural network.

1. Introduction
Today is the era of data explosion. In the face of massive amounts of data, artificially extracting effective information is very time-consuming and inefficient. With the help of natural language processing technology, the programmatic mining of massive data to help users solve the problem of information acquisition is an important direction of artificial intelligence at present. Online shopping has become the main way people to shop, and it's easy to find that reviews are the main factor that affects people's shopping. For a large number of reviews, mechanized information extraction can significantly improve users’ shopping efficiency and save users’ time. At present, text analysis technology is still in the developing stage. Although there are many people studying it, it is still unable to reach a mature level. Now, machines can only recognize text through vectors, but they still can't understand words from the human way of thinking.

Text sentiment analysis is a difficult point in the field of natural language analysis. At present, the most common analysis method is to categorize comments by establishing a sentiment classification model. Few people will analyze the text in detail, which makes users can only intuitively obtain people's general attitude towards the product, but can’t directly obtain other users' specific views on the product. Only obtaining people's emotional tendency towards goods can’t reflect the value of comment data. Therefore, adding keyword extraction in text sentiment analysis can increase the value of review information and enable users to quickly understand people's general views[1]. For users, being able to have a certain understanding of products from reviews can alleviate the user's product selection problem to a certain extent; for merchants, they have ability to extract the user's point of view from product reviews and look at products from the user's point of view, which is conducive to the improvement of the quality of goods and services of merchants. In this way, not only improve the
user's buying experience, but also improve the product quality of the merchant. Finally, a win-win situation between users and merchants is realized.

Badger J et al found that the text effect of random forest and logistics is better after adding keywords[2]. Zaghloul W et al research found that the effect of neural networks on text classification is not worse than that of support vector machines[3]. VM Kreňáková et al uses popular transformation models and traditional TF-IDF algorithm, pre-training word embedding and other methods to compare and evaluate the performance impact of text classification. Finally, the study found that the BERT-BASE uncased model has the best effect[4]. Hammami Linda at al research found that a rule-based method can effectively classify cancer forms from Italian pathology reports[5]. Hemmatian F et al studied the use of various methods to research opinions mining, analyzing the advantages and disadvantages of various methods[6]. Kastrati zenun analyzed the emotional tendency of the teacher's comments on students by using the emotional dictionary. Sentiment analysis based on comments is realized[7]. Dongwen Zhang et al found that using word2vec and SVM can effectively perform sentiment analysis on Chinese comments[8]. Alajlan A A et al research found that machine learning can effectively analyze whether speech is dangerous[9]. Sarah Omar Alhumoud and Asma Ali Al Wazrah research studies and analyzes the performance of RNN model in Arabic sentiment analysis[10]. Oleynik Michel et al found that the shallow method of clinical phenotyping is better than deep learning after pre training[11]. Miao T found that using Ship-Gram model to extract keywords has improved performance than other model[12]. Quan C et al set a built a blog sentiment corpus to analyze the sentiment of Chinese blogs[13].

This article trains a product review sentiment classification model based on product review data. By the model, product reviews can be classified. Then, the neural network-based classification model is trained to extract the key variables according to the model weight, so as to extract the keywords in the review.

2. Comment sentiment classification model

2.1. Text vectorization

After cleaning the text, the text is quantified using the form of one-to-one correspondence between the text and the subscript.

Since the word bag model and One-Hot algorithm don’t extract the text information completely, resulting in the loss of more text information, word2vec algorithm is used for text vectorization. Word2vec uses an ordinary fully connected neural network, which can effectively preserve the relationship between contexts[14].

2.2. Word2vec model

It can be seen from the Figure 1 that the word2vec consists of an input layer, a hidden layer and an output layer. The input layer is the One-Hot encoding of text characters. The output layer is the input encoding of two characters above and below the text characters. A character in the model is used to fit the character of its context, so the vector composed of the value of the intermediate hidden layer is the word vector of the character.

![Figure 1. Word2vec network structure.](image-url)
This article establishes a word2vec model to train text data. In order to visualize the relationship between texts, the PCA algorithm is used to reduce the 100-dimensional text vector to 2 dimensions. This article only visualizes fifty words to observe the results of the model. Through data analysis, we found that pretty good, very quickly and compatible are close together, indicating that these words play the same role[15]. This is in line with the actual situation, the above text is a good comment on the goods. This shows that the model has a good effect. Then, evaluate the effect of the model by analyzing the similarity between words. Use the word GREAT as an example to analyze similar words. The results are shown in Table 1.

| SIMILAR WORDS  | Similarity |
|----------------|------------|
| Needless to say| 0.36       |
| Receive        | 0.35       |
| Work           | 0.34       |
| Not bad        | 0.33       |
| Logistics      | 0.32       |
| Love           | 0.32       |
| Convenient     | 0.31       |
| Happy          | 0.31       |
| Afternoon      | 0.30       |
| As always      | 0.30       |

It can be seen from the Table 1 that words similar to the word GREAT are almost in line with actual life, such as Not bad, Happy, Convenient and other words. It shows that the word vector model has a good effect.

2.3. Model training
This article use tensorflow to train the lstm network. LSTM model is suitable for processing and predicting important events with very long interval and delay in time series. Similarly, it can be used to process comment data with contextual nature. Compared with the rnn network, lstm can be effectively avoiding the phenomenon of gradient dispersion.

The lstm network is processed in units of characters. Each character needs to pass through a neuron. Lstm has the function of long-term memory[16]. After each character passes through the neuron, the long-term memory will be updated. Finally, the sigmoid function is used for activation. Adding the relationship between comment texts to classify texts can effectively improve the accuracy of classification results. [17-18].

After using the lstm model for training, the training effect of the model is shown in Table 2.

| INDEX | Value |
|-------|-------|
| F1    | 0.93  |
| AUC   | 0.95  |
| Precision | 0.95 |
| Recall | 0.95  |
| Accuracy | 0.96 |
The F1 value, AUC value, precision value, recall value and accuracy value are all above 0.9, indicating that the training results of the model have good results.

3. Comment keyword extraction

3.1. Keyword extraction method

There are many methods for text vectorization, but it is difficult to find the words corresponding to the vector based on the vectorization of the text. If you need to find words based on vectors, there are fewer vectorization methods. The most direct method is to use the word frequency method of vectorization. By the way, we can not only quantify the text data, but also find the corresponding words according to the variables[19].

On the premise of having great quality review data, using this vectorization method to find the weights and input feature variables between positive and negative reviews in the neural network, so that the main features that affect the classification results can be analyzed. The variables with larger weights are extracted for inverse solution to obtain keywords that affect good and bad reviews. These words have a guiding effect on positive and negative reviews, so they can be considered as review keywords[20].

Use jieba word segmentation to segment the comment data, then use the stop vocabulary to filter and clean the comment data to remove useless words. Ultimately, use the term frequency algorithm to vectorize the words.

3.2. Building neural network sentiment classification model

After the text word is vectorized, all words in the comments are used as feature input, then the feature vector corresponding to the word is \( x_1, x_2, \ldots, x_n \). Hidden layer neuron is \( y_1, y_2, \ldots, y_n \), output layer neuron is \( z_1, z_2 \)[21].

Connection between hidden layer neurons and feature vectors:

\[
y_i = f(x) = \sum_{i=1}^{n} \beta_i x_i
\]

(1)

Connection between the hidden layer and the output layer:

\[
z_k = \sum_{j=1}^{m} a_j y_j
\]

(2)

Finally, use the Softmax function to predict the result. According to formula (1)(2), we can get formula (3).

\[
z_k = \sum_{i=1}^{n} \gamma_i x_i
\]

(3)

Among \( \gamma = \beta \alpha \), the \( \gamma \) is the weight between the input feature and the output layer, the input feature \( x_i \) corresponding to the larger \( \gamma \) is the keyword in the comment(The algorithm uses Pytorch library based on Python language version to construct neural network for modeling and calculation).

3.3. Model results

This article analyzes the review data of a certain hard disk product. Due to the limited product review data displayed on the platform, only 2000 review data are used for analysis. After establishing the text sentiment classification based on neural network, the model is fitted according to the comment data preprocessed by LSTM model. The accuracy of the model fitting result is 96%. According to the fitted model, the weight of each connection can be obtained. Finally, \( \gamma \) is solved according to \( \beta \) and \( \alpha \) so that the parameter with larger weight in each neuron can be obtained.

The result of extracting keywords from hard disk comments is shown in Figure 2.
4. Conclusion

With the geometric growth of information, information extraction technology plays a vital role in people's life. Quickly obtaining key information from comment data can effectively save people's time and improve people's life efficiency. This article establishes an lstm model based on the comment data to classify the hard disk comment data. According to the sentiment classification results of the lstm model, a neural network is used for fitting. Finally, according to the neuron weight of the fitted model, comment keywords can be found from the positive and negative comments.

![Figure 2. Keyword word cloud diagram.](image)

The keywords extracted by the weight of neural network can effectively extract the words related to the comment emotional tendency. However, when there is a great imbalance in the review data, it is easy to have poor performance. We hope to establish keyword dictionaries about different types of goods in the future. Then, it is believed that the extraction effect will be significantly improved by using the dictionary of corresponding commodities in the information extraction results of different comment data.

Acknowledgments

This work has been supported by university student innovation and entrepreneurship training project of Guangdong Province(S202013675037), characteristic innovation project of Guangdong Province (2019KTSCX217) and Big data research center of Zhuhai College, Beijing Institute of Technology (XJ-2018-05).

References

[1] Kurian D, Sattari F, Lefsrud L, et al. (2020) Using machine learning and keyword analysis to analyze incidents and reduce risk in oil sands operations. Safety Science, 130:104873.
[2] Badger J, Larose E, Mayer J, et al. (2019) Machine Learning for Phenotyping Opioid Overdose Events. Journal of Biomedical Informatics, 94.
[3] Zaghloul W, Sang M L, Trimi S. (2009) Text classification: neural networks vs support vector machines. Industrial Management & Data Systems, 109(5-6):708-717.
[4] VM Kreňáková, Sarnovsky M, Butka P, et al. (2020) Comparison of Deep Learning Models and Various Text Pre-Processing Techniques for the Toxic Comments Classification. Applied Sciences, 10(23):8631.
[5] HI A, Pa B, Pgc D, et al. (2021) Automated classification of cancer morphology from Italian pathology reports using Natural Language Processing techniques: A rule-based approach. Journal of Biomedical Informatics, 116:103712.
[6] Hemmatian F, Sohrabi M K. (2019) A survey on classification techniques for opinion mining and sentiment analysis. Artificial Intelligence Review, 52(1):1-51.

[7] Kastrati Z, Dalipi F, Imran A S, et al. (2021) Sentiment Analysis of Students’ Feedback with NLP and Deep Learning: A Systematic Mapping Study. Applied Sciences, 11(9):3986-3986.

[8] Zhang D, Xu H, Su Z, et al. (2015) Chinese comments sentiment classification based on word2vec and SVMperf. Expert Systems with Applications, 42(4): 1857-1863.

[9] Alajlan A A, Saudagar A. (2020) Machine learning approach for threat detection on social media posts containing Arabic text. Evolutionary Intelligence:1-12.

[10] Alhumoud S O, Wazrah A. (2021) Arabic sentiment analysis using recurrent neural networks: a review. Artificial Intelligence Review:1-42.

[11] Oleynik M, Kugic A, Z Kasá, et al. (2019) Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. Journal of the American Medical Informatics Association, 26(11):1247-1254.

[12] Miao T. (2021) Using the Ship-Gram Model for Japanese Keyword Extraction Based on News Reports. Complexity, 2021(2):1-9.

[13] Quan C, Ren F. (2010) A blog emotion corpus for emotional expression analysis in Chinese. Computer Speech and Language, 24(4):726-749.

[14] Gennaro G D, Buonanno A, Palmieri F. (2021) Considerations about learning Word2Vec. The Journal of Supercomputing:1-16.

[15] Jatnika D, Bijaksana M A, Suryani A A. (2019) Word2Vec Model Analysis for Semantic Similarities in English Words. Procedia Computer Science, 157:160-167.

[16] Hochreiter S, Schmidhuber J. (1997) Long Short-Term Memory. Neural Computation, 9(8):1735-1780.

[17] Xu Z, Chen B, Zhou S, et al. (2021) A Text-Driven Aircraft Fault Diagnosis Model Based on a Word2vec and Priori-Knowledge Convolutional Neural Network, 8(4):112-112.

[18] Shekhar S, Sharma D K, Beg M. (2020) Language identification framework in code-mixed social media text based on quantum LSTM - the word belongs to which language?. Modern Physics Letters B, 34(2):2050086.

[19] Wei C. (2021) Deep Text Information Type Detection Model Based on Statistical Learning in Engineering Application. IOP Conference Series Earth and Environmental Science, 769(4):042070.

[20] Liao Wenxiong, Zeng Bi, Liu Jianqi, Wei Pengfei, Cheng Xiaochun, Zhang Weiwen. (2021) Multi-level graph neural network for text sentiment analysis. Computers and Electrical Engineering 92:107096.

[21] A S D, B O M. (2002) Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics, 35(5-6):352-359.