Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/40/20230651

Text sentiment analysis based on the LSTM method

Zuhe Zhang

School of Electronics and Information Engineering, Tiangong University, Tianjin, 300387, China

zuhezhang@tiangong.edu.cn

Abstract. Sentiment analysis is an important research direction in the field of natural language processing. With the rise of various social network platforms, a large number of network users have begun to express and publish their opinions on X, LINE, Weibo, and other platforms, which makes sentiment analysis more and more important. However, as more and more people use social platforms, the number of comments from Internet users has surged, and Internet buzzwords and Internet popular topics are constantly changing. This phenomenon leads to an increase in the error rate of traditional text emotion detection methods, which in turn leads to a reduction in the operating efficiency of some network platforms. In addition, nowadays, sentiment analysis has a wider and more urgent need in many fields, especially in the business field. In order to cope with this change and demand, this paper focuses on a text sentiment analysis model based on a neural network model.

Keywords: Sentiment Analysis, Neural Network, Machine Learning, RNN, LSTM.

1. Introduction

The long short-term memory network (LSTM) is a variant of the Recurrent Neural Network (RNN), and research on it has continued since it was proposed in 1997. Currently, it plays a role in various fields such as meteorology, civil engineering, and transportation. At the same time, as more and more relevant literature is published, more and more academic and industrial workers are inspired and decide to learn the LSTM network to measure whether it is suitable for their own research or practical use [1]. Furthermore, sentiment analysis (SA) has been considered a popular research direction among researchers since the early 2000s, It is a method of predicting people's attitude toward something, such as a product or a social entity, through sentiment analysis [2]. With the deepening of research, various terms such as opinion mining, sentiment classification, comment mining, emotion mining, and opinion extraction have also been proposed and used in sentiment analysis [2]. At the same time, sentiment analysis can also be called opinion mining [3]. This is a process of extracting useful information from text opinions and analyzing user sentiment. As far as the text is fine-grained, the emotional polarity of the text can usually be classified from the document level, sentence level, and aspect level [4]. Coarse-level analysis can judge the emotional tendency of documents and sentences, while aspect-level analysis is aimed at corpus [4]. With the continuous rise of social platforms, not only the number of Internet users continues to increase, but also a large number and variety of online texts are published. This phenomenon prompts researchers to continue to conduct in-depth research on the sentiment analysis of online comments and conduct a more accurate analysis of them. emotion classification.

DOI: 10.54254/2755-2721/40/20230651

The function of sentiment analysis is to automatically identify and understand text information that is divided into positive, negative, and neutral text to facilitate people's understanding of large amounts of text. The text information of online comments contains huge commercial value, which can reflect the collective intelligence and emotional state of society, and affect the decision-making choices of individuals in social life [5]. Therefore, users' online reviews have become one of the important sources of information for merchants to improve product competitiveness [5]. To meet the challenges in the field of sentiment analysis, researchers have proposed a variety of techniques and methods, from traditional rule-based methods to deep learning models, including LSTM. This paper will focus on how to effectively use the LSTM method to improve the performance of text sentiment analysis and take the IMDB database as an example to discuss sentiment analysis under the LSTM method.

The remaining four parts of this article are presented in the following order. Section 2 introduces related research work in neural networks and other research fields. Section 3 describes the method used in this experiment to solve the text sentiment analysis problem. Section 4 presents and explains the results of the text sentiment analysis experiments. Finally, Section 5 summarizes the conclusions of this experiment.

2. Relate work

This study aims to provide a text sentiment analysis method based on LSTM to provide a reference for future related research. This paper focuses on finding solutions for sentiment analysis from the ML approach and data sets. Recently, deep learning models including LSTM have been used to complete sentiment analysis tasks and improve their efficiency. This section will review previous advanced sentiment analysis methods based on deep learning. In previous studies, researchers not only used methods based on emotional dictionaries and traditional machine learning to solve problems in text sentiment analysis but also explored and used deep learning-based methods to implement text sentiment analysis [6]. However, methods based on sentiment dictionaries are more suitable for sentiment analysis in a single domain. In the current situation that user comments on the Internet continue to grow and change, this method cannot be well adapted to the current network environment. Methods based on traditional machine learning also have similar problems. They also lack high-quality databases to support training, although they have certain scalability and repeatability. However, this accuracy is very dependent on the quality of the training set, and building a high-quality training set requires high labor costs [6]. Therefore, if this method is used to solve the current changing Internet comments, it will consume a lot of manpower and time costs. Compared with the first two methods, the method based on deep learning is relatively more suitable for solving the changing Internet terms and comments. It cannot be denied that it has stronger expressive power and model generalization ability, but the problem of lack of large-scale training data still exists [6]. Furthermore, RNNs process sequential information through internal memory captured by directed loops [7]. Unlike a punch-through neural network, it can also reuse elements from the input sequence because it remembers previously computed information [7]. In addition, LSTM, as a special type of RNN, can not only use long memory as the input of the hidden layer activation function but also overcome the memory limitation problem of traditional RNN [7]. This allows it to process sequence information more efficiently. The key components of LSTM include Cell State which is the core memory unit of LSTM, which can transfer information throughout the sequence [8]. The forget gate and input gate are the keys to LSTM's ability to remember long-term dependencies. The forget gate determines how much previous cell state information should be forgotten in the current time step [8]. It converts the input value to a value between 0 and 1, where 0 means completely forgotten and 1 means completely remembered. The input gate determines how much information about the network state at the current moment needs to be saved to the internal state [8]. It uses the Sigmoid activation function to generate outputs between 0 and 1 and the Tanh activation function to generate outputs between -1 and 1 [8]. The output gate determines how much information is extracted from the cell state for output at the current time step [8]. Similarly, it also uses Sigmoid and Tanh activation functions to generate outputs [8]. The gating mechanism of LSTM not only enables it to effectively process various sequence data but also avoids the limitations of traditional RNN.

During the training process, to obtain better training accuracy, the experimenter usually lets the model go through multiple epochs. However, in some cases, especially when the data set is large, traversing the data set multiple times is likely to lead to overfitting. One of the most effective and straightforward methods to solve this problem is early stopping, which is widely used because of its simplicity and convenience. Furthermore, in many cases, using this method to solve the overfitting problem during training is better than regularization methods [9]. It should be noted that using the early stopping method may sometimes stop training before the model fully converges, resulting in insufficient model performance and underfitting problems. At the same time, when early stopping is used, the performance of the validation set may fluctuate during the training process, making it difficult to determine the moment to stop training. To better use early stopping, the experimenters need a predicate named the stopping criterion to find when is the best time to stop. Researchers are still looking for the criterion that produces the lowest generalization error and is most appropriately used during training [10].

3. Method

This experiment mainly uses the IMDB movie review database from Keras to train the sentiment analysis model based on the LSTM method. The experiment involves the selection of the database, the construction of the training model, and the evaluation of the training model. Finally, the training process will be shown in the form of figures. Visualization of performance metrics. The following is the construction and training process of this experiment.

First, import and use the necessary function libraries to load the IMDB movie review dataset in the Keras database. Then, set relevant parameters to limit the size of the dictionary vocabulary and encode the model. After that, the data is preprocessed through the previously imported function library: a fixed length is set for the movie reviews in the data set to ensure the standardization of the text to facilitate later training. At the same time, the data in the IMDB movie review data set is split into two parts, the training which is used to train the model, and the validation which is used to test the performance of the model. This makes it easier to observe the performance of the trained model later.

After that, we start building the sequence of the model. First, we add an embedding layer to map the words in the text sequence into dense vectors. Second, an LSTM layer is added to process sequence data and capture its temporal dependence. Third, add a fully connected layer to introduce nonlinear properties. This process enables the model to learn complex features and patterns in the input data.

Then, the model needs to be compiled. Specifically, the model parameters are adjusted to minimize the loss function by setting the optimizer Adam, and the loss function “binary_crossentropy” is specified to set the evaluation criteria to evaluate the model’s performance on the classification task.

Finally, set its training period and batch size for the model.

4. Experimental results

Figure 1 shows that on the training set, starting from a certain moment, the training loss continues to decrease while the validation loss increases. At the same time, the training loss is much smaller than the validation loss and the training accuracy is much higher than the validation accuracy. This indicates that overfitting may be occurring.

DOI: 10.54254/2755-2721/40/20230651

Figure 1. Training and validation loss, training and validation accuracy over epochs in the first experiment.

(Picture credit: Original)

To prevent over-fitting, in the second training, the model added the method of early stopping to terminate the training process in advance to avoid over-fitting the model and increasing useless calculation time. Figures 2 and 3 show the training and validation loss over epochs and training and validation accuracy over epochs of the second training, respectively.

Figures 2 and 3 show that after five rounds of epochs, the training loss is approximately 11%, the verification loss is approximately 35%, the training accuracy is approximately 98%, and the verification accuracy is approximately 87%. This shows that the training loss and verification loss are small, and the training accuracy and verification accuracy are high and similar. This indicates that the overfitting problem has been resolved.

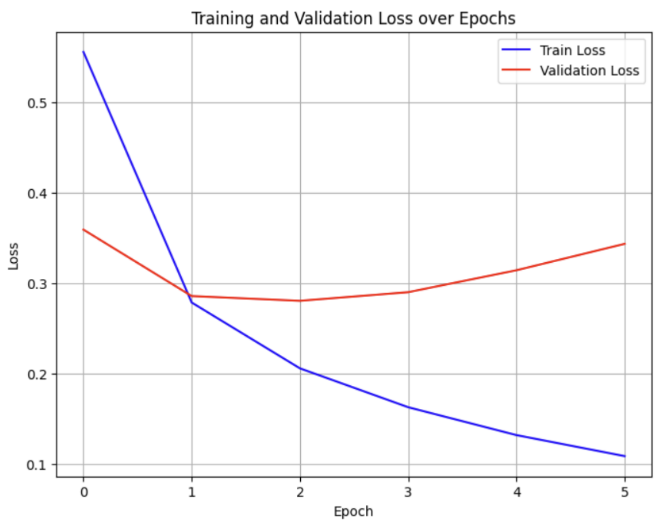

Figure 2. Training and validation loss over epoch in the second experiment.

(Picture credit: Original)

DOI: 10.54254/2755-2721/40/20230651

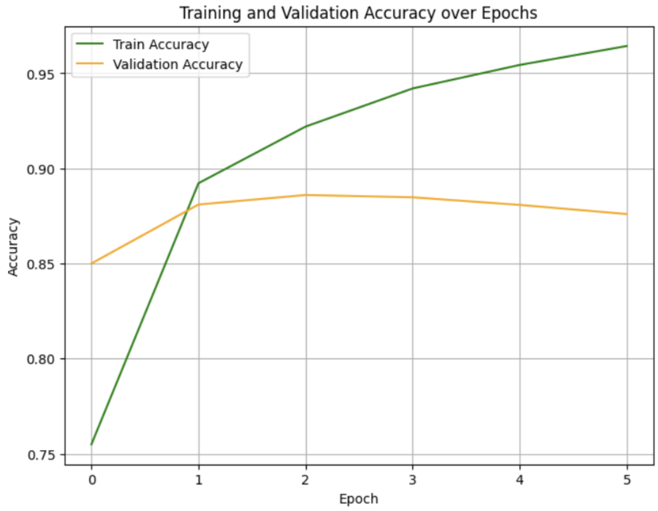

Figure 3. Training and validation accuracy over epoch in the second experiment.

(Picture credit: Original)

5. Conclusion

Nowadays, with the surge in the volume of online comments and the constant changes in online terminology, traditional text sentiment analysis methods are gradually becoming no longer applicable. At the same time, as one of the important ways for merchants to obtain product feedback, online reviews can not only help merchants improve their products and service quality, but also bring huge economic benefits to society. Therefore, to train a suitable text sentiment analysis model, we need to think about the solution from the characteristics of the text. For example, the sequential data structure of the text requires the model to have the ability to consider contextual information. LSTM can not only capture long-term dependencies in text but also has memory units that can remember key information in longer text sequences. Therefore, the experiment of this article chooses to build a neural network through the LSTM model and uses the IMDB movie review data set of Keras as an example to conduct text sentiment analysis. At the same time, to prevent overfitting, the model adopts an early stopping method. Finally, based on the visual data obtained from training, we concluded that in some cases, using the LSTM model can help merchants perform sentiment analysis on movie reviews.

References

[1] Sherstinsky A 2020 Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena*. 404(2020)132306

[2] Behera R K and Jena M and Rath S K and Misra S 2021 Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. *Information Processing & Management*. 58(2021)102435

[3] Chen L and Guan Z Y and He J H and Peng J Y A 2017 Survey on Sentiment Classification. Journal of Computer Research and Development. 54(2017)1150

[4] Joshi M and Prajapati P and Shaikh A and Vala V A 2017 Survey on Sentiment Analysis. International Journal of Computer Applications. 163(2017)34

[5] Lu W C and Xu J 2019 Sentiment Analysis of Network Users' Reviews Based on Bipartite Network. *Data Analysis and Knowledge Discovery*. 3(2019)10

[6] Zhong J W Liu W Wang S L Yang H 2021 Review of Methods and Applications of Text Sentiment Analysis. Data Analysis and Knowledge Discovery. 5(2021)1

[7] Dang N C Moreno-García M N De la Prieta F 2020 Sentiment Analysis Based on Deep Learning: A Comparative Study. *Electronics*. *9(2020)*483

DOI: 10.54254/2755-2721/40/20230651

[8] Landi F Baraldi L Cornia M Cucchiara R 2021 Working Memory Connections for LSTM. Neural Networks. 144(2021)334-341

[9] Finno W and Hergert F and Zimmermann H G 1993 Improving model selection by nonconvergent methods. Neural Networks. 6(1993)771

[10] Prechelt L 2002 Early Stopping-but when?. Neural networks: tricks of the trade. 1524(2002)55