2022 IEEE 5th Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC) | 978-1-6654-7968-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/IMCEC55388.2022.10019969

Sentiment Analysis of Movie Reviews Based on LSTM-Adaboost

Ling Zhang[1,2] , Miao Wang[3], Ming Liu[1,2] ,Haozhan Li[1,2]
1. State Key Laboratory of Media Convergence and Communication, Beijing, China
2. Communication University of China, Beijing, China
3. National Institute of Standardization, Beijing, China
zl_marzes@cuc.edu.cn, 18042921105@163.com, liu_ming@cuc.edu.cn, haozhanli@cuc.edu.cn
Corresponding Author: Ling Zhang Email: zl_marzes@cuc.edu.cn

***Abstract*—With the improvement of public living standards and the rapid development of the Internet, film works as cultural carriers have become an integral part of people's cultural and spiritual life. The booming movie industry has also given rise to a large number of user online reviews. How to effectively identify the positive and negative emotional tendencies of movie reviews is of great significance to the online word-of-mouth research and publicity marketing of movies. Based on the machine learning perspective of deep learning, this paper proposes a LSTM-Adaboost text sentiment analysis method which combines LSTM neural network and Adaboost boosting method. And the experiments are compared with CNN and LSTM models on the IMDB movie review dataset. The results show that the proposed method in this paper has higher accuracy compared with CNN and LSTM methods. In terms of sentiment classification accuracy, the LSTM-Adaboost method improves by 6 percentage points.**

Keywords—Sentiment Classification; Long Short Term-Memory network; Adaptive Boosting Algorithm; Deep Learning; Machine Learning

I. Introduction

With the rapid development of society and the advancement of material living standards, people's demand for a better spiritual culture is growing, and movies have gradually become an important cultural life consumption for the public. As a centralized place for audience feedback and opinions on the content consumption of movie works, the positive and negative sentiment distribution of reviews also reflects the depth of movie content quality to a certain extent. How to dig out valuable sentimental information from the huge volume of review texts is of great significance for grasping the movie's online word of mouth and subsequent promotion and marketing.

Sentiment analysis, also known as comment mining or opinion mining, is a web mining technique that analyzes sentiment or opinion based on published online comments [1]. It aims to extract text from reviews of certain products or services and classify them as positive or negative according to the polarity of the reviews [2-3]. At present, there are two main methods for sentiment classification of comment corpus, one is the text sentiment analysis method based on sentiment dictionary, and the other is the method based on machine learning. The sentiment analysis method based on sentiment dictionaries mainly uses manually hand-written discourse-sentiment dictionary templates to obtain the sentiment polarity of the final text. The classification accuracy of this method is influenced by the quality of the sentiment lexicon. The machine learning based approach focuses on selecting the appropriate features to represent the text. Firstly, the sentiment category of the text data used for training is manually labeled, and then the feature vector is extracted by machine learning method, and the sentiment classifier is constructed for training, and finally the sentiment polarity of the text is obtained. The research based on the former is mainly based on natural language processing, and there are many research difficulties that have not yet been solved. Most of the existing research relies on machine learning methods such as decision trees, naive Bayes, support vector machines, and neural networks.

In this paper, we study deep learning-based sentiment analysis methods, combining long and short-term memory network (LSTM) with adaptive boosting algorithm (Adaboost), and propose a text sentiment analysis method based on LSTM-Adaboost. The IMDB movie review dataset is selected as the experimental object, and conclusions are drawn by comparing the results of CNN and LSTM on performance indicators such as accuracy and precision. The structure is organized as follows: Section 2 briefly describes the neural network model used in this paper. Section 3 establishes LSTM-Adaboost algorithm and performs case study. Finally, Section 4 concludes the paper and points out further work.

II. Related work

A. Long Short Term-Memory Network

As a special type of RNN (Recurrent Neural Network), LSTM (Long Short-Term Memory Network) is a deep neural network proposed by Hochreiter and Schmidhuber to solve the long-term dependency problem faced by RNN [4]. The advantage of RNN is that it will not only learn the information of the current moment, but also rely on the previous information. However, with the complexity of the network structure, the RNN model is calculated and backpropagated during training, the gradient tends to increase or decrease at each moment, and the gradient after many stages of propagation will diverge to infinity or converge to zero. The problem of the exploding or disappearing gradients makes RNNs lose the ability to learn connected remote information. LSTM replaces the modules in the hidden layer of RNN by memory units of cells, while introducing input gates, output gates, and forget gates makes this network remember and update the information of long processes, thereby optimizing the long-term dependency problem [5-6].The structure of the hidden layer of LSTM is shown in Fig 1.

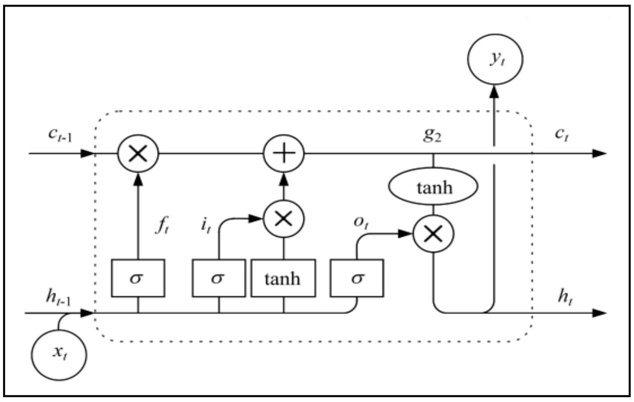


Fig. 1. LSTM neural network structure diagram

The mechanism of the LSTM neural network is as follows:

(1) The LSTM uses a "forget gate" to determine the forgetting or retention of information in the memory cell state. The "forget gate" controls the weight of the "forget gate" between [0-1] by reading the output vector of the previous moment h_{t-1} and the input vector of the current moment x_t , combined with a sigmoid function.

$$f_t=\sigma\left(W_f \cdot [h_{t-1},x_t]+b_f\right) \tag{1}$$

(2) The LSTM uses an "input gate" to determine the new information that needs to be retained in the memory cell state. On the one hand, the sigmoid function determines the information that needs to be updated, and on the other hand, the tanh function creates a vector of alternative values $\widetilde{C}_t$ and adds them to the memory cell state.

$$i_t=\sigma\left(W_i \cdot [h_{t-1},x_t]+b_i\right) \tag{2}$$

$$\widetilde{C}_t=\tanh\left(W_c \cdot [h_{t-1},x_t]+b_c\right) \tag{3}$$

(3) The two parts of information from the previous step are combined to update the states of the memory cells. The old state C_{t-1} is multiplied with the forgotten one f_t to determine the information to be discarded, and the new information $\widetilde{C}_t \cdot i_t$ to be remembered is added to obtain the new candidate values, that is, the updated values are scaled according to the decision of each state.

$$C_t=f_t \cdot C_{t-1} + \widetilde{C}_t \cdot i_t \tag{4}$$

(4) The LSTM uses an "output gate" to determine the value of the output information. The "output gate" first determines the information to be output in the memory cell state based on the sigmoid function, and then adjusts the cell state input tanh function to a value between [-1, 1]. The output of the two components is multiplied to determine the final output vector h_t at the current moment .

$$o_t=\sigma\left(W_0 \cdot [h_{t-1},x_t]+b_0\right) \tag{5}$$

$$h_t=o_t \cdot \tanh(C_t) \tag{6}$$

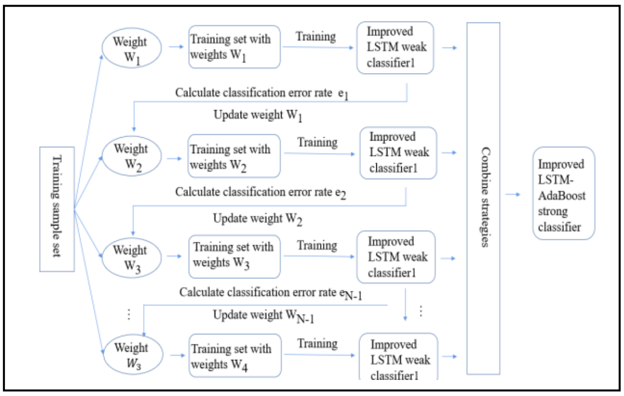
In the above equation, W_f, W_i, W_0 and W_c are the weight vectors of the forget gate, the input gate, the output gate, and the input cell state, respectively, and b_f, b_i ,b_0 and b_c are the corresponding bias vectors.

Considering the advantages of LSTM model in capturing the sequence features and contextual semantic dependencies of text, LSTM is selected as the base classifier for subsequent integrated learning enhancement in this paper.

III. PROPOSED METHOD

A. *LSTM-Adaboost Algorithm*

To improve the accuracy of LSTM model for movie review sentiment classification, this paper proposes a text sentiment classification method based on LSTM-Adaboost, a boosting method applied to binary classification problems proposed by Freund and Schapire. Its basic idea is: for a complex task, the judgment obtained by appropriate synthesis of the judgments of multiple experts is usually better than the judgment of any one of them alone [7]. Since the labels of the text data used for experiments in this paper are divided into two categories, positive and negative, it is a binary classification problem. Therefore, the Adaboost algorithm can be used to integrate the learning of multiple LSTM weak classifiers and update the weights of each training sample by multiple iterations, and combine the strategies to form an improved LSTM-Adaboost strong classifier to improve the accuracy of classifying text sentiment. The overall framework of the LSTM-Adaboost method is shown in Fig 2.



Training sample set
Weight W_1
Training set with weights W_1
Training
Improved LSTM weak classifier1
Calculate classification error rate e_1
Update weight W_1
Weight W_2
Training set with weights W_2
Training
Improved LSTM weak classifier1
Calculate classification error rate e_2
Update weight W_2
Weight W_3
Training set with weights W_3
Training
Improved LSTM weak classifier1
Calculate classification error rate e_{N-1}
Update weight W_{N-1}
Weight W_3
Training set with weights W_4
Training
Improved LSTM weak classifier1
Combine strategies
Improved LSTM-AdaBoost strong classifier

Fig. 2. LSTM-Adaboost method

As a common boosting algorithm, Adaboost has the advantage of paying more attention to the misclassified sample data and emphasizing the weak classifiers with good performance, so that the strong classifiers obtained by integrating Adaboost tend to have higher classification performance. The specific steps of the Adaboost algorithm for boosting LSTM models are given below.

(1) Input sample training set $S=\{(x_1,y_1),\cdots,(x_i,y_i),\cdots,(x_N,y_N)\}$, where x_i is the training data, $y_i \in Y=\{0,1\}$ is the corresponding prediction data. 0 represents negative comments, 1 represents positive comments, and N represents the number of samples.

(2) Initialize the weight distribution of the sample training data.

$$D_1=(w_{11},\cdots,w_{1i},\cdots,w_{1N}),\ w_{1i}=\frac{1}{N},\ i=1,2,\cdots,N \qquad (7)$$

(3) For m=1,2, ···,M ,

① Use the sample dataset with weight distribution for training to get the basic classifier based on the LSTM model: $G_m(x)$: S→Y.

② Calculate the classification error rate on the training dataset.

$$e_m=\sum_{i=1}^{N} P(G_m(x_i)\neq y_i) \qquad (8)$$

③ Calculate the coefficients of the LSTM weak classifier.

$$\alpha_m=\frac{1}{2}\log\frac{1-e_m}{e_m} \qquad (9)$$

④ Update the weight distribution of the training data set.

$$D_{m+1}=(w_{m+1,1},\cdots,w_{m+1,N}) \qquad (10)$$

$$w_{m+1,i}=\frac{w_{mi}}{Z_m}\exp\left(-\alpha_m y_i G_m(x_i)\right),i=1,2,\cdots,N \qquad (11)$$

$$Z_m=\sum_{i=1}^{N} w_{mi}\exp\left(-\alpha_m y_i G_m(x_i)\right) \qquad (12)$$

(3) Integrate the trained LSTM weak classifiers by the corresponding weights of the classifiers.

$$f(x)=\sum_{m=1}^{M}\alpha_m G_m(x) \qquad (13)$$

The final classifier is obtained as follows:

$$G_m(x)=\text{sign}(f(x))=\text{sign}\left(\sum_{m=1}^{M}\alpha_m G_m(x)\right) \qquad (14)$$

B. Datasets and Variable Descriptions

This paper selects the IMDB movie review dataset as the experimental object. The IMDB dataset contains a total of 50,000 movie review texts, which are divided into 25,000 training data and 25,000 test data. Each movie review is marked as "positive review" or "negative review" .

TABLE I. COMMENTS DATASET DESCRIPTION

Type of Data	Training data	*Test Data*
Amount	25000	25000
Label	Positive comments: 1 Negative comments: 0	Positive comments: 1 Negative comments: 0

The text length distribution of the text data was analyzed by performing descriptive statistics on the training data and plotting the text length distribution and the cumulative probability distribution of the IMDB comment data. As can be seen from Fig 3, the shortest text length of the training data is 11, the longest text length is 2494, and the average text length is 239.

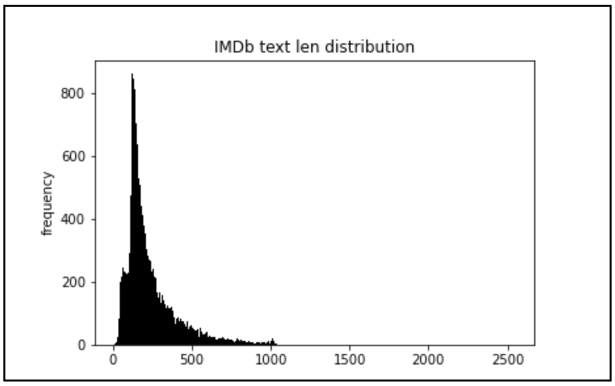


Fig. 3. Text length distribution of IMDB comment training data

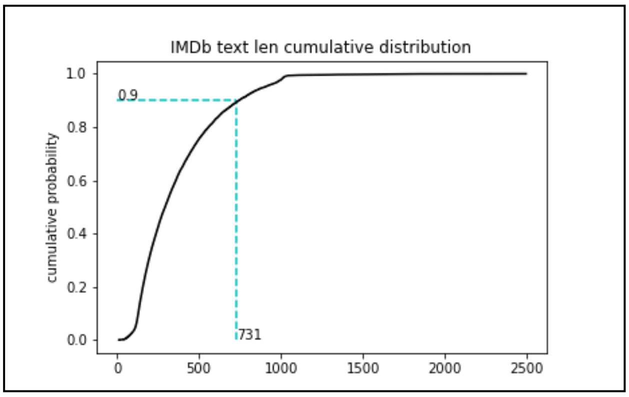


Fig. 4. Cumulative probability distribution of text length for IMDB review training data

As shown by the cumulative probability distribution of the text data, the corresponding quantile text length is 731 when the cumulative probability is taken as 0.9, namely, 90% of the text length is within 731. Given that the construction of the LSTM sentiment classifier requires setting the dimensionality of the input variables, which directly affects the calculation volume and time consuming of the model training. Based on the distribution of text lengths, this study sets the dimensionality of the input variables of the LSTM model to 800, that is, the top 800 words with the highest frequency of occurrence in the comment text are retained.

C. *Experiments and Analysis of Results*

Based on the IMDB review dataset, the LSTM-Adaboost model proposed in this paper is compared with the classical CNN neural network and the original LSTM network model. The parameter configurations of the LSTM-Adaboost model are given in Table 2.

TABLE II. LSTM-ADABOOST MODEL PARAMETER SETTINGS

Parameter	**Model**
Embed Size	64
LSTM Size	[128,64,32]
Dropout	0.2
Epoch	10
Batch Size	256
Number of Integrated Models	10

In order to further verify the effectiveness of the LSTM-Adaboost model proposed in this paper, accuracy, precision, recall and F1 score are used as the evaluation indexes of the classifier in this paper. The specific formulas are as follows.

$$\text{Accuracy}=\frac{\text{TP+FN}}{\text{TP+TN+FP+TN}} \tag{15}$$

$$\text{Precision}=\frac{\text{TP}}{\text{TP+FP}} \tag{16}$$

$$\text{Recall}=\frac{\text{TP}}{\text{TP+FN}} \tag{17}$$

$$\text{F}_1=2\cdot\frac{\text{Precision}\cdot\text{R ecall}}{\text{Precision+Recall}} \tag{18}$$

where TP is true positives, TN is true negtives, FN is false negtives, and FP is false positives.

The accuracy rate examines the percentage of correctly classified samples; the precision rate and recall rate are more concerned with the classifier's ability to recognize positive sentiment text, with the former targeting predicted classification results and the latter focusing on all positive sentiment samples; the F1 score takes into account the precision rate and recall rate of the classification model, and is a harmonic mean of the two. These four evaluation metrics are all positive metrics, that is, the higher the value of the metric, the better the performance of the model classification.

The final experimental results of the three methods for sentiment classification are shown in Fig 5. It can be seen that all three methods have good sentiment classification results for IMDB reviews; the LSTM model has the worst classification effect and the LSTM-Adaboost model has the best classification effect. In terms of specific metrics, it is clear that the prediction ability of the LSTM enhanced by Adaboost is further improved, and the four evaluation index values of accuracy, precision, recall and F1 score are 2.7, 6, 0.4 and 3.2 percentage points higher than the original LSTM model, respectively. Through the integrated learning of Adaboost, the accuracy rate, that is, the recognition ability of positive comments, is effectively improved. In general, the LSTM-Adaboost model proposed in this paper outperforms its alternatives on this dataset.

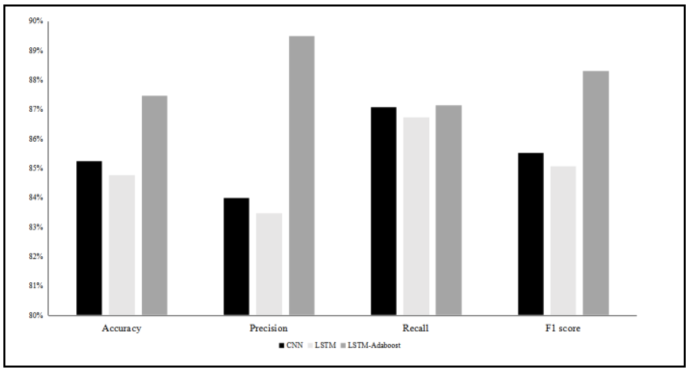


Fig. 5. Performance comparison of 3 classification models

IV. CONCLUSION

Neural network is a common machine learning method in the field of text sentiment analysis. Therefore, this paper proposes an integrated LSTM-Adaboost learning sentiment classification method by combining LSTM neural network and Adaboost algorithm to improve the LSTM model. And the IMDB movie review dataset is used as an example to compare with CNN and LSTM models experimentally. The results show that the LSTM-Adaboost model has the best performance among the three, obtaining up to 87.47% accuracy, 89.51% precision, 84.9% recall and 87.14% F1 score, which means the sentiment analysis method based on the LSTM-Adaboost model can do the sentiment analysis more accurately. Given that the accuracy as well as the number of integrated base classifiers themselves have a great influence on the enhancement effect after integration, their relationship can be further explored in the subsequent research to obtain a better base classifier and parameter selection method.

Acknowledgment

This paper was financially supported by the Beijing Municipal Natural Science Foundation (9202018), the Fundamental Research Funds for the Central Universities (CUC21GZ005), and the Academic Research Project of China Federation of Radio and Television Association (2020ZGLH015).

References

[1] C. Költringer and A. Dickinger, "Analyzing destination branding and image from online sources: A web content mining approach", Journal of Business Research, vol. 68, no. 9, pp. 1836–1843, 2015. doi: https://doi.org/10.1016/j.jbusres.2015.01.011.

[2] E. Cambria, B. Schuller, Y. Xia and C. Havasi, "New Avenues in Opinion Mining and Sentiment Analysis," in IEEE Intelligent Systems, vol. 28, no. 2, pp. 15-21, March-April 2013, doi: 10.1109/MIS.2013.30.

[3] L. V. Casaló, C. Flavián, M. Guinalíu, and Y. Ekinci, "Avoiding the dark side of positive online consumer reviews: Enhancing reviews' usefulness for high risk-averse travelers", Journal of Business Research, vol. 68, no. 9, pp. 1829–1835, 2015, doi: https://doi.org/10.1016/j.jbusres.2015.01.010.

[4] Y. Xie, R. Liang, Z. Liang, C. Huang, C. Zou and B. Schuller, "Speech Emotion Classification Using Attention-Based LSTM," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 11, pp. 1675-1685, Nov. 2019, doi: 10.1109/TASLP.2019.2925934.

[5] A. Graves, "Long Short-Term Memory", in *Supervised Sequence* Labelling with Recurrent Neural Networks, Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 37–45. doi: 10.1007/978-3-642-24797-2_4.

[6] F. A. Gers, J. Schmidhuber and F. Cummins, "Learning to Forget: Continual Prediction with LSTM," in Neural Computation, vol. 12, no. 10, pp. 2451-2471, October. 2000, doi: 10.1162/089976600300015015.

[7] Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119–139, 1997, doi: https://doi.org/10.1006/jcss.1997.1504.