Advances in Social Science, Education and Humanities Research, volume 412
International Conference on Modern Educational Technology and Innovation and Entrepreneurship (ICMETIE 2020)

Sentiment Analysis of Chinese Commodity Reviews Based on Deep Learning

Zifan Zhang

College of software, Taiyuan university of technology, Taiyuan, Shanxi, 030024, China

Corresponding E-mail:angela@cas-harbour.org

Keywords: sentiment analysis, deep learning, Chinese, LSTM, RNN

Abstract: Recent development of the Internet led to the emergence of product reviews with strong emotions on the e-commerce shopping platform. These reviews have become the main channel for people to know about the products. Sentiment analysis, a branch of natural language processing (NLP), is used to evaluate the emotion type described in the text. The establishment of individual emotional marker model allows us to identify the emotional characteristics of sentences based on the deep learning framework, thus obtaining accurate result of sentiment analysis. Results from Tensorflow deep learning framework, RNN and LSTM models are analyzed and compared, showing that the LSTM model has a better performance in the experiment as a guidance for the optimization of related sentiment analysis model.

1. Introduction

Recent development of the 4G network and the popularization of mobile Internet have turned emotional analysis based on social network and other media into a hotspot[1] as an increasing number of consumers re shopping online[2] and are encouraged to comment on products. Product reviews are very important to all businesses. By analyzing product reviews, businesses can classify and understand consumers' opinions to improve their products based on the needs of consumers. Consumers can also easily view the comments by category. Because manual analysis is time-consuming and labor-consuming, it is necessary to carry out sentiment analysis by deep learning.

Sentiment analysis, also known as opinion mining[3], is a branch of NLP of artificial intelligence technology, aiming at establishing the emotion embedded in the texts, namely positive, negative or neutral. For sentiment analysis, a large number of related emotion classification methods have emerged, such as PDTurney[4] and Taboada[5], etc. Pang and Lee applied the traditional algorithm of machine learning to sentiment analysis, where the word bag model was adopted for sentiment analysis of texts in 2002 and 2011 respectively.

With the progress of artificial neural network and researches on deep learning, there have been a number of achievements, including the RNN (recurrent neural network) model proposed by Mikolov[6] and others in 2010 that allows a full use of context information, and sentence semantic features obtained by Socher [7] and others through improved recursive neural network model, bringing more accurate emotion classification.

This paper attempts to study user sentiments embedded in commodity reviews on the e-commerce platform through the natural language processing based on the LSTM deep learning neural network model to assist evaluation of commodity classification of and prove the effectiveness of the LSTM model in handling sentiment analysis by comparing the accuracy of the RNN neural network and LSTM neural network.

2. Theory and technology of Chinese natural language processing

For text-based product reviews, natural language processing (NLP) is needed in the process of computer processing. Natural language processing is an important branch of computer science and artificial intelligence that studies various theories and methods facilitating communication between human and computer with natural language. In a word, NLP is a technology that enables computers to understand human languages.

2.1 Jieba

NLP is based on word segmentation. Word is the smallest meaningful language component that can be independently used. Unlike western languages, Chinese is a word unit without boundary between words such as spaces. In Chinese natural language processing, word segmentation is the basic unit of Chinese text processing. The quality of word segmentation plays a key role in Chinese information processing, in which Jieba is the most commonly tool for this purpose.

Jieba is an excellent third-party Python library for Chinese word segmentation that supports accurate cutting of sentences in text analysis. In this experiment, Chinese sentences can be divided into different parts according to the meaning of words while English sentences can be divided into words according to the spaces.

2.2 Keras

Keras is a high-level neural network API written in Python, which can run as a backend with tensorflow, cntk, or theano. The stunning feature of speed of provides solid supports for fast operation, allowing transformation of idea into experimental result with the minimum delay. Keras allows simple and fast prototyping (thanks to features such as user friendliness, high modularity, and scalability).

2.3 Recurrent Neural Network Model

The concept of deep learning comes from the artificial neural network model. In-depth learning is a combination of low-level features into a superior and more abstract representation of distributed data. The structure of traditional neural networks is relatively simple: input layer - hidden layer - output layer, allowing only one input after another while the former input has nothing to do with the latter. However, some tasks have a greater demand for sequence information processing, where the previous input is usually related to the later one. For example, understanding the meaning of each single word is not enough to capture the meaning of the whole sentence. We need to deal with the whole sequence connected by these words. We need to use another kind of important neural network in the field of deep learning: recurrent neural network.

The reason why RNN can effectively process sequence data is mainly based on its special operation principle. RNN is a kind of basic multilayer feedback neural network, in which the nodes are connected in a ring, with its internal state being indicated by the dynamic time series. Compared with traditional neural network, RNN distinguishes itself as it will bring the previous output to the next hidden layer and train together. The network has a strong internal memory allowing the processing of any input sequence, making it an effective tool in natural language processing.

In the field of NLP, RNN is more suitable for context-based language processing. In natural language processing, word order is very important as the adjacent words in sentences are not independent. Different permutations and combinations form different meanings. For example, "SUVs perform better than cars." "Cars perform better than SUVs.". Just by changing the order, the meaning will be completely different. The output of the current time in RNN is related not only to the output of the previous layer, but also to the output of the last hidden layer. Though a very powerful dynamic system, RNN has some short-term memory problems that may have a greater impact than long-term factors. In the stage of gradient back propagation, the weight of the state transition matrix has a great influence on the learning process. Specifically, the weight of the state transition matrix has a positive correlation with the gradient signal, where the reduced weight of matrix would lead to weaker gradient signal while the increase of matrix weight would deviate the gradient signal, causing the so-called gradient explosion. Therefore, RNN is not capable of handling the long-term dependence of a long input sequence in practice[8].

2.4 Long Short-Term Memory Model

In order to cope with the difficulty of RNN in dealing with long-term dependence caused by gradient disappearance / gradient explosion in practice, the LSTM network with special implicit units was proposed by Hochreiter and schmidhuber (1997). The long-term and short-term memory network, commonly known as LSTM (long and short term memory), is a special RNN that can learn long-term dependence through an exquisite design and store the influence of long-distance context on the current output.

3. Implementation Process of User Sentiment Analysis Based on LSTM

In order to compare the accuracy of LSTM and RNN deep neural network and explore the factors affecting LSTM's accuracy in sentiment analysis, some comparative experiments are designed here. The experimental data comes from the users' comments on different products in real scenarios collected from Taobao, the largest e-commerce platform in the Asia Pacific Region. The LSTM and RNN network models are used for training, learning and prediction respectively, thus revealing the advantages and disadvantages of these two models.

3.1 Introduction to Running Environment

The experiment of this paper is run on high-performance PC mainly based on Keras and Tensorflow which both are deep learning platform. The specific configuration is shown in Table 1.

Table 1 specific configuration of experimental environment

Software name	Software type
Windows 10	Operating System
Python 3.7.0	Programming language
Anaconda	Management tool
Jupyter Notebook	Program running environment
Keras	Python deep learning library
TensorFlow	Python deep learning library
Pandas	Python data analysis module
Numpy	Python Scientific Computing Library
jieba	Natural language processing tools

3.2 Data Set

This paper collects real users' comments on different products posted on the comment page, where each line forms a sentence. The goal of this experiment is to train by building a model that carries out corresponding tests. Each sentence in the test set is classified into "positive" and "negative" sentiment analysis. There are 21100 sentences in total in the dataset, each sentence being marked as "positive" and "negative" respectively. Around 90% of the data is used as the training set while the rest 10% as the test set.

3.3 Training and Prediction of Emotion Analysis

Deep learning requires input data as a vector when training so that matrix operation can be carried out. And it is also the source of multi-layer perceptron. As it is impossible for computers to recognize a group of sentences directly the most important step is to convert words into word vectors, which is usually achieved through word embedding as is used in this experiment. The experiment is basically divided into the following steps:

1. The minimum statistical granularity of Chinese is words, so Jieba is used to segment the sentences into words.
2. Gather the segmented words and calculate their frequency of usage. Sort them out based on the frequency and then index them.
3. Replace all words in the sentences with corresponding indexes, so that each word in a sentence can be represented by a number.
4. The first layer of Keras "embedding()" is used to convert the sentence number arrays into word vectors so that the program can understand the meaning.
5. Train the dataset using Keras, including defining training rounds, setting activation function, adding verification set and so on. Then LSTM is used for the same purpose. After training, the

model is saved.

The specific process is shown in Figure 1.

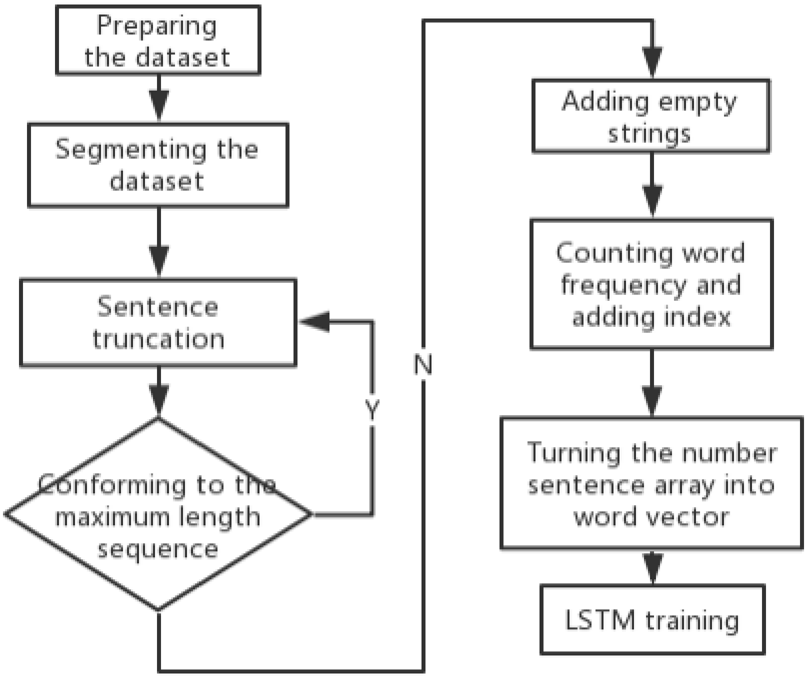

Figure 1 Sentiment analysis flow of LSTM network model

3.4 Comparative experiments

3.4.1 Comparison of LSTM model in Chinese and English sentiment analysis

In order to compare the accuracy of Chinese and English analysis under the same model, the segmentation model in 3.3 is applied to English commodity reviews for comparison. The dataset of English product reviews come from Amazon [9], among which the 5-star reviews in the dataset are marked positive and 1-star reviews as negative, which are segmented in the form of space. The same number of Chinese dataset are screened out for preparation.

3.4.2 The effect of word segmentation on Chinese emotion analysis

The scale of dataset used for training in this experiment is small, with possible errors when analyzing the relationship between texts. Therefore, a comparative experiment is set up to test the accuracy of the two models. Chinese word splitting is to divide a sentence into individual words.

3.4.3 The effect of different maximum training set length on the result

Firstly, the length and frequency distribution of all data in training set are counted, with the statistical results being shown in Figure 2. In this experiment, three values were selected to test the effect of maximum length of different training set on the experimental results. The three selected lengths include: 38 (the length of the sentence with the highest frequency), 140 (the length of the sentence containing most data) and 400 (the length of the dataset containing almost all the data). In this experiment, Chinese dataset was used.

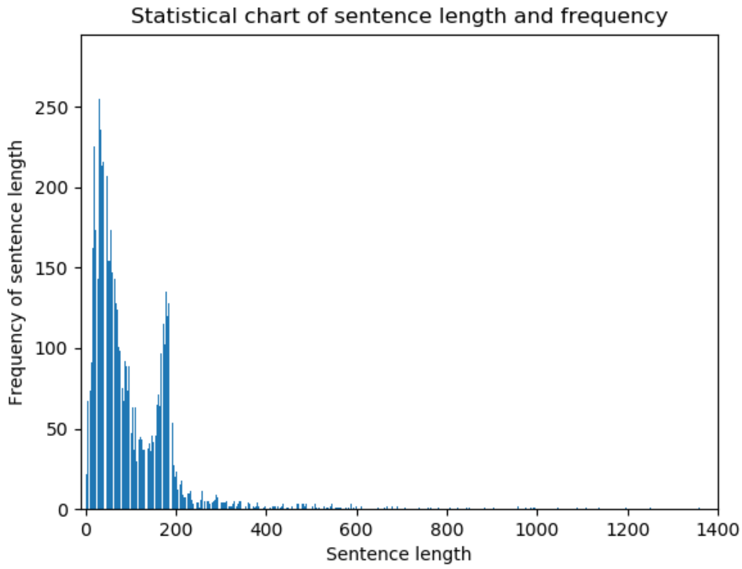

Figure 2. Sentence length and frequency

3.4.4 RNN Network Model

In order to further verify the effectiveness of the LSTM model for sentiment analysis, the paper also applied RNN model in sentiment analysis and classification of the same dataset.

4. Experimental Results and Analysis

Following the above experimental steps and methods, different results of three experiments are obtained and the experimental results are shown in Table 2:

Table 2. Experimental results

Experimental name	Accuracy(in same epochs)
LSTM Chinese dataset (sentence_max_length=100)	92.81%
LSTM English dataset(sentence_max_length=100)	84.33%
LSTM Chinese dataset with a maximum length of 38	91.90 %
LSTM Chinese dataset with a maximum length of 140	90.81 %
LSTM Chinese dataset with maximum length of 400	51.60%
LSTM Chinese dataset (word splitting) (sentence_max_length=100)	93.94%
RNN Chinese dataset(sentence_max_length=100)	86.85%

Under the same model, the accuracy of the Chinese dataset is 92.81% while that of the English dataset is %. The experimental results show that the same model has different accuracy for different languages, implying the need to select data processing method based on the specific language.

In Experiment 2, different maximum length leads to different training speed. The greater the maximum length, the longer the training time. However, dataset of shorter length has no obvious impact on the results. When the length reaches up to 400, the training speed is extremely slow, with an accuracy half of the rest under the same condition.. Therefore, when choosing the maximum sentence length, minimum length is preferred that covers most data.

In the results of Experiment 3, the training time of single word segmentation is twice as long as that of the words segmentation model. In the model of single word segmentation, the loss value does not fall as fast as the word segmentation, and the accuracy is slightly higher. In the case of small dataset, the single word segmentation is recommended for data processing.

The results of Experiment 4 show that LSTM model has higher accuracy than RNN model and is more suitable for long sequence text analysis.

5. Conclusion and Suggestion

Through the above experimental verification and analysis, it is concluded that the LSTM deep neural network model enjoys greater accuracy than RNN network model in sentiment analysis classification and prediction.

Different LSTM model parameters will lead to different experimental results. In the experiment, parameters should be adjusted constantly to achieve the best effect.

Different data processing approaches will affect the results while data preprocessing can be used to improve the accuracy. Stop word refers to words that have no obvious influences on semantics in the texts. , these stop words can be removed when word segmentation is working to improve the performance of text classification. However, the small scale of dataset used for training makes it difficult for the paper to deal with the useless words , which might lower the accuracy of analysis of Chinese texts. Therefore, in the following experiments, the dataset could be further processed by removing useless information, such as stop words, punctuation, special characters, escape characters and so on. At the same time, the scale of datasets should be expanded and manual audit of datasets could be carried out to make it more accurate, thus ensuring that all datasets come from the same distribution and are reliable.

Finally, the expansion of network capacity and complexity of the neural network model without over fitting will be the focus of next stage of researches to further improve its accuracy.

References

[1] Blazyte, A. (2019) Alibaba: cumulative active online buyers Q2 2014-Q2 2019. Accessed time: Sep 27, 2019 https://www.statista.com/statistics/226927/alibaba-cumulative-active-online-buyers-taobao-tmall/

[2] Tan, P. N., Steinbach, M., Kumar, V. (2006) Introduction to data mining. Pearson Addison Wesley. Boston.

[3] Pang, B., Lee, L. (2008) Opinion mining and sentiment analysis. J. Foundations and Trends in Information Retrieval, 2(1-2):1−135.

[4] Liu, L. (2012) Sentiment Analysis and Opinion Mining(Introduction and Survey. In: Graeme H. (Eds.), Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

[5] Pang, B., Lee, L., Vaithynanthan, S. (2002) Thumbs up? Sentiment classification using machine learning techniques. In: Proceedings of the conference on Empirical Method in Natural Language Processing. Philadelphia. 79-86.

[6] Pang, B., Lee, L (2004) A sentimental education: Sentiment analysis using subjectively summarization based on minimum cuts. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA. 271-278.

[7] Turney, P. D. (2002) Thumbs up or thumbs down? Semantic Orientation Applied to Unsupervised Classification of Reviews. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics(ACL'02). Philadelphia. 417-424.

[8] Taboada, M., Brooke, J., Tofiloski, M., et al. (2011) Lexicon-based methods for sentiment analysis. J.Computational linguistics, 37(2): 267 -307.

[9] Bengio Y, Ducharme R, Vincent P, et al. (2003) A neural probabilistic language mode. J. Journal of Machine Learning Research, 3(6):1137-1155.

[10] Mikolov, T., Karafiát, M., Burget, L., et al. (2010) Recurrent neural network based language model. In: INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari.

[11] Socher, R., Perelygin, A., Wu, J. Y,. et al. (2013) Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Seattle. 1631-1642.

[12] Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient Flow in Recurrent Nets: The Difficulty of Learning LongTerm Dependencies. In: John F. Kolen, Stefan C. Kremer (Eds.), A Field Guide to Dynamical Recurrent Networks. Wiley-IEEE Press, 237-243.

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

[14] Cornell University.(2013) Amazon reviews - Full[DB/OL]. https://s3.amazonaws.com/fast-ai-nlp/amazon_review_full_csv.tgz