A Practical Application for Text-Based Sentiment Analysis Based on Bayes-LSTM Model

A Practical Application for Text-Based Sentiment Analysis Based on Bayes-LSTM Model

Jiawen Li and Huaping Zhu*

Department of Electronics and Information Engineering, Tongji Zhejiang College, Jiaxing, China
Email: huapingzhu@tongji.edu.cn

Abstract. Text-based sentiment analysis algorithms have now become one of the active research areas in emotional analysis which has gained much attention nowadays. Text emotion classification can be widely used in social public opinion analysis, product use feedback, harmful information filtering, etc. In this paper, we first developed a robotic crawler to gather data about comment on Huawei cellphone from Sina weibo microblog sites (Chinese twitter). Then we generate the data text to be trained according to the input requirements of the Keras module, and perform formal training and learning on the model after data preprocessing. Subsequently, the classifier was constructed based on the Bayes-LSTM model in which TF-IDF model was used for feature selection. The LSTM model can be characterized by the ability to self-evaluate the usefulness of the information obtained, which makes up for the shortcoming of naive Bayes formula that only applies to two independent events. We finally have a practical application that generates a word cloud from text, showing frequently used words in larger font sizes, effectiveness of the algorithm was also verified by experiment.
Keywords. Chinese text; sentiment analysis; Naive Bayes; LSTM model.

1. Introduction

The unstructured or semi-structured comments on the Internet, such as product comments, news comments, stock comments, etc., are the main carriers for internet users to express their opinions, attitude feelings and emotions. Text classification was initially classified by texts themes according to political, economic, military, sports and others. Currently, the text analysing focuses on analysing the emotion reflected in the text and making category judgments on the emotional tendency of the text. Researches on textual emotion categories usually classifies 2 categories (i.e., positive, negative) or 3 categories (i.e., positive, neutral and negative), and more research focuses on two categories [1]. Text emotion classification can be widely used in social public opinion analysis, product use feedback, harmful information filtering, etc.

Statistical method is the mainstream technology of text emotion classification. In this research, Naive Bayes, maximum entropy model and support vector machine were adopted for emotion classification on English file comment [2]. Another research [3] conducted an emotion classification study on news and comment corpus using Naive Bayes and maximum entropy model. As a statistical separator, the Naive Bayes is one of the easy-to-understand classification algorithms in text classification. Otherwise, it seems that the Bayes classifier has a good classification effect and performance for processing large-scale data [4, 5].

As a widely used RNN (recurrent neural network) model, Long Short-Term Memory (LSTM) has been designed to address the vanishing and exploding gradient problems of conventional RNNs [6]. Because of LSTM is powerful for modelling sequences, i.e., language modelling,

RNNs have cyclic connections making them powerful for modelling sequences. They have been successfully used for sequence labelling and sequence prediction tasks, such as handwriting recognition, language modelling, novel LSTM based RNN architectures also used effectively in text classification. The schematic diagram of LSTM is shown in figure 1, the split line will be replaced with the main line according to the importance of the input word before which were analysed. Forgetting control and input control determines the update of the main line, while the main line and the split line together determine the output control.

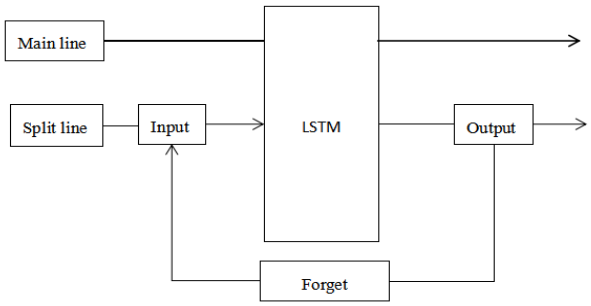

Figure 1. LSTM model schematic.

In this paper, we used a novel LSTM based RNN architectures which make more effective use of model parameters to train text analysed models for large vocabulary text classification. We trained Bayes-LSTM models in which TF-IDF model was used for feature selection. We show that LSTM models converge quickly and showing frequently used words by the word cloud from weibo comments.

2. Sentiment Analysis Based on Bayes-LSTM

2.1. Data Collection

Date were crawled more than 1000 comments about Huawei products from Weibo sites, such as, comment text, comment time, and the number of thumbs ups. The data with CSV format were acquired in a short time for the relatively small data. In order to remove and modify CSV data which are incorrect, incomplete, irrelevant, duplicated, or improperly formatted, we wrote python programs for data cleaning. The classification model we use in this paper is the naive Bayesian model which is supervised learning method. All of the training text data in our model, 2 categories (i.e., positive emotion and negative emotion) were used for text emotion classification, positive emotions marked 1 and negative emotions marked 0 respectively. As shown in figure 2, some text data were manual marked which include some fields such as comments, post time and the number of thumbs ups.

评论内容	点赞数	时间	sentiment
华为发布P30都给我起来high	296	3月26日 2	1
一直在用华为产品，支持	139	3月26日 2	1
正在用p20pro的同学路过～	67	3月26日 2	1
这个是不是一摔就碎	5	3月26日 2	0
还是4鸡？	3	3月26日 2	0
这么贵！很难入手啊！华为化	3	今天 00:0	0
要钱就不考虑	2	3月26日 2	0

Figure 2. Some text data were manual marked (This picture lists a few Chinese comments and the number of likes, time and emotional value of other people.)

2.2. Text Data Processing

As the text data stored in CSV format, the Pandas module in Python was used to read the text for the next judgement step of the correct processing. Due to the operation of feature vectorization of text data requires Chinese word segmentation of comment text, the Jieba module was used in our program that can support three different word segmentation modes: precise mode for text analysis, full mode and search engine mode for search engines. Accordingly, we used the space notation as a separator between each word with the precise text segmentation model. Two classes of data (i.e., training sets and test sets) were then collected in which Three-quarters of the text data were classified as training sets [7]. After a successful segmentation, the number of words was manually entered for each comment text in order to ensure that the resulting matrix is of the same size. The number of comments in this paper was set to 100 words limited by that Weibo comments is less than 120 words. To solve the problem of feature vectorization for Chinese text, this paper adopted TF-IDF model to select text features. In order to the same length of each text data, there are enough space needs to be added in each comment data, that is, each text data made up of 100 words. Finally, the data text to be trained is generated according to the input requirements of Keras module, and the formal training and learning of the model will be done after the data pretreatment.

2.3. Model Training

Firstly, the Bayesian model was used to classify the training set in which the frequency P (Ai) of each category get in the previous stage was calculated separately [8]. Similarly, the Classification of the characteristics of each sample was also calculated: the conditional probability and the emotions of the text are initially classified as 0 and 1. Bayesian model does not need many times of iterative training and has good classification efficiency for big data, it is best to be used as a model for the first classification. As an improved model of RNN, three controllers, such as input control, output control and forgetting control were added in the LSTM model. Further, the LSTM model can be characterized by the ability to self-evaluate the usefulness of the information obtained, which makes up for the shortcoming of naive Bayes formula that only applies to two independent events. Therefore, LSTM was used in this paper for secondary classification. For the connection between the two models, a for-loop structure was used to read each layer along with the new model reconstructed layer by layer.

As shown in figure 3, the text classified in the previous step was firstly converted to a word vector matrix using the word embedding model, which was input into the initialized LSTM model. The final new training model that according to the emotional predisposition values of the first classification was built. Similar to the previous research [9], the optimal parameters were obtained by neural network model training after more than 30 iterations.

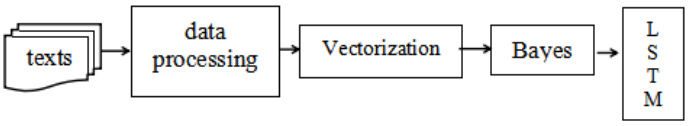

Figure 3. Naive Bayes-LSTM model sentiment analysis flow chart.

2.4. Experimental Result

After cross-validation of training set and test set with LSTM model, the accuracy of training set and test set is 94.62% and 83.10% respectively. As for the corpus of public comments on Huawei, the experimental results clearly show that the positive sentiment of reviewers is 57.77%, which represents that the overall public maintains a relatively positive opinion of Huawei brand. The detailed classification results were shown in table 1.

As shown in figure 4, the word cloud graph was generated using raw text data, in which the topic vocabulary of these comments includes Huawei brand, mobile and smart glasses. An example of a classification of the emotion analysis model is shown in figure 5.

Table 1. Confusion matrix for sentiment analysis.

	Actual positives	Actual negatives
Predicted positives	576	58
Predicted negatives	156	477

Figure 4. Word cloud graph (These words that appear frequently in the text. The higher the frequency, the larger the font size. The top three largest words are: Huawei, mobile phone, and price.).

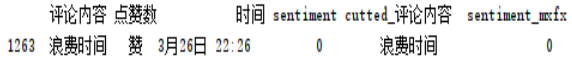

Figure 5. An example of a classification of the emotion analysis model (This picture shows a negative comment. Use the model to perform sentiment analysis on this comment, and the result is 0, which is negative.).

3. Conclusion

In the experiment of this paper, it is found that the accuracy of emotion classification was seriously affected by meaningless digital characters as feature words, which proves that the selection of feature words has a great impact on the effect of emotion classification. This paper used the TF-IDF model for Word Segmentation, which is helpful to select more refined characteristic words and improves the accuracy of emotion analysis. In order to improve the time independence of the model and improve the iteration efficiency, the naive Bayes-LSTM model was also established in this paper. Finally, the experimental results showed that the model can classify the emotion of Chinese text and obtain relatively accurate and stable analysis results.

Acknowledgments

This work was supported by Zhejiang province commonweal projects (ID: LGG20F020004) and General scientific research projects of Zhejiang Education Department (ID: (0218506) Y201840424).

References

[1] Yang Y and Pedersen J 1997 A comparative study on feature selection in text categorization *Proc. Int. Conf. on Machine Learning* pp 412-420.
[2] Pang B, Lee L and Vaithyanathan S 2002 Thumbs up? sentiment classification using machine learning techniques *Proc. Emnlp* pp 79-86.
[3] Wang S, Jiang L and Li C 2015 Adapting naive Bayes tree for text classification *Knowledge & Information Systems* **44** (1) pp 77-89.
[4] Arulampalam S, Maskel S, Gordon N and Clapp T 2002 A tutorial on pfs for on-line non-linear/non-gaussian Bayesian tracking *Scientific Programming* **50** (2) 174-188.
[5] Subramanian A, Alias B and Ramasamy R 2009 Effective and efficient feature selection for large-scale data using Bayes theorem *International Journal of Automation and Computing* **6** (1) 62-71.
[6] Chen G 2018 Text sentiment analysis based on polarity transfer and bidirectional long-short term memory *Information Technology* **2** 149-152.

[7] Zhang J, Zhuo W and Verma N 2018 In-Memory computation of a machine-learning classifier in a standard 6T SRAM Array. *IEEE Journal of Solid-State Circuits*, vol 99 pp 1-10.
[8] Zhang M, Peña J and Robles V 2009 Feature selection for multi-label naive Bayes classification *Information Sciences* **179** (19) 3218-3229.
[9] Yi J, Wen Z, Tao J, et al. 2017 CTC Regularized model adaptation for improving LSTM RNN based multi-accent mandarin speech recognition *Journal of Signal Processing Systems* **90** (2) 1-13.