ISSN: 2688-9323

Comparison of Answer Classification Methods in Question Answering Platform based on Deep Learning

Yujie Shi, Enpeng Hu*

School of Management, Shanghai University, Shanghai 200444, China

Abstract

As an important channel to obtain information and knowledge, community question answering platform plays an important role in people's life. It is very important to improve the efficiency of people's access to information in the community question answering platform, improve the user experience, and help users screen high-quality answers. This paper trains and compares four models of BERT CNN, BERT LSTM, CNN and LSTM on the legal domain question answering dataset. The experimental results show that the CNN model has the best classification result on the legal domain question answering dataset used in the training, and the Accuracy and Macro F1 values are both 0.94. In the cross domain dataset experiment, there is a gap with the original experiment, but it can provide a reference for the follow-up research. The answer quality classification model of community question answering platform based on deep learning can help people automatically select high-quality answers, which is of great significance to the platform and users.

Keywords

Community Question Answering; Deep Learning; Classifier.

1. Introduction

With the rapid development of information technology and the Internet, people increasingly rely on the Internet to obtain the information and knowledge they need to solve the problems encountered in life, study and work. The emergence of the community-based Q&A platform has greatly improved the user's experience of obtaining information. Compared with search engines, searching or asking questions on the community-based Q&A platform can easily obtain more relevant and accurate answers. Although users' spontaneous questioning and answering behaviors have contributed a lot of valuable content to the platform, this large-scale, disorderly information production has also led to information overload, and the amount of information people are exposed to has exploded.

Now, community-based Q&A platforms (such as Yahoo Answers, Zhihu, Baidu Know, Chunyu Doctor, Auto Home, etc.) have become important channels for users to share and acquire knowledge. With the increasing amount of information faced by users and the uneven quality of question answers, how to enable users to quickly find valuable answers has become an urgent problem to be solved. Helping users find the information they need more accurately and quickly can not only improve user experience, but also help increase user stickiness and the continuous development of the platform. Many community-based Q&A platforms use internal search engines and rely on information retrieval techniques such as query intent identification and query recommendation to provide search services. However, this method can only match user intent with questions and cannot find out which answers are high-quality. And display these answers on the front end of the page. In the field of natural language processing, the use of deep learning models can effectively capture text features. Therefore, exploring the use of

ISSN: 2688-9323

deep learning-based natural language processing technologies to identify high-quality answers in community-based question answering platforms is a very valuable research question.

This article is based on deep learning technology, using BERT (Bidirectional Encoder Representations from Transformers) as a pre-training model, Convolutional Neural Network (CNN), and Long-Short Term Memory (Long-Short Term Memory, LSTM) as a classification model to identify The high-quality answers in the community-based question-and-answer platform help the platform to better screen out the high-quality answers and present them to users, alleviate the information redundancy problem of the platform, and reduce the user's information acquisition cost.

2. Literature Review

Community Q&A, also known as interactive Q&A or social Q&A, refers to the way people ask others to obtain information on social networking sites, forums or using social search engines. The community-based Q&A platform provides a new platform for people to obtain information, promotes people's knowledge exchange, sharing and accumulation, and enables some knowledge that cannot be retrieved by search engines and stored in the human brain to be displayed, which makes community-based Q&A The platform has gradually become one of the important sources for users to obtain information. Therefore, how to quickly and accurately present the information that users need to users is one of the key points of community Q&A research.

At present, a lot of research is concentrated in the field of Question and Answering System (QA). The purpose is to hope that the question and answer system can analyze users' oral questions and organize relevant data through information retrieval, answer extraction, etc., and automatically Relevant answers are fed back to users. Among them, the Transformer model has achieved good results in the field of QA. Qingqing Cao et al. believe that the QA model based on Transformer is slower due to its internal self-attention mechanism, and will take up huge memory [1], so They introduced the DeFormer model to replace the full self-attention mechanism with the self-attention mechanism for the question and the self-attention mechanism for the answer at the bottom, which greatly reduced the amount of calculation during calculations.

For the community-based Q&A field, Yang Deng et al. studied the length and redundant questions of community-based Q&A platforms [2]. If the answer is too long, it may cause users to read difficulties and misunderstandings, and affect the efficiency of users to obtain information. They build A large-scale community-based question and answer corpus, and a new joint learning model is proposed to solve the problems of answer selection and answer summary generation. It has achieved good results on both questions. For the community-based Q&A platform, it is very important for the Q&A platform to match the highest quality answers with the corresponding questions. Xiao Yang et al. defined a binary classification question [3], which divides the question and the answer into Relevant and irrelevant, and proposed a confrontation training model to train the classifier to determine whether the question and the answer are relevant, so as to match the high-quality answer with the question, which is extremely important for improving the user experience. When users face the massive answers in the question-and-answer platform, users may not want to filter out all the answers, but only want to see the most important ones. Adi Omari et al. conducted a research on the method of sorting answers in the community-based question-and-answer platform [4]. They believe that not only consider the relevance of questions and answers, as well as diversity and novelty. They proposed a new sorting algorithm for answers, adding diversity on the basis of relevance, and coverage of important information in the question. Answer selection is an important question in community-based Q&A. Most existing methods deal with this question as a text matching

ISSN: 2688-9323

task. However, they ignore the influence of community users in voting for the best answer. Shanshan Lyu et al. believe that the quality of answers is highly related to semantic relevance and user expertise [5]. From the perspective of user expertise, they comprehensively consider the semantic relevance of the question-and-answer pair and the user expertise of the question-and-answer pair. Formalized the answer selection question, and designed a new matching function to explicitly simulate the impact of user expertise on community acceptance. With the accumulation of users and content, more and more attention has been paid to the efficiency and answer quality of the community-based Q&A platform. Xiang Cheng et al. proposed a concept of Question Routing, which aims to recommend new questions to appropriate answerers [6], These respondents have a higher probability of answering and professional ability. Ding Heng and Li Yingxuan proposed a query recommendation method based on deep learning to optimize the search recommendation within the platform [7]. Under the condition of the user's given query, natural language with a small distance from the user's question intention is found from the corpus Questions are collected and recommended to users as a result.

According to previous studies, among community-based Q&A-related research, research related to improving user experience and optimizing platform content has attracted more and more attention. Based on deep learning technology, this article compares the answer quality classification methods of community-based Q&A platforms. I hope to filter out high-quality answers in the platform through automatic methods instead of manual methods.

3. Experimental Data and Related Technologies

3.1. Experimental Data and Processing

The experimental data of this study comes from a public data set. The original data is the question and answer text data of Baidu Know Q&A platform. The data set belongs to the field of legal knowledge question and answer, and the question field and reply field are desensitized. The data set contains 36000 legal question and answer data. , Each piece of data contains four fields: title, question, reply and is_best. The is_best field is a label, a label value of 1 represents a high-quality answer, and a label 0 represents not a high-quality answer. This paper selects answer and label data as the data pair. The labels in the data set are manually labeled, and the data set is divided into training set and test set according to the ratio of 90% and 10%, with 32400 training sets and 3600 test sets. Examples of raw data are shown in Table 1.

Table 1. Data set example

Title	Question	Reply	Is_best
I turned right into the main road normally on the secondary road. There were two cars on the main road. The car behind occupied the opposite lane to overtake. Who is responsible for the collision?	NaN	In the event of a traffic accident, report to the police in time. The traffic police shall divide the responsibilities, issue a traffic accident certification, and negotiate and resolve it according to the division of responsibilities and damage...	0
Can I sue if I owe money and there is no IOU with a recording?	NaN	If the recording materials have not been edited, pieced together, tampered with or fabricated after verification, and are confirmed by other relevant evidence, their validity can be determined.	1

The data must be preprocessed before model training. When using BERT as a pre-training model, after the text segmentation is serialized, the [CLS] and [SEP] tags must be added at the beginning and the end of the text, so that it can be used as the input of the model. In addition,

ISSN: 2688-9323

in the preprocessing stage, we need to observe the descriptive statistical characteristics of the training set data, such as observing the distribution of text length in the training set, and setting appropriate hyperparameters based on the observed data.

3.2. Related Technology

The main implementation process of the experiment consists of three stages: data preprocessing, loading pre-trained models, and text classification. The pre-training model is BERT, and the text classification model is CNN and LSTM.

In the experiment, BERT is used as the pre-training model, and the lightweight RBT3 Chinese pre-training model is used. The pre-training corpus includes data such as Chinese Wikipedia, other encyclopedias, news, and Q&A. The goal of the BERT model is to use large-scale unlabeled corpus training to obtain a vector representation of the text that contains rich semantic information, that is, the semantic representation of the text, and then fine-tune the semantic representation of the text in a specific natural language processing task, and finally apply it For this natural language processing task, such as entity recognition, sentiment classification, similarity calculation, etc. BERT is derived from the encoder of the Transformer model. Compared with other language models, BERT introduces a large amount of external knowledge information and extracts the representation information of the word from the input context by using the attention mechanism therein, and finally obtains the more complete the characterization information of word, so as to obtain richer semantic information.

The classifiers selected in the experiment are the CNN model and the LSTM model in deep learning. Convolutional Neural Network (CNN) is a classic representative of artificial neural networks. Compared with ordinary neural network structures, its main advantage lies in the shared weight structure, which greatly reduces the parameters that need to be trained and improves the training speed. It was originally used in the image field. Later, CNN also achieved very good results in many natural language processing tasks, and the parallel computing power of convolutional neural networks in processing text is not available in cyclic neural networks. The CNN model constructed in this paper contains three convolutional layers, a pooling layer, a Dropout layer, a fully connected layer and an output layer.

The long short-term memory model (LSTM) is a variant of the Recurrent Neural Network (RNN). During the calculation of the internal structure of the RNN, the calculation at the current time step is to compare the activation value output by the previous time step with the current the input values of is spliced together, and the activation value of the current time step is output through the tanh activation function. The disadvantage is that the RNN unit can only focus on the state at the previous time. If the length of the input text is too long, it will cause the back propagation path to be too long. Long-distance gradient disappears, which eventually leads to the loss of long-distance text information. LSTM is proposed to solve the problem of long-distance text information memory. The internal structure of LSTM is more complex than RNN, and new calculation processes are added. LSTM introduces memory cells and gating mechanism, and uses memory cells to store information, and uses input gates, forget gates, and output gates to maintain and control information. In addition to the tanh function, the activation function also has a sigmoid function, which adds a summation operation to reduce the possibility of gradient disappearance and gradient explosion.

4. Experiment Implementation

4.1. Experiment Preparation

Community-based question answering platform answer quality classification research needs to build a classifier to achieve high-quality answers in the automatic recognition platform. First, you need to obtain and process the training data set. As mentioned above, the data set selected

ISSN: 2688-9323

in this article is the legal field knowledge question and answer data, a total of 36000 data, 32400 training sets, and 3600 test sets. Secondly, set the initial parameters of the model. Since the parameters need to be modified continuously in order to achieve better results during the model iteration process, the initial parameter values will not be given here, and the final parameter settings will only be given after the test.

4.2. Experiment Procedure

In the training model stage, for the question of answer quality classification of the community-based question and answer platform studied in this article, the deep learning model was selected in the experiment. The deep learning model has achieved very good results in many natural language processing tasks, among which CNN and LSTM are based the classifier effect is the most prominent, so this article chooses CNN and LSTM as experimental models to compare the classification effects of the two models. In addition, due to the outstanding performance of the BERT model in the field of natural language processing in recent years, and even the performance on some tasks surpassed the human level, this article introduces BERT as a pre-training model, which is combined with CNN and LSTM respectively into BERT CNN and BERT LSTM conducts a comparative test.

4.2.1. BERT CNN

CNN is a commonly used classifier in text classification tasks. In the experiment, BERT is used as a pre-training model to obtain semantic vectors containing rich semantics. Since BERT has trained a set of parameters through a large number of corpus, the trained BERT is equivalent to a powerful The feature extractor only needs to input training data to fine-tune the parameters of our CNN classifier model. The BERT CNN model constructed in this study uses three convolutional layers, each of which contains a convolutional layer and a maximum pooling layer. After the convolution operation is completed, the outputs of the three convolutional layers are spliced, In order to enhance the feature extraction ability of the model, the number of neurons in each convolutional layer is 256, the size of the convolution kernel is 3, 4, and 5 respectively, and all the convolutional layers use the Relu activation function. After the convolution operation, add a layer of Dropout layer to prevent over-fitting, followed by a fully connected layer to further extract text features, the number of neurons is 512, the Relu activation function is used, and the last is the output layer, the number of classifications is 2, use The softmax activation function, the optimizer is Adam optimizer, and the learning rate is set to 5e-7.

4.2.2. BERT LSTM

LSTM is a classic model in the field of natural language processing. It is good at capturing long-distance text information. It uses a pre-trained BERT model combined with LSTM to form a new text classifier BERT LSTM. The BERT LSTM classifier constructed in this paper uses a single-layer LSTM with a number of neurons of 64. After the input is extracted from the LSTM, the output feature vector is input to the fully connected layer. The number of neurons in the fully connected layer is 256, and the activation function is Relu, followed by the Dropout layer to prevent overfitting, and finally the output layer, the optimizer is Adam optimizer, and the learning rate is set to 8e-6. During the experiment, it was found that too many LSTM neurons will cause the training effect to deteriorate and the time will be longer. After many experiments to ensure that the model training effect is improved, the above parameter combination is finally determined.

BERT CNN and BERT LSTM model structure diagram is shown in Figure 1.

ISSN: 2688-9323

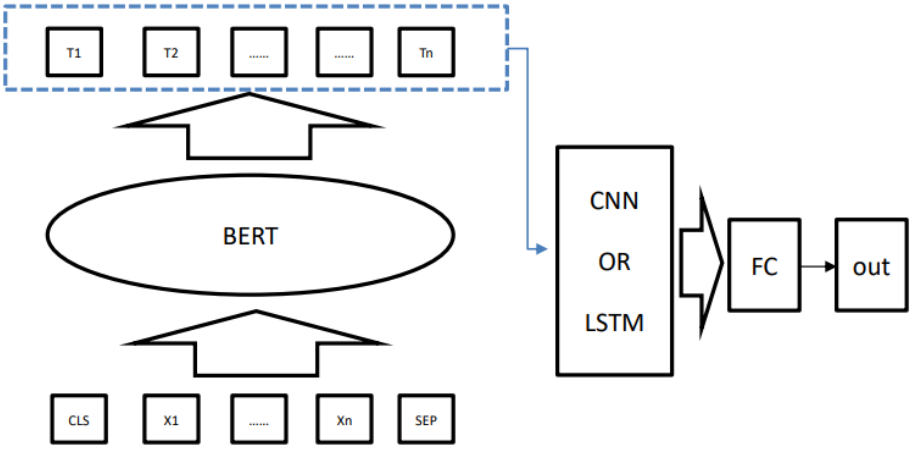

Figure 1. BERT CNN, BERT LSTM model structure

4.2.3. CNN

The CNN classifier model architecture is basically the same as the BERT CNN. Without the BERT make the input text is converted into a vector. The CNN model adds the Embedding layer to obtain the word vector, followed by the same three-layer convolutional layer, Dropout layer, and full connection as BERT CNN For layer and output layer, the optimizer is Adam optimizer, and the learning rate is set to 5e-7. The CNN model achieved the best results among the four models.

4.2.4. LSTM

Similar to a pure CNN, a separate LSTM model requires an Embedding layer to obtain the word vector, and the word vector is input to the model to extract features. The number of neurons in the LSTM layer is 256, which enhances feature extraction capabilities, fully connected layer, Dropout layer and output The layer is consistent with BERT LSTM, the optimizer is Adam optimizer, and the learning rate is set to 8e-6.

The above parameters are obtained by constant adjustment and comparison in the experiment. Although the model converges and the effect is good, due to time and computer performance limitations, there are still many parameter combinations that have not been tested. Among them, due to computer memory limitations, the batch sample size of BERT CNN and BERT LSTM Hyperparameters cannot exceed 16, and the maximum length of the text serialized representation should be less than 600. The batch sample size and sequence length will also affect the experimental results to a certain extent, so the above four models may not be in the current the best results are achieved on the data set.

The structure diagram of the CNN and LSTM model is shown in Figure 2.

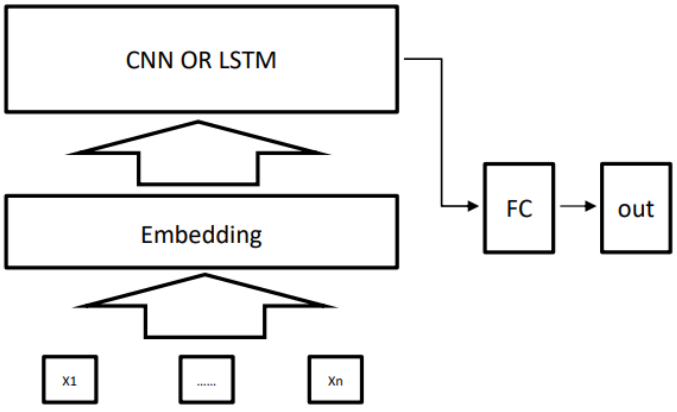

Figure 2. CNN, LSTM model structure

ISSN: 2688-9323

4.3. Evaluation Index

Accuracy: Accuracy is the ratio of correctly classified samples to the total number of samples. In this experiment, it is the ratio of the sample predicted as label 1 to the total number of samples.

Precision: The precision rate, that is, the proportion of the positive data predicted to be correct to the positive data predicted. In this experiment, it is the proportion of the number of samples that are correctly predicted to be label 1 to the number of samples that are predicted to be label 1.

Recall: That is, the proportion of positive data predicted to be correct to actual positive data. In this experiment, it is the proportion of data correctly predicted to be label 1 to all label 1 data. It has a trade-off relationship with Precision.

F1-score: It is the reciprocal of the harmonic average of "Precision" and "Recall". Taking the two into consideration, it is a more compromised way.

Macro F1-score: The F1-score of each category is averaged, and it is not affected by data imbalance.

5. Result Analysis

5.1. Experimental Model Evaluation

Table 2 shows the final results of the four models to classify the answer quality of the community-based Q&A platform. In general, according to the experimental results, in the legal field question and answer data set, only the traditional deep learning model has achieved better classification results than the combined model using BERT as the pre-training model, and the CNN model is better than LSTM model. The CNN model achieved the best results, achieving the best performance on the Precision, F1-score, Accuracy and Macro F1 indicators, followed by the BERT CNN model. The Accuracy and Macro F1 values both reached 0.92, and the optimal result was 0.94. It is the closest, and achieves the best performance in the Recall value. After that is the LSTM model, the Accuracy and Macro F1 values are both 0.88, and the worst performing is the BERT LSTM model, all indicators are the lowest, and the Accuracy and Macro F1 values only reach 0.86. The experimental results show that for the legal field question and answer data set used in this study, CNN can extract features in the text more effectively than LSTM, and the effect of traditional CNN is slightly better than BERT CNN, but in terms of model convergence speed, BERT CNN is slightly better than traditional CNN. Based on various indicators, only using the CNN model to classify the answer quality of the community Q&A platform has the best effect, and the model is the best.

Table 2. Experimental results

Model	Precision	Recall	F1-score	Accuracy	Macro F1
BERT CNN	0.89	0.96	0.92	0.92	0.92
BERT LSTM	0.81	0.94	0.87	0.86	0.86
CNN	0.93	0.95	0.94	0.94	0.94
LSTM	0.83	0.95	0.89	0.88	0.88

5.2. Cross-domain Application Evaluation

Whether the model can be applied across domains is an important issue in many studies. Therefore, this article adds a set of comparative experiments to compare the performance of the CNN model and the BERT CNN model that performed well above on the insurance field question and answer data set. The data source is a public data set, and the original data is Baidu Know Q&A platform's insurance field question and answer text data. A total of 4181 pieces of

data are selected for evaluation. The experimental results on the question-and-answer data set in the insurance field are shown in Table 3. It can be seen that the two models do not perform well in Cross-domain data sets. The highest Accuracy value is only 0.71, which is quite different from the highest Accuracy value of 0.94 in the original experiment, indicating that the model does not yet have the ability to directly apply across domains. , A more comprehensive data set is needed to train the model and enhance the generalization ability of the model. In the Cross-domain comparison experiment, the BERT CNN model is better than the traditional CNN model in all indicators. In the original experiment, the BERT CNN model is slightly inferior to the CNN model, which highlights the role of the BERT pre-training model. The semantics of the word vectors obtained by the language model trained on the corpus are richer and have the advantage of Cross-domain application.

Table 3. Results of CNN and BERT CNN models on the insurance question and answer data set

Model	Precision	Recall	F1-score	Accuracy	Macro F1
BERT CNN	0.63	0.91	0.74	0.71	0.71
CNN	0.61	0.82	0.70	0.68	0.68

6. Conclusion

Based on deep learning technology, this paper studies and compares the traditional deep learning text classifier model and BERT as a pre-training model combined with the traditional deep learning model, and compares the CNN classifier model and the LSTM classifier model in the community. The effect of Q&A platform answer quality classification task. In the end, it was found that the CNN classifier model performed best in the question and answer data set in the legal field. The four experiments are all through the conversion of text data into the feature space to extract the high-dimensional semantic relationship between words, thereby obtaining a vectorized representation of the text. The difference lies in the way the BERT pre-training model and the traditional Embedding layer extract text features different.

This article also adds comparative experiments on Cross-domain data sets to verify the Cross-domain application capabilities of the models constructed in the experiments. The experiment compared the performance of the BERT CNN and CNN models in the insurance field question-and-answer data set with better results in the original experiment. The results show that the two models do not perform well in the data set of the new field and do not have the ability to directly use across fields. The performance of the BERT CNN model in the new data set is better than that of the traditional CNN model. The BERT pre-training model can make the entire model have stronger generalization capabilities.

References

[1] Q.Q. Cao, H. Trivedi, A: Balasubramanian, et al. DeFormer: decomposing pre-trained transformers for faster question answering, Proc. of the 58th Annual Meeting of the Association for Computational Linguistics (Online, May 2, 2020). p.4487.

[2] Y. Deng, W. Lam, Y.X. Xie, et al. Joint learning of answer selection and answer summary generation in community question answering, Proc. of the Thirty-Fourth AAAI Conference on Artificial Intelligence (New York, USA, April 3, 2020). Vol. 34, p.7651.

[3] X. Yang, M. Khabsa, M.S. Wang, et al. Adversarial training for community question answer selection based on multi-scale matching, Proc. of the Thirty-Third AAAI Conference on Artificial Intelligence (Hawaii, USA, July 17, 2019). Vol. 33, p.395.

[4] A. Omari, D. Carmel, O. Rokhlenko, et al. Novelty based ranking of human answers for community questions, Proc. of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval (Pisa, Italy, July 7, 2016). Vol. 16, p.215.

ISSN: 2688-9323

[5] S.S. Lyu, W. Ouyang, Y.Q. Wang, et al. What we vote for? answer selection from user expertise view in community question answering, Proc. of the 19th World Wide Web Conference (San Francisco CA, USA, May 13, 2019). Vol. 19, p.1198.

[6] X. Cheng, S.G. Zhu, S. Su, et al. A multi-objective optimization approach for question routing in community question answering services, IEEE Transactions on Knowledge and Data Engineering, Vol. 29 (2017) p.1779-1792.

[7] D. Heng, Y.X. Li: Improving online Q&A service with deep learning, Data Analysis and Knowledge Discovery, Vol. 4 (2020) No.10, p.37-46.