ISSN 2959-6157

LSTM-Based Forum Topic Classification Model

Leo Liang

Abstract:

With the rapid development of information technology, forums have become an important platform for information exchange. However, the manual classification of forum topics consumes a significant amount of human resources and is prone to classification errors. To address this issue, this study proposes a forum topic classification model based on Long Short-Term Memory (LSTM) networks. By leveraging LSTM's capability in text processing, the accuracy and efficiency of topic classification are significantly improved. This study used approximately 68,000 entries from 29 topic categories, scraped from Zhihu, for the experiments. Preprocessing steps such as text cleaning, tokenization, and word vectorization were performed, and a classification model with LSTM and Dropout layers was designed. The experimental results indicate that the model performs well in most topic classifications, although overfitting remains an issue for certain categories. The paper concludes by summarizing the advantages and limitations of the model and discusses the potential for improving classification accuracy through increasing data volume and optimizing the model in the future.

Keywords: Deep Learning, Natural Language Processing, Long Short-Term Memory Network, Forum, Text Multi-classification

1. Introduction

With the rapid advancement of information technology, online activities have gradually become an integral part of our social activities. The internet connects people worldwide, gathering numerous experts and scholars, significantly promoting knowledge sharing and exchange. Consulting various questions on the internet often leads to responses from experts in the field, not only solving our problems but also providing additional assistance.

However, sometimes the individuals asking questions might not be familiar with the field and fail to post their questions in the correct category. Traditional forums generally rely on moderators to manually manage the sections, but this method is labor-intensive and still prone to errors.

To improve efficiency, machine learning methods can be used to alleviate the burden of review. By using LSTM-based form question classification, the forum moderators can be alerted if a topic is misclassified, thus enhancing the efficiency of issue response, making it easier for users to find relevant information, and promoting knowledge sharing and community collaboration. Additionally, automatic question classification can optimize resource allocation and improve the timeliness of issue handling, providing a better user experience. Such technology helps build a more intelligent and efficient social interaction platform.

1.1 Domestic and International Overview

In recent years, significant advancements have been made in the field of text classification through deep learning. Han Mei demonstrated the advantages of deep learning models in sentiment analysis, achieving precise classification of Chinese bullet screen (bullet screen comments) by combining Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. Additionally, Liang Dengyu focused directly on the application of LSTM in Chinese text classification, confirming LSTM's advantages in handling long sequence text data, especially in retaining long-term dependency information. Siyu Xin in explored the application of deep learning in multi-class classification of Chinese news texts. By using the THUCNews dataset and combining word embedding techniques, TextCNN, and RNN models, efficient text classification was achieved. Zhang et al. further validated the broad applicability of deep learning models in multi-language text classification tasks, demonstrating their robustness and efficiency in different language environments. Li's research explored how to improve the accuracy and efficiency of text classification by optimizing the structure and parameters of deep learning models. The results showed that the optimized models outperformed traditional methods across multiple datasets. Feng, Guozhong, et al. discussed an innovative text classification algorithm, assigning appropriate values to each category to improve text classification by converting textual document content into vector space. Extensive experiments confirmed the superior performance of this algorithm in various application scenarios. Machová proposed a novel deep learning-based text classification method, utilizing advanced algorithms and model structures to achieve efficient and accurate text classification.

2. Theoretical Basis and Methodology

2.1 LSTM

LSTM (Long Short-Term Memory) is a special type of recurrent neural network (RNN) that is particularly suitable for processing and predicting sequence data. Therefore, it is frequently used in natural language processing (NLP) tasks, including topic classification. The reasons for selecting LSTM are primarily threefold:

Traditional RNNs encounter gradient vanishing or exploding issues when dealing with long sequences. LSTM, however, has a special gating mechanism (forget gate, input gate, and output gate) that can effectively retain and utilize long-term dependency information. This enables LSTM to better understand the relationships and dependencies between different parts of the text, thereby improving classification accuracy.

LSTM can handle input sequences of varying lengths, which is crucial for topic classification since document or paragraph lengths may vary significantly. Unlike traditional methods that require strict preprocessing like padding or truncation, LSTM preserves more of the original information.

The gating mechanism in LSTM allows it to selectively ignore irrelevant or distracting information while processing input sequences. This means it can retain more useful and critical information, enhancing text classification accuracy.

2.2 Data Preprocessing

Before classifying Chinese text, preprocessing is typically required to support subsequent tasks better. During preprocessing, we first clean the text of meaningless characters like punctuation marks to reduce noise and improve processing efficiency.

2.3 Dropout Layer

Dropout is a structure that can reduce overfitting in neural networks. The larger the neural network, the more likely it is too overfit. Therefore, by randomly deleting some neurons, we can prevent overfitting, ensuring the fitted results are not overly accurate.

2.4 Dense Layer

Since this is a text classification model, the activation function selected is softmax, and the output dimension is 30. This means the model attempts to classify input data into 30 different categories. Each neuron's output is converted by the softmax function so that the sum of all output values is 1. This allows each neuron's output to be interpreted as the predicted probability for the corresponding category, achieving model classification.

3. LSTM-Based Forum Topic Classification Model

Firstly, we need to determine the number of categories. Secondly, we proceed with text cleaning and tokenization on the acquired dataset. Next, we vectorize the text and use a classification model to train a usable model. Thirdly, we use this model to attempt text classification.

3.1 Data Acquisition and Preprocessing

The dataset was obtained through web scraping, mainly from topics on Zhihu and their corresponding categories. There are 29 topic categories (including digital, technology, internet, business finance, workplace, education, law, military, automotive, humanities and social sciences, etc.), with approximately 68,000 entries to simulate a forum usage scenario. We use Long Short-Term Memory (LSTM) networks to train and achieve forum question classification, ensuring that questions are correctly classified into the corresponding tags based on their content.

3.2 Model Design and Construction

This study adopts an LSTM-based model architecture to achieve forum topic classification. LSTM has the ability to handle long sequence data, making it highly suitable for natural language processing tasks. The overall architecture of the model is as follows:

Input Layer: The model's input is preprocessed and tokenized text data. The representation of each text is a sequence of word indices.

Embedding Layer: Word2Vec is used to convert the word index sequences into word vector matrices. The purpose of the embedding layer is to map high-dimensional sparse word indices into a low-dimensional dense vector space, capturing the semantic relationships between words.

LSTM Layer: The LSTM layer is the core of the model, used to capture contextual information in the text sequence. Choosing LSTM over a simple RNN is based on its effective solution to the gradient vanishing problem and its suitability for handling long sequence text. In constructing the LSTM-based forum topic classification model, we use 256 LSTM units, aiming to balance complexity and performance.

Dropout Layer: The Dropout layer is used to prevent overfitting. By randomly discarding a certain proportion of neurons, the model's generalization ability is enhanced. The dropout rate is set to 0.1, meaning 10% of neurons are randomly dropped to improve the model's generalization ability.

Dense Layer: The Dense layer maps the output of the LSTM layer to the classification space. We selected 30 neurons and used the softmax activation function to generate the probability distribution for each category.

3.3 Experiment and Analysis Results

We input the nearly 70,000 data entries from 29 categories mentioned above into the model, setting the training period to 10 epochs to train and experiment with the model.

In the figure below, we can see that while the training error continuously decreases, the test set error initially rises and then falls, eventually stabilizing. This indicates overfitting, as the limited number of samples per category reduces the model's generalization ability, leading to poorer performance on new or unseen data.

Figure 1 Accuracy rate as a function of the training cycle

The experimental results show that the LSTM-based model performs well in most topic classification tasks, with an overall accuracy of 79.29%. However, there are significant differences in classification performance across different topics. Since we cannot evaluate a multi-class model solely based on accuracy, we also consider recall rate and F1-score. The table below shows the model's accuracy, recall rate, and F1-score for each topic category:

Table 1 Model evaluation results

Class	Accuracy	Recall Rate	F1-Score	Sample Count
Gender	0.74	0.91	0.82	32
Internet	0.59	0.69	0.64	32
Humanities and Social Sciences	0.82	0.71	0.76	78
Sports and E-sports	0.95	0.66	0.78	29
Health	0.86	0.81	0.83	37
Military	0.88	0.92	0.9	38
ACG	0.7	0.82	0.76	40
Business finance	0.74	0.91	0.82	35
Pets	0.96	0.9	0.93	29
Home	0.9	0.9	0.9	31
technology	0.71	0.75	0.73	32
Film and Entertainment	0.71	0.81	0.76	31
Psychology	0.76	0.71	0.74	35
Emotions	0.84	0.76	0.79	41
Education	0.75	0.84	0.79	32
Digital	0.85	0.65	0.74	43
Travel	0.86	0.83	0.85	30
Fashion	0.89	0.83	0.86	30
Mother and Baby	0.98	0.78	0.87	54
Automotive	0.79	0.86	0.83	36
Law	0.86	0.72	0.78	43

Technology	0.52	0.69	0.59	39
Food	0.82	0.93	0.88	30
Workplace	0.9	0.76	0.82	37
Natural Sciences	0.59	0.84	0.69	31
Art	0.89	0.75	0.81	32
Design	0.78	0.88	0.82	32
Fitness	0.81	0.73	0.77	30
Music	0.94	0.91	0.92	33

Overall, the model performs best in the “Military” and “Pets” categories, with F1 scores of 93% and 90%, respectively. This indicates that the LSTM model has advantages in processing structured information and text with dense technical terms. In contrast, the F1 scores for the “Internet” and “Technology” categories are below 65%, showing relatively poor performance. This may be due to significant overlap between these categories, making it difficult for the model to distinguish their respective features.

We conducted an in-depth analysis of the model’s classification differences and identified the main reasons as follows:

Data Distribution Imbalance: Some topic categories have fewer data samples, resulting in insufficient learning by the model in these categories. For example, the “Psychology” category has fewer data samples, and the content is diverse, making it difficult for the model to extract uniform features.

Topic Overlap: There is significant overlap between some topic categories. For instance, posts in the “Digital,” “Technology,” and “Internet” categories may have overlapping content, leading to confusion in the model when distinguishing these categories.

Text Length and Complexity: Different topics have varying text lengths and complexity, affecting the LSTM model’s adaptability. Longer technical discussion texts may be more suitable for the LSTM model’s long sequence processing capabilities, while short literary reviews may lead to insufficient feature extraction by the model.

4. Conclusion and Future Work

This study proposes a forum topic classification model based on Long Short-Term Memory (LSTM) networks. It can automatically identify and classify forum topics, enhancing moderators’ efficiency and optimizing user experience. The experimental results show that the model performs well in most topic classifications, with high accuracy and efficiency. Especially in handling long sequence text data, the LSTM model demonstrates its unique advantages. However, the experiments also reveal that the model exhibits overfitting in certain categories, resulting in lower classification performance for these categories. This is mainly due to the insufficient number of training data samples and significant overlap between some topics.

References

[1] Han Mei. Research on Sentiment Analysis of Chinese Bullet Screen Text Based on Deep Learning [D]. Nanchang University, 2024.

[2] Liang Dengyu. Application Research of Chinese Text Multi-Classification Based on LSTM [J]. Journal of Shanghai University of Electric Power, 2020.

[3] Xin S .Multi-classification application of Chinese news text based on deep learning[J].Journal of Physics Conference Series, 2020, 1549:022011.DOI:10.1088/1742-6596/1549/2/022011.

[4] Hua, Zhang, Jiawei Qin, Yan Wang, Yuan Ma, L. Yao and Jun Lei. “Research on Android Multi-classification Based on Text.” Journal of Physics: Conference Series 1828 (2021): n. pag.

[5] Lei, Tao, Regina Barzilay and T. Jaakkola. “Molding CNNs for text: non-linear, non-consecutive convolutions.” Conference on Empirical Methods in Natural Language Processing (2015).

[6] Feng, Guozhong, Shaoting Li, Tieli Sun and Bangzuo Zhang. “A probabilistic model derived term weighting scheme for text classification.” Pattern Recognit. Lett. 110 (2018): 23-29.

[7] Machová, Kristína, Martin Mikula, Xiaoying Gao and Marián Mach. “Lexicon-based Sentiment Analysis Using the Particle Swarm Optimization.” Electronics (2020): n. pag.

[8] Zhu Lili. Research on Chinese Text Classification Based on Attention Mechanism and LSTM-CNN [D]. Chongqing University of Technology, 2023.

[9] Kong Weize, Liu Yiqun, Zhang Min, et al. Research on the Evaluation Method of Answer Quality in Question and Answer Communities [J]. Journal of Chinese Information Processing, 2011.

[10] Chang Lei, Wang Yilun, Chen Yanping. Application Research of Text Multi-Classification Based on Bert Model [J].

Computer Knowledge and Technology, 2023.

[11] Zhang Xin, Zhai Zhengli, Yao Luyao. Chinese News Text Classification Based on CNN and LSTM Hybrid Model [J]. Computer and Digital Engineering, 2023.

[12] Xu Peng. Research on News Text Classification Methods Based on Deep Learning [D]. Nanjing University of Information Science and Technology, 2024. DOI:10.27248/d.cnki.gnjqc.2023.000230.

[13] Shi, Xingjian, Zhourong Chen, Hao Wang, D. Y. Yeung, Wai-Kin Wong and Wang-chun Woo. “Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting.” Neural Information Processing Systems (2015).

[14] Behera, Ranjan Kumar, Monalisa Jena, Santanu Kumar Rath and Sanjay Misra. “Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data.” Inf. Process. Manag. 58 (2021): 102435.

[15] Xue Jincheng, Jiang Di, Wu Jiande. Research on Automatic Patent Text Classification Based on Word2Vec [J]. Information Technology, 2020, 44(02): 73-77. DOI:10.13274/j.cnki.hdzj.2020.02.015.