Proceedings of the 3rd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/2025.23319

Sentiment Analysis Method for Douban Movie Reviews Based on Prompt Learning

Yining Zhang

Henan University, Kaifeng, China
1571669656@qq.com

Abstract: Movie reviews reflect the viewers' evaluations of films. Conducting sentiment analysis on these reviews not only helps viewers select films they enjoy but also provides guidance for filmmakers in their creative processes. The sentiment analysis task for movie reviews has evolved from sentiment lexicon construction and machine learning to deep learning. These methods all rely on the fine-tuning paradigm, which has performance limitations when the downstream task objectives differ from the pretraining objectives. To address this issue, this paper adopts a prompt learning-based model. By designing task-description-based prompt templates, the downstream task is reformulated as a masked language prediction task, making full use of the semantic understanding of pre-trained language models. Experimental results show that compared to existing fine-tuning methods, the prompt learning-based approach improves accuracy by 6.8%-7.1% on the Douban movie review dataset, demonstrating the excellent performance of the prompt learning paradigm.

Keywords: Prompt learning, Movie reviews, Sentiment analysis

1. Introduction

In the age of advanced internet technology and social media, movies have become one of the most important forms of entertainment for people [1]. After watching a movie, people often post reviews on social media to express their opinions. As one of the most influential movie websites in China, Douban[1] provides an extensive range of features, including movie introductions, reviews, ratings, and more. The goal of movie review sentiment analysis [2] is to analyze the subjective sentiment of the reviewer from the movie reviews, determining the polarity of the sentiment expressed. This task holds wide application value in the movie industry [3]. For viewers, sentiment analysis helps them quickly understand the content and quality of movies, allowing them to choose films that suit their preferences. For filmmakers, it provides insight into audience feedback, helping them create more popular films and advance the development of the movie industry.

Early movie review sentiment analysis methods evolved from manually constructed sentiment lexicons [3] to the development of machine learning [4] technologies. Currently, deep learning [5] has become the mainstream research method. Researchers use neural network models such as Convolutional Neural Networks (CNN) [6], Recurrent Neural Networks (RNN) [7], and Long Short-Term Memory (LSTM) [8] networks to process text. Through the combination and improvement of network architectures, progress has been made in the performance of hybrid neural network models.

[1] https://movie.douban.com

DOI: 10.54254/2755-2721/2025.23319

However, deep learning models learn significant features of text from large-scale labeled data, and their understanding of semantics relies on statistical patterns in the training data, which limits their ability to fully capture higher-level concepts such as syntax and semantics.

In recent years, the rise of pre-trained language models like BERT [9] and GPT [10] has provided new solutions for movie review sentiment analysis tasks. The fine-tuning paradigm, which adjusts pre-trained language models to adapt to downstream tasks, can achieve good classification results but requires modifications to the model's network structure based on the specific task's objective function, reducing computational efficiency and limiting model performance [9]. Prompt learning is an emerging paradigm in the field of natural language processing [11]. Guided by task-description-based prompt templates, downstream tasks are reformulated as masked language prediction tasks familiar to pre-trained language models during training. This method does not require fine-tuning of the pre-trained language model, allowing it to better utilize its learned semantic knowledge and further improve model performance.

This paper uses a prompt learning-based pre-trained language model to complete the sentiment analysis task of movie reviews, aiming to verify the effectiveness of this method on the Douban movie review dataset and compare it with previous mainstream methods, exploring the advantages of prompt learning. First, task-relevant prompt templates are designed and integrated into the input sentences. For example, for the original data "A rare masterpiece recently, the plot is very touching," the prompt template converts it to "The sentiment of the following review is [MASK] oriented. A rare masterpiece recently, the plot is very touching." The input data is processed using the masked language prediction model to fill in the blank, and the sentiment analysis result is output through answer mapping. This process reformulates the task into a specific form, using the model's learned method to better complete the task, effectively reducing learning costs and error rates.

Experimental results on the Douban movie review dataset show that the model in this paper outperforms other baseline models, with an accuracy improvement of 6.8%, a recall rate improvement of 6.6%, and an F1-score improvement of 6.7% compared to the next-best model. This demonstrates that the prompt learning method can more accurately determine the sentiment polarity of text in movie review sentiment analysis tasks.

2. Related work

2.1. Sentiment analysis of movie reviews

In past research, sentiment dictionaries [3] and machine learning [4] methods have been applied to the sentiment analysis of movie reviews. With the development of deep learning technologies, researchers began using neural network models to process movie review data. Nedjah [12] and others analyzed the impact of hyperparameters on model performance and proposed a sentiment analysis classifier based on convolutional neural networks, achieving significant progress on movie review datasets. Liu [13] and others proposed CNN–BiLSTM, which integrates convolutional neural networks and bidirectional long short-term memory networks for feature extraction, effectively improving text classification accuracy. Deep neural networks can automatically learn text features from large-scale training data, capturing semantic information effectively without complex feature engineering. However, the model's understanding of contextual semantic relationships depends on the quality of the labeled data and lacks mastery of general world knowledge, which limits its accuracy in sentiment classification tasks. In recent years, the rapid development of pre-trained language models has brought new breakthroughs to sentiment analysis tasks for movie reviews. Researchers have utilized large amounts of text from daily life to train models, allowing them to learn word probability distributions and general language features, resulting in good performance on downstream tasks.

DOI: 10.54254/2755-2721/2025.23319

2.2. Prompt learning paradigm

Prompt learning is an emerging paradigm in the field of natural language processing. Unlike the fine-tuning paradigm, prompt learning does not require modifying the pre-trained language model itself but instead changes the format of the downstream task with the help of prompt templates. The prompt model is designed as a natural language prompt containing unfilled slots, describing the downstream task goal while incorporating the original data to be classified. By masking the emotional label words in the template, the masked language prediction model is guided to fill the slot with an emotionally charged word. Finally, the predicted answer is mapped to the emotional label, outputting the classification result. Through this design, the original text classification task is transformed into a cloze task of predicting the masked word, which aligns well with the task the model solved during pre-training. This helps preserve and utilize the semantic knowledge learned by the pre-trained language model, leading to a more accurate understanding of deep semantic information.

Prompt learning has garnered attention from researchers and has made progress in several domains. For example, in the multimodal domain [14], prompt templates have successfully been applied to image classification and text retrieval tasks; in machine translation [15], prompt templates have enabled efficient cross-language text generation; in question-answering systems [16], prompt templates help reduce model dependence on labeled data. This paper applies prompt learning to the sentiment analysis of movie reviews, verifying its effectiveness through the design and transformation of prompt templates.

In summary, research on sentiment analysis of movie reviews has evolved from early sentiment dictionary construction [3] to machine learning [4] technologies and made significant progress with the development of deep learning [5] and pre-trained language models [17]. However, current research still has some limitations. To accommodate the objective functions of downstream tasks, the fine-tuning paradigm requires modifying the model to add extra output layers, which not only affects the prior knowledge it has learned but also limits the model's ability to fully understand the semantic relationship between labels and text. Based on these issues, this paper applies prompt learning to the sentiment analysis of movie reviews, which can effectively alleviate the limitations of the fine-tuning paradigm and achieve significant performance improvements.

3. Experimental framework

The overall architecture of the sentiment analysis method for movie reviews designed in this paper is shown in Figure 1. It consists of four parts: the input layer, embedding layer, model layer, and output layer. First, the prompt template with the mask identifier [MASK] is concatenated with the movie review as input for the model. Then, word embedding technology is used to convert the input data into word vectors, and a masked language prediction model is applied to predict the word at the [MASK] position. The resulting hidden vector is passed through the MLM head to output the probability distribution over the entire vocabulary. Finally, answer mapping is applied to convert the prediction results into three sentiment labels: positive, negative, and neutral, outputting the sentiment analysis result of the movie review.

DOI: 10.54254/2755-2721/2025.23319

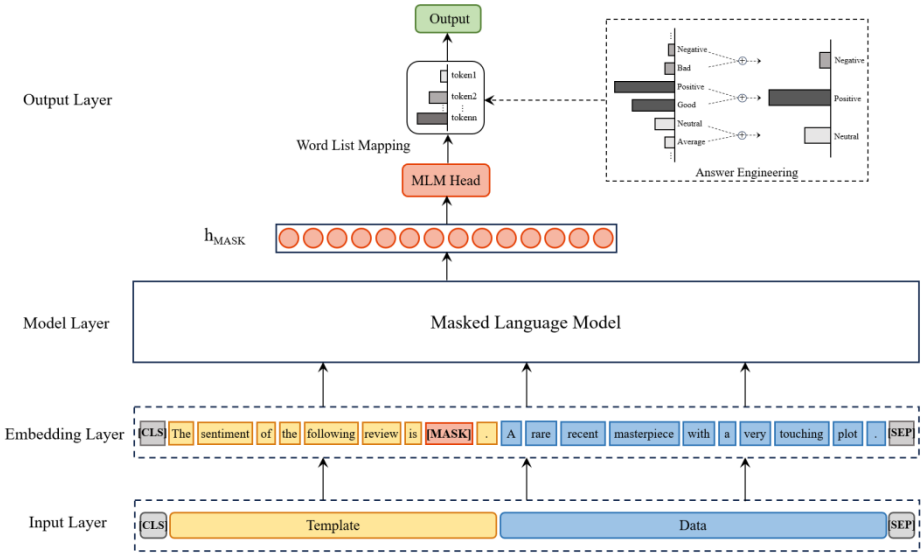

Figure 1: Sentiment analysis architecture based on the prompt learning paradigm

3.1. Prompt template engineering

The core of prompt learning lies in the transformation of the downstream task format, and the key to this process is the construction of task description-like prompt templates, that is, the creation of the prompt function $f_{prompt}(\cdot)$. This paper adopts a method of manually constructing prompt templates, which can fully leverage human language knowledge and understanding of the task, enabling stable performance on downstream tasks [18]. The first step is to determine the shape of the prompt. Since pre-trained language models have learned a substantial amount of language representations through masked language prediction tasks, cloze-style prompts are a good choice. They embed a slot to be predicted in the prompt, aligning well with the pre-training task. Next, the downstream task is restructured. For the sentiment analysis of movie reviews, the training objective of past models was to predict the probability distribution of sentiment labels based on raw input data. In prompt learning, the classification task is reformulated as a cloze task, enabling the model to address the problem in a format consistent with its pretraining. Accordingly, when constructing the masked word, it is necessary to extract the salient features of the target classification labels—"positive," "negative," and "neutral"—and then mask the initial character, restructuring the phrase into a word combination with an answer slot: "[MASK] sentiment." The integrated prompt template is designed as: "The sentiment of the following review is [MASK] sentiment." This template not only facilitates the model's understanding of the task but also reuses the model's inference and filling ability. After the prompt template is concatenated with the data, x' is used as the model's input for further processing.

3.2. Model training

Once the raw data is shaped into a format that the model can understand, word embedding technology is used to map it into high-dimensional word vectors, which include the initial word vectors, segment vectors, and position vectors. The resulting sentence vector representation is as follows:

$$E = E_{token} + E_{segment} + E_{position}$$

Where $E \in R_{n*k}$, n is the length of the input sequence, and k is the dimension of the word vector.

The vectorized text is then passed through the model for processing. The masked language prediction model fills in the cloze-style prompts defined specifically for this task and predicts the

DOI: 10.54254/2755-2721/2025.23319

word at the [MASK] position. The goal of the model training process is to minimize the prediction error at the masked position, thus obtaining the predefined sentiment label word as accurately as possible. The loss function used is the cross-entropy loss, which measures the discrepancy between the predicted probability distribution and the true labels. The loss function is expressed as:

$$L = -\sum_{i=1}^{N} log\, P(y_i|x'_i)$$

Where N is the number of training samples, and $P(y_i|x'_i)$ is the probability of category y_i. To optimize the training effectiveness, the Adam optimizer is used, which dynamically adjusts the learning rate based on gradient information, achieving faster convergence and higher stability.

3.3. Answer engineering

After filling the prompt template as required, the masked language prediction model outputs the hidden vector h_{MASK} at the [MASK] position, which is then used to derive the probability distribution over the entire vocabulary. If the model directly selects the token with the highest score as in conventional approaches, the result may not match the true label. This is because the model searches for potential answers across the full vocabulary, whereas the standard answer space is limited. Consequently, there is often a deviation between the predicted masked token and the target classification label. For instance, if the expected standard answer is "positive," but the model substitutes a semantically similar synonym, the score of the standard answer may be reduced, leading to a misclassification. To calibrate the original probability distribution and enhance the model's predictive performance, answer engineering is required to implement specific relational mappings.

For the task addressed in this paper, the model is expected to fill in one of the three words: "positive," "negative," or "neutral." To ensure that semantically similar responses yield the correct output, it is essential to expand the label vocabulary and enhance semantic coverage, thereby increasing the diversity and flexibility of the answer space. We define a mapping function $f_{answer}(\cdot): z \to y$, which maps an answer z to an emotional label y. For example, if the model predicts "bad," it can be further mapped to the emotional label "negative" based on semantic relationships and mapping rules. After adjusting the probability distribution accordingly, the label with the highest score is selected as the filled token, thereby determining the sentiment orientation of the movie review.

This paper constructs a sentiment analysis model for movie reviews based on the prompt-learning paradigm. First, prompt engineering is used to construct prompt templates, converting the classification task into a masked prediction task. Then, the word sequence is transformed into word embeddings, and a masked language model is employed to predict the masked word. Finally, label mapping is achieved through answer engineering, enabling the model to output the prediction result and complete the sentiment analysis task for movie reviews.

4. Experimental results and analysis

This chapter introduces five parts in sequence: experimental data and preprocessing, parameter settings, evaluation metrics, result analysis, and ablation study. The effectiveness of the proposed method is verified through evaluation using several mainstream metrics and comparison with baseline models on the Douban movie review dataset.

DOI: 10.54254/2755-2721/2025.23319

4.1. Experimental data and preprocessing

The Douban movie review dataset used in this study was obtained from the public dataset platform Kaggle[2]. Due to hardware limitations, only one year's worth of user review data was selected for the experiment. The original data includes fields such as movie title, username, date, user review, and rating. Following standard preprocessing procedures, data cleaning was first performed to remove duplicate and empty user comments and eliminate special characters in the text. Ratings were then mapped to the three sentiment labels "positive," "negative," and "neutral" based on the following rules: 1–2 stars as negative sentiment, 3 stars as neutral sentiment, and 4–5 stars as positive sentiment. After preprocessing, the dataset used in this experiment contains more than 600,000 movie reviews. Finally, the labeled data was split into training, validation, and test sets at a ratio of 8:1:1.

4.2. Experimental parameter settings

Due to hardware limitations, this study uses the 12-layer Chinese BERT pre-trained language model bert-base-chinese. Based on the experimental results, the optimal parameter settings are selected as shown in Table 1. To fully capture the complete information in each sentence, the maximum sentence length was set to 120, which corresponds to the maximum length in the movie review data. In line with methods that have shown the best performance on similar tasks, Adam was used as the optimizer, and cross-entropy loss was chosen as the loss function.

Table 1: Model parameter settings for sentiment analysis of movie reviews

Parameter	Value
Epoch	8
Learning rate	1e-5
batch_size	32
max_length	120

4.3. Evaluation metrics

Following previous research, four commonly used evaluation metrics were selected to comprehensively assess the model's performance: accuracy, precision, recall, and F1 score. Accuracy is the proportion of correctly classified samples out of the total samples. Precision refers to the proportion of truly positive samples among those predicted as positive. Recall measures the proportion of correctly predicted positive samples among all actual positive samples. The F1 score is the harmonic mean of precision and recall. The formulas for these metrics are as follows:

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

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall}$$

[2] https://www.kaggle.com

DOI: 10.54254/2755-2721/2025.23319

Where TP and TN denote the numbers of correctly predicted positive and negative classes, respectively. FP and FN refer to the numbers of incorrectly predicted positive and negative classes, respectively.

4.4. Experimental results and analysis

This study uses current mainstream deep learning models as baselines, replicating their frameworks according to relevant literature, and selecting the best-performing results on this dataset. The CNN model uses Word2Vec for word embedding and employs a convolutional neural network for feature extraction. The CNN–BiLSTM model also uses Word2Vec for word embedding, with CNN capturing local features of the text and bidirectional long short-term memory (BiLSTM) capturing global features. The comparison results are shown in Table 2.

Table 2: Comparison of experimental results

Model	Accuracy	Precision	Recall	F1
CNN	0.710	0.708	0.710	0.709
CNN–BiLSTM	0.713	0.714	0.712	0.712
Ours	**0.781**	**0.783**	**0.778**	**0.779**
%Improv.	6.8%	6.9%	6.6%	6.7%

The results demonstrate that, assuming balanced class distribution in the dataset, the proposed model achieves an accuracy 6.8% higher than CNN–BiLSTM and 7.1% higher than CNN. This intuitively reflects a significant improvement in the model's ability to judge overall data compared to the single neural network model CNN, and it also shows a notable improvement over the hybrid CNN–BiLSTM model. The proposed model leverages BERT for embedding layer transformation, dynamically adjusting word vectors based on varying contextual information. This offers better adaptability to the polysemous nature of Chinese words compared to static word embedding methods. Additionally, prompt learning utilizes deep semantic associations learned during the pre-training phase, alleviating the limitations of insufficient semantic modeling caused by training neural networks from scratch. This proves that the proposed method can more accurately extract semantic features from the text, achieving superior classification performance. From the F1 score, the proposed method improves by 6.7% over the next-best CNN–BiLSTM, indicating that prompt learning provides more precise recognition and differentiation of sentiment classes. This significantly reduces both false positives and false negatives, resulting in more stable overall performance.

4.5. Ablation study

To further validate the effectiveness of prompt-based learning, this section conducts an ablation study to examine the contribution of prompt learning to the overall performance, by comparing it with the fine-tuning paradigm for downstream tasks. The results of the ablation study are shown in Table 3.

Table 3: Ablation study results

Model	Accuracy	Precision	Recall	F1
BERT	0.758	0.757	0.759	0.755
Ours	**0.78**	**0.782**	**0.778**	**0.779**

A comparison of the experimental results from fine-tuned BERT and our method reveals that the accuracy of the fine-tuning paradigm is lower than that of prompt-based learning. This is because the

DOI: 10.54254/2755-2721/2025.23319

discrepancy between pretraining and downstream tasks in the fine-tuning paradigm hinders the model's ability to grasp the deeper semantic information contained in sentiment labels, limiting its capacity to fully leverage pretrained knowledge for text comprehension. After incorporating prompt learning, the model's accuracy, precision, and recall all improved by 2%–3%. Prompt-based learning, by adopting a cloze-style format consistent with the pretraining task, enables the model to reuse existing semantic knowledge and accurately capture the semantic relationships between sentiment keywords and their contextual information. Moreover, the prompt template provides rich contextual background, making it easier for the model to understand and process downstream tasks. This not only improves accuracy but also enhances the balance and stability of classification performance, as evidenced by the significant improvement in the F1 score.

In summary, compared to the baseline models, the proposed approach achieves the highest accuracy and outperforms others in terms of precision, recall, and F1 score, demonstrating that prompt learning can enhance both the accuracy and stability of the model, delivering outstanding performance on the Douban movie review dataset.

5. Conclusion and future work

This study constructs a sentiment analysis model for movie reviews based on the prompt learning paradigm, aiming to better leverage the semantics of labels and bridge the gap between pretraining and downstream tasks. By transforming the task into a cloze-style prompt format, the model directly utilizes its capability for masked language prediction. Experimental results on the Douban movie review dataset show that our model achieves the best performance across four metrics—accuracy, precision, recall, and F1 score—when compared with baseline models, demonstrating that prompt learning can more accurately capture semantic content and excels in the sentiment analysis of movie reviews.

Due to hardware limitations, this study employs a relatively small-scale language model and dataset. In the future, we plan to explore the performance of larger language models and conduct experiments using more comprehensive datasets. Additionally, we aim to evaluate the effectiveness of various prompt templates in movie review sentiment analysis, and further investigate how prompts containing different semantic cues impact the comprehension abilities of pretrained language models.

References

[1] *Kaplan A M, Haenlein M. Users of the world, unite! The challenges and opportunities of Social Media[J]. Business horizons, 2010, 53(1): 59-68.*

[2] *Pang B, Lee L. Opinion mining and sentiment analysis[J]. Found. Trends Inf. Retr., 2007, 2(1-2): 1-135.*

[3] *Hu M, Liu B. Mining and summarizing customer reviews[C]. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2004, 168-177.*

[4] *Pang B, Lee L, Vaithyanathan S. Thumbs up? Sentiment classification using machine learning techniques[C].Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, EMNLP 2002, 2002, 79–86.*

[5] *KIM Y. Convolutional Neural Networks for Sentence Classification[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, ACL, 2014, 1746–1751.*

[6] *Kalchbrenner N, Grefenstette E, Blunsom P. A convolutional neural network for modelling sentences[C]. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, ACL, 2014, 655–665.*

[7] *Liu P, Qiu X, Huang X. Recurrent neural network for text classification with multi-task learning[C]. Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, IJCAI/AAAI Press, 2016, 2873–2879.*

[8] *Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification[C]. Proceedings of the 2015 conference on empirical methods in natural language processing, EMNLP 2015, ACL, 2015, 1422–1432.*

DOI: 10.54254/2755-2721/2025.23319

[9] Devlin J. Bert: Pre-training of deep bidirectional transformers for language understanding[J].Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, ACL, 2019, 4171-4186.
[10] Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. OpenAI Technical Report, 2018.
[11] Liu P, Yuan W, Fu J, et al. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing[J]. ACM Computing Surveys, 2023, 55(9): 1-35.
[12] Nedjah N, Santos I, de Macedo Mourelle L. Sentiment analysis using convolutional neural network via word embeddings[J]. Evolutionary Intelligence, 2022, 15(4): 2295-2319.
[13] Liu Z, Zhang D, Luo G, et al. A new method of emotional analysis based on CNN–BiLSTM hybrid neural network[J]. Cluster Computing, 2020, 23: 2901-2913.
[14] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]. Proceedings of the 38th International Conference on Machine Learning, ICML 2021, PMLR, 2021, 8748-8763.
[15] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer[J]. Journal of machine learning research, 2020, 21(140): 1-67.
[16] Schick T, Schütze H. Exploiting cloze questions for few shot text classification and natural language inference[C]. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, ACL, 2021, 255-269.
[17] Qiu X, Sun T, Xu Y, et al. Pre-trained models for natural language processing: A survey[J]. Science China technological sciences, 2020, 63(10): 1872-1897.
[18] Schick T, Schütze H. Exploiting cloze questions for few shot text classification and natural language inference[C]. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, ACL, 2021, 255-269.