GHGA-Net: Global Heterogeneous Graph Attention Network for Chinese Short Text Classification

Meimei Li[1,2], Yuzhi Bao[1,2], Jiguo Liu[1,2(✉)], Chao Liu[1,2], Nan Li[1,2], and Shihao Gao[1,2]

[1] Chinese Academy of Sciences, Institute of Information Engineering, Beijing, China
liujiguo@iie.ac.cn
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

Abstract. As an important research content in the field of natural language processing, Chinese short text classification task has been facing two challenges: (i) existing methods rely on Chinese word segmentation and have insufficient semantic understanding of short texts; (ii) there is lacking of annotated training data in practical applications. In this paper, we propose the Global Heterogeneous Graph Attention Network (GHGA-Net) for few-shot Chinese short text classification. First, we construct the global character and keyword graph representations from the entire original corpus to collect more text information and make full use of the unlabeled data. Then, the hierarchical graph attention network is used to learn the contribution of different graph nodes and reduce the noise interference. Finally, we concatenate embedding with text vector and fuse the keyword and character features to enrich the Chinese semantics. Our method is evaluated on the Chinese few-shot learning benchmark FewCLUE. Extensive experiments show that our method has achieved impressive results in the classification tasks of news text and sentiment analysis, especially in minimal sample learning. Compared with existing methods, our method has an average performance improvement of 5% and less training consumption, which provides a new idea for few-shot Chinese natural language processing without relying on pre-training.

Keywords: Chinese short text classification · Few-shot learning · Heterogeneous graph · Hierarchical graph attention · Feature integrate

1 Introduction

Short text classification (STC) is applied in many research scenarios, such as sentence pair matching [1], news classification [2] and sentiment analysis [3]. Different from the long text which includes several paragraphs, short text generally only contain one or a few sentences. Due to its length limitation, short

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2024
F. Liu et al. (Eds.): PRICAI 2023, LNAI 14326, pp. 172–184, 2024.
https://doi.org/10.1007/978-981-99-7022-3_16

text cannot carry as rich semantic and grammatical information as long text. The fragmented text makes it difficult to obtain information beyond single word semantics, and it is almost impossible to understand text in combination with context. So STC task is much harder than long text when proper nouns appear in the text or some words have multiple meanings. Many studies based on graph neural network [4] aim to enrich the semantic information of short texts. The HGAT introduced [5] HIN structure builds graph based on the topic, entity and documents information and STGCN [6] uses words. However, topic acquisition and entity recognition methods cannot achieve high accuracy and requires additional training consumption. Others introduce part-of-speech (POS) tags [7] or use external wiki knowledge [8]. But these methods ignore the global text information in the original documents and have deviations in semantic understanding while Chinese texts carry more complex semantic information.

In natural language processing (NLP), the biggest difference between Chinese and English is that the character in English do not express meaning in most cases but Chinese did. For example, a text in TNEWS is "现实中的大司马是什么样的? (What is Da Sima like in reality?)", its category belongs to the game because "大司马(Da Sima)" is a game anchor. However, the "司马(sima)" was a type of official position in ancient China, and "马" is translated to horse directly. So that the complex meanings of Chinese words and characters are the biggest difficulty in Chinese STC and separate words from sentence in Chinese is much harder than English. The main way to solve this gap is to combine learning word and character features from Chinese text [2,9]. And there are also methods integrate sentences and words feature [10]. Lexicon [11] can match word through the tree structure more accurate, but it rely on external vocabulary.

General neural network methods [1] rely on large amount of training data to learn text features and perform poor while lacking labeled data. However, the cost of manually annotating all texts is unacceptable in practical STC tasks, while the extreme zero-shot learning rely heavily on pre-training and unable to adapt to multiple domains. In contrast, few-shot learning [12] only need a small amount of annotated texts and could achieve similar performance as normal.

To address the aforementioned problems, in this paper, we propose a **G**lobal **H**eterogeneous **G**raph **A**ttention Networks (GHGA-Net) for few-shot Chinese STC. By building the global heterogeneous graph, we make full use of the unlabeled texts information from entire original corpus to better fit few-shot learning. Then, we use the hierarchical graph attention networks to learn the contributions of different nodes to text categories and integrating word and character features to achieve deep understanding of the semantics of Chinese short texts.

The main contributions of this paper are summarized as follows :

- We propose the GHGA-Net method, which constructs heterogeneous graph to integrate keyword and character features to better represent the semantic information of Chinese short text. Graph attention mechanism is used to learn the contribution of different nodes and reduce noise interference.
- The unlabeled data are fully used by generating the global graph representation, which deeply collect the global semantic information of the original

data without pre-training and optimize the classification learning in the small number of annotation scenarios.
– The experimental results on the FewCLUE datasets show that the proposed method significantly improves the classification performance compared with other existing models.

2 Related Work

Graph Network for STC: In the text classification task, the global structure of the graph can model the complex semantic relationship between words in the text, and it is one of the most effective research methods to transform the text into a document graph and then use the graph neural network for learning and training. The graph convolutional neural network (GCN) [13] adds convolution operation to graph network, which can effectively compress the size of the model and increase the input size of text. The SHINE [7] model use GCN to combine documents, entities and position features, but it ignore the global information of short text. Addressing the lack of short text information, SimpleSTC introduce the external wiki text to enrich the global information, which benefits the STC task effectively. Attention mechanism is also applied to graph neural networks [14]. HyperGAT [15] introduces the concept of supergraph into text representation and uses dual attention mechanism to learn the nodes and edges of the graph respectively. Methods based on graph neural network can better represent the various feature information of short text. However, there is lacking in-depth research for Chinese STC based on graph neural network.

Pre-training for STC: In order to reduce training costs and adapt to more NLP tasks, pre-training models have been widely used in recent years [16–18]. These models are usually pretrained on large-scale corpora, enabling them to be more generalized and adaptable to few-shot learning scenarios. Thus, simply fine-tune the target dataset can achieve good results. However, most of pre-training models have large parameters and there are limitations in the actual deployment and operation process. Moreover, many models based on BERT has not made special optimizations in Chinese word segmentation, and it is still character segmentation, which hinders the understanding of Chinese semantics.

Chinese STC: Due to the particularity of Chinese text, the research based on integrate the word and character features of the text [9] has achieved good application results, and there are also methods to hierarchical learning sentence and words [10]. In addition, since the radicals of Chinese characters also belong to hieroglyphs, the radicals can also be added as a feature to the construction of Chinese text representation [2], but these methods are limited by embedding special word vectors. It is a valuable way to express Chinese text in the form of text map and integrate Chinese character and word features for learning.

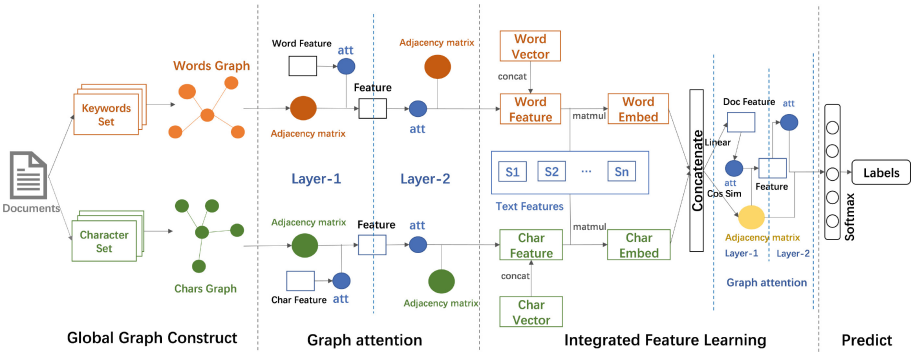


Fig. 1. The overall architecture of GHGA-Net Model.

3 Proposed Method

First, we give the task definition. For the given Chinese short text set $S_{doc} = (text)_n$ and its training set $S_{train} = (text, label)_m$, where $m << n$. Our goal is to train the classification model under the training set S_{train}, and finally predict the class label of remaining texts in S_{doc}.

The architecture of our GHGA-Net is shown in Fig. 1. Our idea is to extract the keywords and characters from each text in the whole corpus S_{doc} to construct the global graph representation. The hierarchical graph attention network is introduced to learn the graph features, which weighted the original graph representation and word embedding. Then our method fuse the heterogeneous features to document feature. Another hierarchical graph attention layer update the feature and the final prediction is made through softmax.

3.1 Global Graph Representation

In the case of only a small number of sample annotations, relying solely on training data to construct text features is clearly not enough. The unlabeled text in the original dataset can also be learned as implicit features to better obtain the semantic and category features of the text. So we choose to use the entire text set to construct the global graph representations.

Not all words contain specific information in Chinese. Therefore, we traverse each text in S_{doc}, extract and segment words of different parts of speech based on the term frequency-inverse document frequency (TF-IDF) and finally construct a global words vocabulary, only nouns, gerunds and some proper nouns under Chinese grammar are retained. Then, S_{doc} is cleaned according to the obtained global vocabulary. Next, we use point-wise mutual information (PMI) to calculate the word co-occurrence relationship between each keyword in vocabulary [13]. Let v_i, v_j be different keyword nodes in the global vocabulary, the relationship calculation method between them follows the following formula :

$$[C_{word}]_{ij} = max(PMI(v_i, v_j), 0) \tag{1}$$

C_{word} is a vector space with vocabulary length dimension in both rows and columns, which records the relationship between each node and other nodes in the global vocabulary. For each word in the global vocabulary, we match it with the pre-trained word2vec word vector to construct the word vector map.

According to the differences in grammar structure between Chinese and English. Besides word features, using character as the feature input of Chinese text classification can enrich the semantic and grammatical information of text. Therefore, we also propose and construct a global character vocabulary. For each short text in S_{doc}, we remove the numbers and symbols, only remain common words with word frequency above 10 and match with pre-trained character vectors. Similarly, the relationship $[C_{char}]_{ij}$ in character vocabulary is calculated by formula 1. Finally, we obtain the global features of keywords G_{gword} and characters G_{gchar} of documents with matched word vectors.

3.2 Hierarchical Graph Attention

In short text, not all words contribute same to the category information, especially in the case of lacking text information. To better focus on key features and reduce the interference of noise, we added the attention layers to update the weights of different nodes and perform weighted summation output.

For the constructed global heterogeneous graph representations C_{word} and C_{char}, the word vector V_{word} and V_{char}, we update the node vector H based on the two-layer graph attention networks:

$$H = GAT(C, ReLU(GAT(C, V))) \tag{2}$$

where RELU is the activation function, representing $[ReLU(x)]_i = max([x]_i, 0)$.

We directly introduce the pre-trained word vector here. Specifically, we regard the relation graph matrix as the input node vector, and the word vector embedding as the node feature. Performing a linear transformation on the node embedding $h_i^{(l)}$ in l-layer, similar to direct weighting in convolution operations [4], $W^{(l)}$ is a trainable weight parameter :

$$z_i^{(l)} = W^{(l)} h_i^{(l)} \tag{3}$$

Unlike concatenating the embedding of two nodes [14], our method uses a similar self-attention mechanism to calculate the original attention score for word nodes and character nodes respectively :

$$e_i^{(l)} = LeakyReLU(\vec{a}^{(l)T} z_i^{(l)}) \tag{4}$$

The attention weight is obtained by applying the softmax operation to the original attention score of the node. Finally, the features of all adjacent nodes are weighted and summed based on the attention weight:

$$a_i^{(l)} = \frac{exp(e_i^{(l)})}{\sum_{k \in N(i)} exp(e_k^{(l)})} \tag{5}$$

$$h_i^{(l+1)} = \sigma(\sum_{j \in N(i)} a_i^{(l)} z_i^{(l)}) \tag{6}$$

For the weighted word encoding, we concatenate it with the word vector again to make better use of the semantic information between words. Finally, the graph representation is transformed into character embedding and keyword embedding.

$$E = concat(E, E_{bedding}) \tag{7}$$

3.3 Integrated Heterogeneous Feature Learning

At last, we learn the text features based on global heterogeneous text graphs. For each content of the original Chinese short text, the text graph G_{text} is constructed by transforming the text into a vector. The word features of the text are encoded as aggregation nodes and embedded into h_{wi}. The text relationship after word segmentation is calculated by:

$$h_{wi} = \beth(E^T s_i), [s_i]_m = TF - IDF(v_m, x_i) \tag{8}$$

where T stands for matrix transpose operation, $\beth$ stands for regularized $x/||x||_2$, The TF-IDF vector is calculated by [19], where v_m represent the nodes in G_{gword} and x_i represent the nodes in G_{text}. Words in text but not belong to the global vocabulary will not be calculated. Note that the character feature encoding h_c is also calculated using the same way. The final fusion text representation is concatenate encoded for word embed and character embed:

$$h = concat(h_w, h_c) \tag{9}$$

Our original intention is to use a similar way to graph attention network, where word and character embedding are input as adjacent nodes, and an additional attention layer is added to achieve feature fusion. However, due to the difference between Chinese words and characters, the attention method did not lead in all test datasets, while the concatenate operation generally achieved good results. The specific ablation study will be discussed in Sect. 4.4.

After obtaining the fused text coding, we first use linear transformation to obtain the feature vector F of the text, and then calculate the adjacency matrix A of the text based on cosine similarity :

$$F = linear(h) \tag{10}$$

$$[A]_{ij} = ReLU(cos(h_i, h_j) - \tau) \tag{11}$$

τ is the correlation threshold, and the final text category prediction is also updated by the two-layer GAT method we proposed:

$$Prediction = SoftMax(GAT(A, ReLU(GAT(A, F))) \tag{12}$$

SoftMax represents $[softmax(x)]_i = exp([x]_i)/\sum_j exp([x]_j)$. Finally, we use the cross entropy as loss function for optimization process of the model.

$$Loss = -\sum_{i \in \iota_l} (y_i)^T log(y_i) \tag{13}$$

The complete procedure of GHGA-Net is described in Algorithm 1:

Algorithm 1: GHGA-Net Algorithm

Input: short text dataset S_{doc}, global graph set G, pretrained embedding E_{pre}
Output: predict label list $L = l_1, l_2, ..., l_n$ and trained model

```
for G=G_gword, G_gchar do
    update and generate the word embedding E_w and character embedding E_c by (2)
    for E=E_w, E_c do
        concatenate with the pretraind embedding by (7)
    end
end
for E=E_word,E_char do
    obtain the aggregated heterogeneous text graph feature by (8)
end
fuse the word and char embedding to final text embedding h by (9)
generate the text feature F and adjacency matrix [A]_ij by (10), (11)
update learning final text representation and predict the label by (12)
optimize model parameter by (13)
```

4 Experiments

4.1 Datasets

We conducted experiments on short text classification datasets from the Chinese few-shot learning benchmark FewCLUE [12] (Table 1):

1. **TNEWS**: The headline Chinese news short text classification dataset for few-shot learning tasks contains a total of 15 categories.
2. **EPRSTMT**: E-commerce product sentiment analysis dataset for sentiment polarity binary classification.

Table 1. Summary of used FewClue datasets.

Dataset	TrainSingle	TrainAll	DevAll	Classes	Unlabeled	LenAvg
EPRSTMT	32	160	160	2	20000	22
TNEWS	240	1185	1098	15	19565	36

4.2 Experimental Setup

BaseLines. We compare our method with the following three kinds of baselines:

- **General Method**: (1) **TextCNN**: Sentence classification method based on convolutional neural network [1]. (2) **BiLSTM-Att**: Bidirectional long short-term memory network with attention mechanism [20]. (3) **Transformer**: Encoder-decoder structure with multi-head attention [21].
- **Pre-training Model**: (1) **BERT**(Bert-wwm-Chinese): Pre-training model based on bidirectional Transformer architecture [16]. (2) **BERT-CNN**: Text encode by BERT and use CNN to train. (3) **RoBERTa**(RoBerta-wwm-Chinese): A robustly optimized BERT pre-training approach [18]. (4) **ERNIE**: Baidu's Pre-training model for Chinese natural language processing [17].
- **Graph Based Method**: (1) **HyperGAT**: Hypergraph attention neural network classification method based on LDA algorithm to extract text topics [15]. (2) **SimpleSTC**: GCN based short text classification method with external wiki knowledge [8].

4.3 Performance Comparison

Table 2 shows the performance. It can be seen that TNEWS is harder to classify due to its larger amount of categories. Our GHGA-Net achieves optimal results in almost all tasks and reaches an average improvement of about 5% compared with the second best baseline. Original methods achieve the worst average performance. All pre-training models perform well, and the RoBERTa model has achieved the highest accuracy on TrainAll set in TNEWS, which proves the advantages of using a large amount of corpus for pre-training in few-shot Chinese STC tasks. For graph based methods, HyperGAT performs obviously worse under small samples while SimpleSTC improves a little by external wiki knowledge. Besides, both of them are unable to deeply understand the complex semantics contained in Chinese. Our GHGA-Net is optimized for the semantic features of Chinese text, which integrates the heterogeneous graph features and introduce the hierarchical graph attention, receives the best result.

In the case of minimal training samples (TrainSingle), our method achieves state-of-the-art in both news multi-classification and sentiment binary classification tasks, which outperforms the second best baseline model by 6%. Almost all non pre-trained methods have a significant reduction in accuracy under extremely few-shot learning, which indicates their strong dependence on training

Table 2. Test performance (%) mesured on FewCLUE datasets. Normally trained under TrainAll data, the * mark represant trained in TrainSingle data. The best results are marked in bold, and the second-best results are underlined. The last row records the relative improvement of GHGA-Net over best results among other methods.

	TNEWS		EPRSTMT		TNEWS*		EPRSTMT*	
Model	ACC	F1	ACC	F1	ACC	F1	ACC	F1
TextCNN	44.08	44.36	50.78	50.78	23.59	21.18	49.38	47.89
BILSTM-ATT	14.30	9.18	41.41	40.97	7.74	3.28	45.62	45.52
Transformer	14.31	13.11	53.12	50.77	6.83	1.01	53.75	53.68
BERT	51.09	49.67	50.00	44.59	44.54	43.56	49.38	44.65
BERT-CNN	46.08	44.90	51.56	34.02	44.46	44.03	53.75	34.69
RoBERTa	**52.55**	51.16	49.22	45.12	45.26	44.69	48.13	44.85
ERNIE	51.73	51.13	47.66	34.71	42.17	40.22	46.88	35.58
HyperGAT	33.70	32.99	65.62	65.46	14.21	12.52	54.37	54.07
SimpleSTC	35.33	35.62	59.37	59.01	20.67	20.45	50.00	40.47
GHGA-Net	51.45	**51.91**	**68.75**	**68.03**	**47.17**	**47.13**	**58.74**	**57.63**
relative↑(%)	−2.13	1.47	4.77	3.93	4.22	5.46	8.04	6.58

data. Although the pre-training model has undergone a large amount of corpus training, there is still a gap in accuracy compared with our method. The results strongly prove the influential contribution of our global heterogeneous graph constructing based on the original documents information.

Table 3. Training cost compare with pre-training models. Evaluated in TNEWS dataset with 200 epochs.

Mode	Parameters	Hidden size	Layers	Times
BERT	102.28M	768	12	9 m 37 s
RoBERTa	102.28M	768	12	9 m 40 s
ERNIE	99.88M	768	12	5 m 02 s
GHGA-Net	0.605M	256	6	58.88 s

For training cost, we compare GHGA-Net with pre-training models. Table 3 shows the results. Our method has much fewer training parameters and less time consumption, but it achieves better performance. A lightweight structure makes GHGA-Net more efficient for real task and deployment.

4.4 Ablation Study

Recall that the proposed global heterogeneous graph and attention mechanism, we designed ablation experiments with different variants of GHGA-Net: (1)

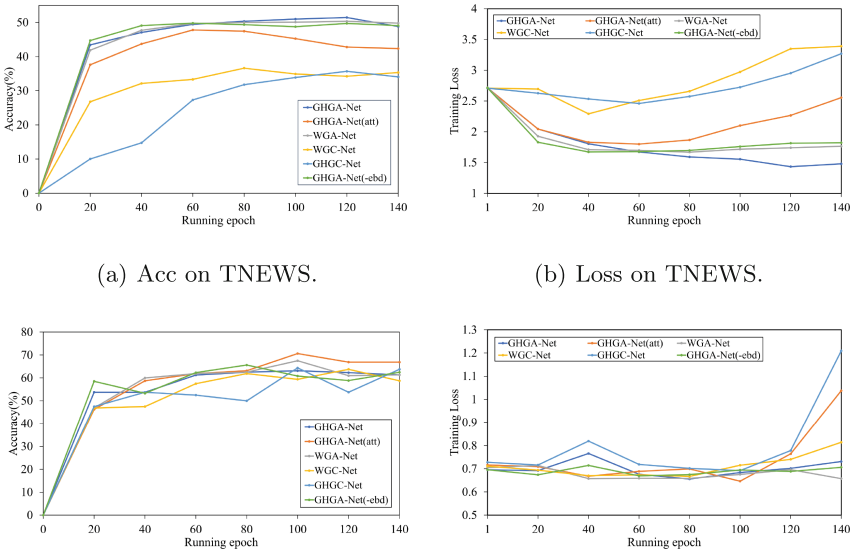
WGC-Net: without character features and attention layers, use GCN for training; (2) **HGC-Net:** without attention layers and use GCN for training; (3) **WGA-Net:** without character features; (4) **GHGA-Net(-ebd)**: using identity matrix instead of pre-trained vector in 2; (5) **GHGA-Net(fuse method)**: As mentioned in Chap. 3, besides concatenate operation in 9, we test the effect of linear interpolation and attention network for the fusion of features.

Table 4. Ablation Study (%) mesured on FewCLUE datasets. Trained under TrainAll set for 500 epoch. The - mark means unable to fit. The best results are marked in bold.

	TNEWS		EPRSTMT	
Method	ACC	F1	ACC	F1
WGC-Net	36.24	36.37	66.25	65.71
GHGC-Net	35.7	35.96	66.87	66.65
WGA-Net	50.36	51.08	67.5	66.61
GHGA-Net(-ebd)	49.82	49.90	68.12	67.66
GHGA-Net(linear)	–	–	46.25	31.62
GHGA-Net(att)	47.81	47.54	**70.62**	**70.28**
GHGA-Net(ours)	**51.45**	**51.91**	68.75	68.03

Table 4 lists the results, we can see the improvement of introducing pre-trained word vectors compared with initial encoding in both datasets. Figure 2(a) shows the significant effect of our proposed graph attention mechanism for graph representation learning. Compared with the graph convolution method, the accuracy rate is improved by more than 15%. Due to the fact that the simple convolution does not pay attention to all key category features. As can be seen from the loss curves in Fig. 2(b) and Fig. 2(d), with the increase of training rounds, the loss of ordinary convolution methods will rise, and the introduction of attention mechanism can effectively solve this problem. Among all the curves, our proposed GHGA-Net is the smoothest and also the most stable, which strongly proves that we have adopted the optimal method.

In terms of embedding fusion, the linear interpolation method has the worst performance, which indicates that the simple weighted average will lose the original information. As shown in Fig. 2(c), the attention-based fusion method achieves the best accuracy on the EPRSTMT dataset. Although the performance on the TNEWS dataset is slightly worse, it proves the feasibility of using neural network based methods to fuse text features. However, it cannot be ignored that with the increase of the number of training rounds, the accuracy rate of the att-fusion method has declined and the loss has increased, which may be caused by overfitting and needs further experiments in future research.



(a) Acc on TNEWS. (b) Loss on TNEWS.

(c) Acc on EPRSTMT. (d) Loss on EPRSTMT.

Fig. 2. Performance in the first 140 epoch training.

5 Conclusions

In this paper, we propose the GHGA-Net for Chinese STC without relying on pre-training. By constructing heterogeneous global graph, we can make full use of the unlabeled texts, and the finally feature fusion of character and word is more suitable for the classification task of Chinese text. Experiments results show that our method outperforms existed models on few-shot learning in Chinese STC scenario, especially in case of minimal training data. The additional ablation study strongly prove that our graph representation learning based on attention mechanism can effectively reduce the noise and highlight the key information. Despite those achievements, there are also some limitations to improve: (i) we have tested that remove some high frequency words in different domains may help reduce noise. (ii) we could create embedding by diagonal matrix for words out of vocabulary to capture rare semantics. (iii) radicals and some implied features of Chinese can be added to heterogeneous graph. (iv) we intend to adapt our hierarchical attention to transformer-like, which could further benefit the text feature learning. We will conduct in-depth research in future works.

Acknowledgement. This work was supported by the Research Funds for the Institute of Information Engineering, Chinese Academy of Sciences (No. BMKY2021B04, No. BMKY2023B04, No. E1R0141104).

References

1. Kim, Y.: Convolutional neural networks for sentence classification. Eprint Arxiv (2014)
2. Tao, H., Tong, S., Zhao, H., Xu, T., Jin, B., Liu, Q.: A radical-aware attention-based model for Chinese text classification. In: AAAI Conference on Artificial Intelligence (2019)
3. Wankhade, M., Rao, A.C.S., Kulkarni, C.: A survey on sentiment analysis methods, applications, and challenges. Artif. Intell. Rev. **55**, 5731–5780 (2022)
4. Defferrard, M., Bresson, X., Vandergheynst, P.: Convolutional neural networks on graphs with fast localized spectral filtering. In: NIPS (2016)
5. Hu, L., Yang, T., Shi, C., Ji, H., Li, X.: Heterogeneous graph attention networks for semi-supervised short text classification. ACM Trans. Inf. Syst. (TOIS) **39**, 1–29 (2019)
6. Ye, Z., Jiang, G., Liu, Y., Li, Z., Yuan, J.: Document and word representations generated by graph convolutional network and bert for short text classification. In: European Conference on Artificial Intelligence (2020)
7. Wang, Y., Wang, S., Yao, Q., Dou, D.: Hierarchical heterogeneous graph representation learning for short text classification. arXiv e-prints (2021)
8. Zheng, K., Wang, Y., Yao, Q., Dou, D.: Simplified graph learning for inductive short text classification. In: Conference on Empirical Methods in Natural Language Processing (2022)
9. Zhou, Y., Xu, B., Xu, J., Yang, L., Li, C., Xu, B.: Compositional recurrent neural networks for Chinese short text classification. 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI), pp. 137–144 (2016)
10. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E.H.: Hierarchical attention networks for document classification. In: North American Chapter of the Association for Computational Linguistics (2016)
11. Zhang, Y., Yang, J.: Chinese NER using lattice LSTM. ArXiv arXiv:1805.02023 (2018)
12. Xu, L., et al.: Fewclue: a Chinese few-shot learning evaluation benchmark. ArXiv arXiv:2107.07498 (2021)
13. Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. ArXiv arXiv:1809.05679 (2018)
14. Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio', P., Bengio, Y.: Graph attention networks. ArXiv arXiv:1710.10903 (2017)
15. Ding, K., Wang, J., Li, J., Li, D., Liu, H.: Be more with less: hypergraph attention networks for inductive text classification. In: Conference on Empirical Methods in Natural Language Processing (2020)
16. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. ArXiv arXiv:1810.04805 (2019)
17. Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., Liu, Q.: ERNIE: enhanced language representation with informative entities. In: Annual Meeting of the Association for Computational Linguistics (2019)
18. Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach. ArXiv arXiv:1907.11692 (2019)

19. Aggarwal, C.C., Zhai, C.: A survey of text classification algorithms. In: Mining Text Data (2012)
20. Tamekuri, A., Nakamura, K., Takahashi, Y., Yamaguchi, S.: Providing interpretability of document classification by deep neural network with self-attention. J. Inf. Process. **30**, 397–410 (2022)
21. Vaswani, A., et al.: Attention is all you need. In: NIPS (2017)