Neural Processing Letters (2024) 56:143
https://doi.org/10.1007/s11063-024-11595-z

DialGNN: Heterogeneous Graph Neural Networks for Dialogue Classification

Yan Yan[1] · Bo-Wen Zhang[2] · Peng-hao Min[1] · Guan-wen Ding[1] · Jun-yuan Liu[1]

Accepted: 12 March 2024

Abstract
Dialogue systems have attracted growing research interests due to its widespread applications in various domains. However, most research work focus on sentence-level intent recognition to interpret user utterances in dialogue systems, while the comprehension of the whole documents has not attracted sufficient attention. In this paper, we propose DialGNN, a heterogeneous graph neural network framework tailored for the problem of dialogue classification which takes the entire dialogue as input. Specifically, a heterogeneous graph is constructed with nodes in different levels of semantic granularity. The graph framework allows flexible integration of various pre-trained language representation models, such as BERT and its variants, which endows DialGNN with powerful text representational capabilities. DialGNN outperforms on CM and ECS datasets, which demonstrates robustness and the effectiveness. Specifically, our model achieves a notable enhancement in performance, optimizing the classification of document-level dialogue text. The implementation of DialGNN and related data are shared through https://github.com/821code/DialGNN.

Keywords Dialogue classification · Heterogeneous graph neural network · Pre-trained language model

1 Introduction

In recent years, dialogue systems have been prevalently applied in customer services, online health consultation, chatbots, etc. Dialogue classification, which aims at assigning predefined labels to an entire dialogue, is a fundamental task for many applications, including dialogue theme recognition, customer satisfaction analysis, service quality, etc. [1].

Most existing researches on classification in dialogue systems focus on the intent of users in each turn within a dialogue [2, 3]. These methods, taking the sentence-level user utterance as input and output of predicted intent, are not appropriate to classify the entire dialogues

✉ Bo-Wen Zhang
bwzhang@baai.ac.cn

1 Department of Artificial Intelligence, China University of Mining and Technology, Beijing 100083, China

2 Beijing Academy of Artificial Intelligence, Beijing 100084, China

Published online: 08 April 2024

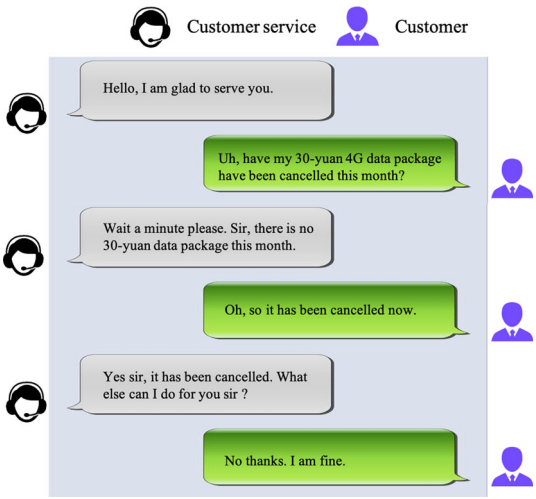

Fig. 1 An example dialog from telecom customer service, is defined as “consulting" rather than “business cancelling"

at document-level because sentences in dialogues are meant to be understood with the help of the context of all messages in the dialogue. The dependence on extended context requires that the classification process must regard a large block of utterances as input, which should be classified as a whole [4].

An intuitive solution to solve the above problem is to treat the whole dialogue as a document and use document classification methods. These methods either concatenate sentences into a long sequence [5, 6] or combine them hierarchically [7]. The main challenge is that a dialogue may contain multiple semantic topics, some of which are irrelevant to the business of the application task. Such irrelevant topics, regarded as noise, may be meaningless or misleading for classification models to predict correct results. For example in Fig. 1, the customer is consulting whether the 30 yuan data package is cancelled. The ground truth category should be business consultation. However, the mentioned canceling topic might mislead the models to identify the category of business cancellation. Therefore, existing models can hardly identify the noise in dialogues to determine the accurate categories.

We propose DialGNN, a generic framework based on heterogeneous graph neural networks for document-level dialogue classification. Firstly, a heterogeneous graph is constructed for each dialogue to represent the latent relationships among the sentences and the words within the dialogue. The sentences and words in each dialogue are regarded as nodes of different types in the graph. Then we combine graph neural networks and pre-training language models to learn latent representations of nodes and edges in the dialogue graph. During the messages passing over the graph, the representations of word-nodes and sentence-nodes are updated together, which helps to learn more implicit relationships among words and sentences. To validate the effectiveness, we conduct a set of experiments based on a public dataset[1] and an e-commerce customer service dataset contributed by ourselves. The comparison results show that the proposed method outperforms the sota methods.

[1] http://www.cips-cl.org/static/CCL2018/call-evaluation.html.

2 Related Work

2.1 Dialogue Classification

Dialogue classification involves assigning predefined labels to dialogues or their segments, such as utterances or turns, based on their functional or intentional significance within the conversation [8, 9].

Most existing studies on dialogue classification focus on sentence-level or utterance-level intent recognition of user statements [10]. These studies commonly employ hierarchical neural networks to model the sequential and structural information within words, characters, and utterances [11]. However, these approaches fail to explicitly account for the transition of speakers during the dialogue, which can impact the interpretation of dialogue acts. For instance, when speaker A poses a question, the subsequent utterance from speaker B is more likely to be an answer. Conversely, if the speaker remains the same, the following act is less likely to be an answer.

Tavabi [12] proposed to integrate the turn changes in conversations among speakers when modeling dialogue acts. They learned conversation-invariant speaker turn embeddings to represent the speaker turns in conversation; the learned speaker turn embeddings were then merged with the utterance embeddings for the downstream task of dialogue act classification. They showed that their model outperformed several baselines on three benchmark public datasets.

Another challenge for dialogue classification is that a dialogue may contain multiple semantic topics, some of which are irrelevant to the business of the application task [13]. In some cases, these topics may be irrelevant to the primary objective or business task of the application. This complexity arises from the natural flow of conversation, where participants may introduce unrelated or tangential subjects alongside the main focus of the dialogue. Consequently, accurately classifying dialogues requires the ability to identify and filter out irrelevant topics, ensuring that the assigned labels reflect the pertinent information and align with the specific objectives of the application.

Kumar [14] addressed this problem by augmenting small data to classify contextualized dialogue acts for exploratory visualization. They collected a new corpus of conversations, CHICAGO-CRIME-VIS, geared towards supporting data visualization exploration, and they annotated it for a variety of features, including contextualized dialogue acts. They applied data augmentation techniques to the training data, such as paraphrasing and back-translation, to increase the diversity and robustness of the data. They ran experiments with different classifiers and found that conditional random fields outperformed other methods.

Guo [15] recognized the importance of removing redundant information from dialogue text and thus adopted a long text segmentation method based on resampling, which solves the limitations of the BERT input length as well.

2.2 Heterogeneous Graph Network

For news classification, Kang [16] proposed a heterogeneous graph called News Classification Graph to represent the relationships between multiple news, such as their relevance in time, place and people. Moreover, they proposed a Joint Heterogeneous graph Network (JHN) to properly embed the News Classification Graph.

For aspect-based sentiment analysis, trying to capture the sentiment relationship among aspect terms, Niu [17] constructs a heterogeneous graph that models the inter-aspect relationships and aspect-context relationships simultaneously.

To combine multiple aspects of a review together and make use of the link between a sentence and its words, Yang [18] propose a dual-level attention-based heterogeneous graph convolutional network, including node-level and type-level attentions.

For short text classification, Yang [19] proposed a word-concept heterogeneous graph convolution network to avoid regarding introduced concepts as noises and learn the representations with interactive information. Kong [20] considers the lack of labeled data. Adopting an uncertainty-aware mechanism, they proposed a heterogeneous graph attention network. Furthermore, the lack of context, the sparsity of short text features and the inability of word embedding and external knowledge bases to supplement short text information are also challenges for short text classification. Aiming to improve classification accuracy and reduce computational difficulty, Zhang [21] built a text, word and POS tag-based graph convolutional network which does not require pre-training word embedding as initial node features.

For utterance-level dialogue classification, many graph-based methods are applied to capture the implicit feature information in the dialogue structure, Qin [6] designed the co-interactive graph interaction layer to capture contextual information and interaction information, which are important information hidden in dialogue. Shen [22] proposed a neural network based on directed acyclic graph to better represent dialogue information flow and combine the advantages of graph neural network and recurrent neural network models.

For dialogue-level dialogue classification relevant models in this field are scarce, but graph-based models are still concerned by relevant researchers. Pang [23] regarded speakers, local discourses and utterances as main information and used graphs to model them to construct a multi-factor graph.

3 Methodology

DialGNN encompasses three essential modules that collectively contribute to its functionality: DialGraph Construction, Node Representation, and Heterogeneous Graph Network, as illustrated in Fig. 2.

The DialGraph Construction module plays a crucial role by transforming a given dialogue into a heterogeneous graph. This graph enables the capturing of intricate relationships among words, sentences, and the overall dialogue structure. By representing the dialogue in this manner, DialGNN gains a comprehensive understanding of its underlying dynamics.

Thus, with a pretrained BERT model to initialize the representation of nodes in DialGraph, DialGNN takes the contextual semantics into consideration and can effectively combine the contextual information as well as the structures within dialogues. The Node Representation module within DialGNN undertakes the task of initializing the node representations within the DialGraph. This is achieved by employing BERT-based embeddings, which are pretrained contextual representations capable of capturing rich semantic information. Through this initialization process, the Node Representation module equips the graph with meaningful and informative node representations.

The final module, Heterogeneous Graph Network, is responsible for encoding the heterogeneous graphs generated by DialGraph Construction. It employs graph attention networks to capture relevant dependencies and interactions among nodes within the graph. By updating the representations of the nodes based on these learned relationships, the Heterogeneous

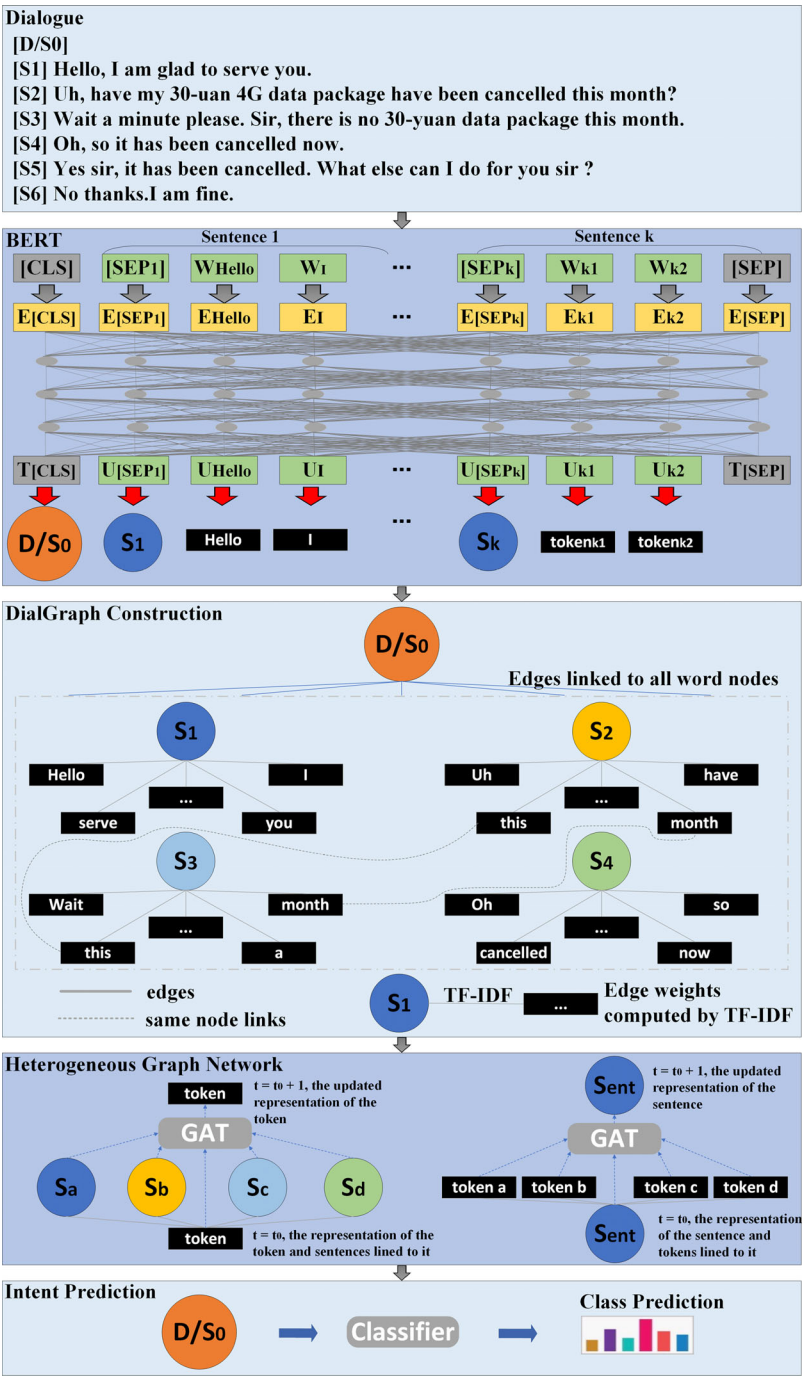

Fig. 2 The architecture of DialGNN framework. The forward process of an example dialogue (with 6 sentences) contain 4 stages: initialization of representation with BERT, construction of DialGraph, update node representation with Heterogeneous Graph Network, and classification of dialogue node to perform intent prediction

Graph Network module enhances the graph's ability to handle downstream tasks effectively. Now let's delve into the detailed descriptions of each section.

3.1 DialGraph Construction

There are several efforts to convert a dialogue into a topological graph [24]. They majorly regard each sentence as a node and construct a homogeneous graph where the edges between nodes are formed with contextual relations. That is, only sentences within a fixed window size have edges. Such methods might fail to capture the relations among sentences with long distance and may ignore the impact of some words which significantly contribute to the predicted categories.

To this end, we construct a heterogeneous graph named DialGraph with word nodes, sentence nodes and dialogue nodes. The edges between sentence nodes and word nodes represent the containing relations. Then more implicit relations among different sentences can be derived from the relations between sentences and words, such as the co-occurrence, semantic distance, and term frequencies. Inspired by the usage of [CLS] tag in BERT, we add a 0^{th} sentence node as the dialogue node, rather than using a pooling layer of sentence node embedding. Formally, the heterogeneous graph DialGraph is defined as follows.

Given a dialogue $C = \{s_1, s_2, \ldots, s_n\}$, the DialGraph is denoted as $G = \{V, E\}$, where $V = V_w \cup V_s \cup V_c$, $E = \{e_{10}, e_{11}, \ldots, e_{mn}\}$ represents the node set and the edge set respectively. Here, $V_w = \{w_1, w_2, \ldots, w_m\}$ denotes m unique words, V_s corresponds to the n sentences, V_c is the dialogue node. E is a real-value edge weight matrix and $e_{ij} (i \in [1, m], j \in [0, n])$ indicates the j^{th} sentence contains the i^{th} word. Note that the dialogue node V_c connects to all word nodes.

The node updates in DialGNN are determined by considering the features of neighboring nodes and the associated edge weights. In this regard, the word nodes update their representations based on the features and edge weights of the corresponding sentence nodes. Similarly, the sentence nodes update their representations by considering the features and edge weights of the word nodes connected to them. Furthermore, the dialogue nodes update their representations by incorporating the features and edge weights of the sentence nodes connected to them. This approach ensures that the node representations in DialGNN are iteratively refined, taking into account the contextual information from neighboring nodes and their respective edge weights.

3.2 Node Representation

We denote $\mathbf{X}_\mathrm{w} \in \mathbb{R}^{m \times d_w}$, $\mathbf{X}_\mathrm{s} \in \mathbb{R}^{n \times d_s}$, and $\mathbf{X}_\mathrm{c} \in \mathbb{R}^{n \times d_c}$ as the input feature matrices representing word, sentence, and conversation nodes, respectively. Here, d_w, d_s, and d_c refer to the dimensions of the word embeddings, sentence representation vectors, and dialogue representation vectors, respectively.

Here, we use BERT-based [25] embeddings to get the initialized representations of the words, the sentences and the dialogue. Note that other embedding models and other pre-trained language models can also be utilized.

To incorporate the varying importance of relationships between nodes, we employ TF-IDF (Term Frequency-Inverse Document Frequency) values to initialize the weights of the edges. TF-IDF is a statistical measure commonly used in natural language processing to evaluate the significance of a term in a document relative to a collection of documents. By assigning TF-

IDF values as the edge weights, we can capture the importance of the connections between nodes in the graph structure.

3.3 Heterogeneous Graph Network

Given the constructed DialGraph with node features $\mathbf{X}_{\mathrm{w}} \cup \mathbf{X}_{\mathrm{s}} \cup \mathbf{X}_{\mathrm{c}}$, we leverage the graph attention networks [26] to update the representations of nodes.

We refer to $h_i \in \mathbb{R}^{d_h}, i \in [0, m+n]$ as the hidden states of input nodes. The graph attention(GAT) layer is designed as follows:

$$\alpha_{ij} = \mathrm{softmax}(\mathrm{LeakyReLU}(W_a[W_q h_i; W_k h_j])) \tag{1}$$

$$u_i = \sigma\left(\sum_{j \in N_i} \alpha_{ij} W_v h_j\right) \tag{2}$$

where W_a, W_q, W_k, W_v are learnable linear transformation matrices and α_{ij} is the attention weights between h_i and h_j. The multi-head attention can be denoted as follows:

$$u_i = \|_{k=1}^{K=1} \sigma\left(\sum_{j \in N_i} \alpha_{ij}^k W^k h_i\right) \tag{3}$$

Furthermore, we also add a residual connection to avoid gradient vanishing. Therefore, the final output can be formulated as follows:

$$h_i^{'} = u_i + h_i \tag{4}$$

Besides, we modify the GAT layer to infuse the scalar edge weights e_{ij}, which are mapped to the multi-dimension embedding. Hence, the Eq. 1 is modified as follows:

$$z_{ij} = \mathrm{LeakyReLU}(W_a[W_q h_i; W_k h_j; e_{ij}]) \tag{5}$$

After each GAT layer, we introduce a feed-forword network that includes two linear project layer as Transformer [27].

$$\mathrm{FFN}(x) = \mathrm{ReLU}(x W_1 + b_1) W_2 + b_2 \tag{6}$$

3.4 Training and Optimization

During the training stage, the representations of the dialogue node, sentence nodes and word nodes are updated alternately. Since the dialogue node can be regarded as the 0^{th} sentence connected with all words. The process of updating the dialogue node is the same as the process of updating the sentence nodes. Thus, one iteration of the training includes a sentence-to-word update process and a word-to-sentence update process.

In the sentence-to-word update process, the dialogue node and the sentence nodes are updated in the t^{th} iteration based on their connected word nodes via the GAT and FFN layer as follows:

$$U_{s \leftarrow w}^{t+1} = \mathrm{GAT}(H_s^t, H_w^t, H_w^t) \tag{7}$$

Table 1 The statistics of CM and ECS datasets

Dataset	Labels	Train	Dev	Test	Avg Len	Max Len
CM	37	15,791	1996	1997	680	5059
ECS	14	68,759	8697	8698	5023	98749

$$H_s^t = \mathrm{FFN}(U_{s\leftarrow w}^{t+1} + H_s^t) \tag{8}$$

where $H_w^0 = \mathbf{X}_\mathrm{w}$, $H_s^0 = \mathbf{X}_\mathrm{s}$ and $U_{s\leftarrow w}^1 \in \mathbb{R}^{m\times d_h}$. GAT() denotes that H_s^0 is used as the attention query and H_w^0 is used as the key and value.

Then in the sentence-to-word update process, the word nodes are updated through the new dialogue node and the sentence nodes.

$$U_{w\leftarrow s}^{t+1} = \mathrm{GAT}(H_w^t, H_s^t, H_s^t) \tag{9}$$

$$H_w^{t+1} = \mathrm{FFN}(U_{w\leftarrow s}^{t+1} + H_w^t) \tag{10}$$

Finally, classification for the dialogue node determines the label of the whole dialogue and cross-entropy loss is used to optimize the model [28].

4 Experiments

In this section, we perform several experiments to assess and analyze the effectiveness of our proposed dialogue classification approach. Our objectives are to address the following research questions:

- How does our approach compare with existing methods on the dialogue classification task? (Section 4.3.1)
- How does the heterogeneous graph information affect the dialogue classification performance? (Section 4.3.2)
- What are the contributions of each component in our approach? (Section 4.3.3)

4.1 Datasets and Experiment Settings

We use two datasets for our experiments: China Mobile Dataset (CM) and E-commerce Customer Service Dataset (ECS). CM is a dataset of phone call dialogues between customers and service staff, where the goal is to identify the business type requested by the customers. ECS is a dataset of online chat dialogues between customers and sellers, staff or AI systems, where the goal is to classify the dialogue acts or emotions. The statistical information of them is shown in Table 1.

China Mobile Dataset(CM). This dataset assumes a scenario where the customer service staff answer the phone calls from different customers. The aim is to determine which business is the accurate request of the calls given the whole dialogue history. The contents are the ASR texts from customer service dialogues of phone calls. The labels are pre-defined business types. The dataset contains 19,784 labeled conversation segments, with 37 different human–machine dialogue intent categories. Table 2 shows the business and conversation intention types.

Table 2 Business types and conversation intentions

Business types	Conversation intentions
Consulting	Subscription information inquiry, regulations, tariff, processing mode
	Account information, marketing activities information
	Number status, e-commerce product information, etc
Processing	Download/setup, cancel, move/install/uninstall
	Change, open, print/mail, pay, cancel/reopen
	Replace/exchange, reset/modify/reissue, etc
Complainting	Uninformed customization, business usage, business processing
	Business regulations dissatisfaction, information security
	Network problems, marketing problems, cost problems, etc

Table 3 Samples of ECS dataset

Dial Id	Sentence Id	Sentence Info	Category
07dc2	1	text: 我是您的客服,很高兴为您服务。	异常评论: 评论泄露隐私 (Abnormal comments: Comments Leak Privacy)
		(I'm your service guy	
		I'm happy to help you.)	
		id: 1	
		member_type: 3	
	2	text: 订单的评论泄露了我的手机号。	
		(The review of the order	
		gave away my phone number.)	
		id: 2	
		member_type: 1	
	3	text: 请发送给我订单编号。	
		(Send me the order number.)	
		id: 3	
		member_type: 2	

E-commerce Customer Service Dataset (ECS). ECS is contributed by ourselves to the community. The dialogues took place between a customer and a seller, a staff, or an AI system. The user goal is relatively straightforward, that is, to complain about an unsatisfied experience. The labels are event types, such as malicious refunding, counterfeit, right infringement, etc.

Table 3 displays a representative ECS sample, where each column represents distinct elements. The first column serves as a unique dialogue key. The second column contains a sequence of sentence IDs associated with the dialogue. The third column comprises a JSON list with keys such as "id" (the JSON ID linked to the sequence), "text" (the sentence content), and "member_type" (1 for customer, 2 for customer service, 3 for automatic AI customer service). The fourth column indicates the dialogue category, classified into coarse and fine levels.

We adopt several widely used evaluation metrics, which are accuracy and F1-score, to evaluate the performance of DialGNN. We choose categorical cross entropy as the loss

function for our model on two datasets. The learning rate (lr) is set to $5 \times e^{-5}$ and batch size is set to 128. We set the multi-head graph attention network 3 layers, 2 heads and 1024 hidden units. For embeddings, we take 768-dimensional BERT and Chinese RoBERTa embeddings as word embeddings. To increase the training and inference efficiency, we adopted mini-batch processing by processing a subset of the data simultaneously during training. This allows for parallel computation across different graph instances, further enhancing efficiency. All the codes for our experiments are available on Github (https://github.com/821code/DialGNN).

4.2 Baselines

To evaluate the performance of our proposed framework, we compare it with several baseline models that use different sequence encoders.

- **TextRNN** [29] is a type of recurrent neural network that can handle text data and take into account the order of words. TextRNN uses a recursive way to pass the output of the previous time step as the input of the current time step, so as to transfer the context information to the next time step.
- **TextRNN-Att** [30] is a text classification model that combines recurrent neural networks (RNNs) and attention mechanisms. TextRNN-Att uses a bidirectional RNN to encode the input text into hidden states, and then applies an attention layer to aggregate the hidden states into a sentence representation. The attention layer assigns different weights to different parts of the text, depending on their relevance to the classification task.
- **TextCNN** [31] short for Text Convolutional Neural Network, is a deep learning model designed for text classification and sentiment analysis tasks. TextCNN can handle variable-length sentences and learn complex semantic features from the text.
- **CNN-LSTM** [32] is a widely-used model consisting of regional CNN and LSTM. By combining the regional CNN and LSTM components, the CNN-LSTM model can leverage both the local spatial information captured by the CNN and the sequential dependencies captured by the LSTM. This hybrid approach allows the model to effectively extract meaningful features from input data and capture complex relationships within sequential data.
- **BERT** [25] is a transformers-based language model, which is pre-trained on large-scale corpus and has achieved remarkable success in many NLP tasks. The use of transformers in BERT enables it to capture contextual dependencies in a more comprehensive manner. The transformer architecture utilizes attention mechanisms to weigh the importance of different words in a sentence based on their relevance to each other. This attention mechanism allows BERT to consider the entire context when representing a word, rather than just relying on its immediate neighbors.
- **Roberta** [33] is a robustly optimized version of BERT, a pre-trained language model that uses bidirectional transformers to learn contextual representations of text.
- **ERNIE** [34] stands for Enhanced Representation through kNowledge IntEgration, which indicates its ability to incorporate various types of knowledge into the pre-training process of language models.
- **Han** [7] is a hierarchical attention network containing two levels of attention mechanisms applied at word-level and sentence-level, which is similar to the graph. The hierarchical nature of Han's attention mechanisms allows it to effectively model relationships and dependencies between words and sentences. The model can capture not only the local interactions between words but also the broader interactions and contextual dependencies between sentences.

- **DAG** [22] is an acronym for Directed Acyclic Graph. DAG can be used to model the structure and context of a conversation, where each node represents an utterance and each edge represents the dependency or influence between utterances. DAG can capture the information flow and the long-distance dependencies in a conversation.
- **InductGCN** [35] constructs a graph based on the statistics of training documents only and represents document vectors with a weighted sum of word vectors. It then conducts one-directional GCN propagation during testing.
- **TextGCN** [36] incorporates semantic information and relationships from text data by constructing a text graph and applying graph convolution operations. It can perform text classification without the need for external embeddings, making it a valuable approach for specific text classification tasks.
- **AttentionXML** [37] introduced an attention mechanism and a probabilistic label tree (PLT). Attention mechanism ensures that the model captures the subtle and context-dependent associations between text and labels, enhancing classification accuracy.

4.3 Results and Analysis

4.3.1 DialGNN Comparing with Baseline Methods

Table 4 presents the performance evaluation of different models on two distinct datasets, CM and ECS, for dialogue classification. Notably, DialGNN(BERT) denotes the amalgamation of the BERT model with the DialGNN structural framework, referred to subsequently as DialGNN.

On the CM dataset, DialGNN demonstrates remarkable performance with 70.2% accuracy and 59.3% F1 score. This outcome underscores the DialGNN structure's innate capacity to capture intricate linguistic patterns embedded within dialogues. Meanwhile, the ECS dataset reveals similar excellence, with an accuracy rate of 60.3% and an equally robust F1 score of 54.9%. This success can be primarily attributed to the indispensable role played by the DialGNN structure in facilitating the modeling of intricate dependencies and contextual information across the sequence of dialogue turns.

The results obtained underscore the inherent limitations of conventional baseline models, such as TextRNN, TextCNN, and CNN_LSTM. These models, rooted in their sequential processing architecture, consistently display a comparatively inferior performance on both the CM and ECS datasets.

Comparing to BERT, Roberta, and ERNIE, DialGNN outperforms them due to its heterogeneous graph architecture designed for dialogue classification, especially in understanding the flow of conversations, tracking changing topics, and capturing user intents that evolve across sentences.

Futhermore, the underperformance of the DAG model on the ECS dataset can be attributed to a fundamental mismatch between the model's design and dialogue classification task. The DAG model primarily operates at the sentence level, focusing on classifying individual sentences, while the datasets in question require dialogue-level classification.

InductGCN's performance on both the CM and ECS datasets is relatively lower than DialGNN. Its limitations may stem from its capacity to capture nuanced linguistic patterns, contextual dependencies, and user intents in dialogues, which are pivotal in dialogue classification tasks.

In the case of dialogue text, it contains interactions and transitions between speakers, and TextGCN's design does not explicitly address the complex relationships between speakers.

Table 4 Comparsion with baseline methods on CM and ECS datasets

Models	CM		ECS	
	Accuracy	F1	Accuracy	F1
TextRNN	58.3	34.4	54.7	48.4
TextRNN-Att	61.5	39.0	57.7	51.9
TextCNN	60.6	43.4	56.8	50.8
CNN_LSTM	47.4	28.0	55.2	50.0
BERT	69.2	53.8	59.6	54.8
Roberta	67.7	53.9	58.8	53.0
ERNIE	67.7	50.6	57.9	51.2
Han	62.7	46.6	58.5	53.3
DAG	11.5	1.6	14.3	3.3
InductGCN	43.7	35.7	49.1	42.7
TextGCN	43.3	28.6	49.9	42.9
AttentionXML	30.7	30.9	43.1	43.1
DialGNN (BERT)	**70.2**	**59.3**	**60.3**	**54.9**

It tends to deal with individual sentences or text blocks and may not capture the context and relationships between speakers adequately, limiting its ability to understand and model the overall semantics of the dialogue.

As for AttentionXML, if there are weak label correlations or label relationships that have minimal impact on classification in a given dataset, the model might not fully leverage the advantages of the attention mechanism. AttentionXML relies on learning label relationships to better capture the relationships between text and labels. If these relationships are not significant in the dataset, the model's performance could be constrained.

Crucially, DialGNN incorporates nodes at different levels of semantic granularity, allowing for flexible integration of various pre-trained language representation models. The use of BERT within the graph structure endows DialGNN with powerful text representational capabilities.

To further validate the generalizability of DialGNN, Table 5 presents the comparison results of different sequence encoders with and without DialGNN. We can find that for all baseline models, combining with DialGNN achieves significant improvements. Even on the strong baseline model BERT, the DialGNN gains 5.5% and 6.4% F1 scores on the CM and ECS datasets, respectively.

As shown in Table 5, models with DialGNN-seg have better performance on the ECS dataset. DialGNN-seg refers to a trick to handle too-long dialogues. The BERT model restricts the maximum length of the input sequence to 512 due to computational issues. For a dialogue from the ECS dataset with more than 512 tokens, we truncate it into 512 tokens in the basic DialGNN settings. In the DialGNN-seg settings, we obtain the initial embeddings by sliding a context window of 512 tokens. So the DialGNN-seg incorporates more contextual information into node embeddings and gains a better performance.

4.3.2 Comparisons on Graph Designs

Table 6 displays the performance evaluation of various graph designs that utilize pre-trained models on CM Dataset. The comparison groups consist of different design variations,

Table 5 The performance comparisons of baseline models and combining DialGNN

Models	CM		ECS	
	Accuracy	F1	Accuracy	F1
CNN_LSTM	47.4	28.0	55.2	50.0
+ DialGNN	**54.1**	**41.8**	**57.2**	**52.3**
Han	62.7	46.6	58.5	53.3
+ DialGNN	**63.1**	**48.1**	59.1	54.3
+ DialGNN-seg	–	–	**63.2**	**59.3**
BERT	69.2	53.8	59.6	54.8
+ DialGNN	**70.2**	**59.3**	60.3	54.9
+ DialGNN-seg	–	–	**65.4**	**61.2**

Table 6 The results of Different Graph Designs

Models	Accuracy	F1
Base(BERT-tiny)	65.1	41.0
+ context graph	63.2	38.6
+ DialGNN	**68.5**	**57.4**
Base2(BERT)	69.2	53.8
+ async init	60.9	41.2
+ sent nodes agg	61.9	45.4
+ DialGNN	**70.2**	**59.3**

including those with context relation modeling (specifically DialogueGCN, which requires a substantial amount of GPU memory, and BERT-tiny as the base model), asynchronous initialization, and designs without a dialogue node.

The results of the comparison reveal that alternative graph designs tend to compromise the quality of the latent representations provided by pre-trained models. This suggests that the mentioned designs are not able to effectively capture and incorporate the contextual relationships present in the data. In particular, the performance metrics indicate that these alternative designs result in a degradation of the pre-trained model's ability to represent and understand the underlying patterns in the China Mobile Dataset.

These findings highlight the importance of preserving the latent representation quality obtained from pre-trained models when designing graph structures for natural language processing tasks. The results strongly suggest that the approaches involving context relation modeling, asynchronous initialization, and the inclusion of dialogue nodes are essential for maintaining the integrity and effectiveness of the pre-trained models when applied to the China Mobile Dataset. By utilizing these design elements, the models can leverage the full potential of the pre-trained representations and achieve better performance in capturing the intricacies of the dataset.

4.3.3 Ablation Study

To validate the contribution of each component, a series of experiments are designed to observe the performances and the results are summarized in Table 7, "w/o" indicates that this component is not included in the model.

Table 7 The results of Ablation study on CM Dataset

Models	Accuracy	F1
w/o TF-IDF	68.7	53.6
w/o sent-word update	69.4	55.8
w/o word-sent update	69.2	52.0
BERT + DialGNN	**70.2**	**59.3**

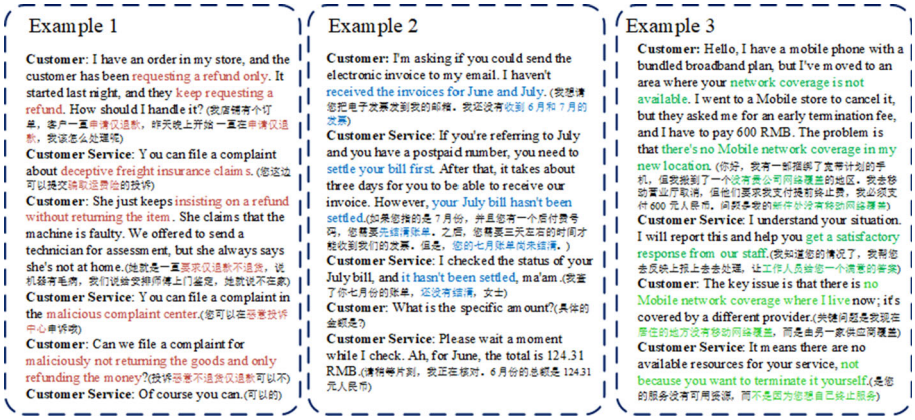

Fig. 3 The Examples of Case Study on CM and ECS Datasets. Keywords in different examples are marked in color respectively

The results presented in the table demonstrate that the TF-IDF initialization approach for edge weights significantly enhances the overall performance of the system. This initialization technique, which utilizes TF-IDF algorithm, provides a valuable foundation for establishing the weights of connections between nodes in the graph structure.

Moreover, both the sentence-to-word updating step and the word-to-sentence updating step have been found to play crucial roles in the functionality of the DialGNN system. These steps are integral in facilitating the flow of information and the exchange of knowledge between sentences and words in the graph. By ensuring a bidirectional and iterative updating process, these steps enable the system to capture and incorporate relevant information from both sentence-level and word-level representations.

The findings underscore the significance of each component in the overall performance of the system. The TF-IDF initialization contributes substantially to the quality of the edge weights, enhancing the accuracy and effectiveness of the system. Additionally, the sentence-to-word updating step and the word-to-sentence updating step are deemed essential for the optimal functioning of DialGNN, enabling seamless information propagation and integration between sentence and word representations.

4.4 Case Study

Drawing upon the preceding model analysis, a case study was conducted utilizing samples extracted from the dataset. As depicted in Fig. 3, Table 8 presents the classification outcomes achieved by DialGNN in comparison to other models. Remarkably, DialGNN demonstrated unerring classification accuracy for all the examples. DialGNN's superiority resides in its

Table 8 Classification results of different models for Examples. Bold indicates correct classification

Model	Example		
	1	2	3
DialGNN	Malicious non-returns only refunds 恶意不退货仅退款	Reset/modify /reissue 重置/修改/补发	Dissatisfaction with operational requirements 业务规定不满
Roberta	Fraudulent shipping insurance 骗取运费险	Business use issues 业务使用问题	Removal/installation /dismantling 移机/装机/拆机
ERNIE	Fraudulent shipping insurance 骗取运费险	Business process issues 业务办理问题	Removal/installation /dismantling 移机/装机/拆机
TextRNN	Exploit evaluation blackmail 利用评价要挟	business requirement 业务规定	Product/Business Functions 产品/业务功能
TextRNN_Att	Exploit evaluation blackmail 利用评价要挟	Business process issues 业务办理问题	Product/Business Functions 产品/业务功能
TextCNN	Exploit evaluation blackmail 利用评价要挟	Business process issues 业务办理问题	Modification 变更
InductGCN	Fraudulent shipping insurance 骗取运费险	Print/Mail 打印/邮寄	Removal/installation /dismantling 移机/装机/拆机

capacity to transcend mere feature-based word-level classification. It distinguishes itself by comprehending the nuanced semantics of words within the broader context and the underlying dialogue structure.

Referring to Fig. 3, Example 1 showcases a distinctive dialogue structure. In the conversation, both the customer and the customer service initially hold different interpretations of the event, which evolve over the course of the discussion. DialGNN inspired by BERT, introduces a novel approach by incorporating a “0th sentence node" as a dialog node. This innovative addition allows the model to integrate various discourse features within the entire dialogue, leading to a more comprehensive comprehension of the dialogue at the macro level. In this specific case, DialGNN can synthesize the customer’s descriptions of the event, thus obtaining a deeper understanding. Consequently, it can accurately identify the correct label, even when faced with conflicting opinions, which distinguishes it from other models that may primarily focus on direct opinion information.

Moving on to Example 2, while other models may be influenced by the frequency of terms like ‘July’, ‘Bill’, ‘Settlement’, and quantities mentioned in intermediate exchanges, they could interpret the core intent as a general ‘Business Processing’. In contrast, DialGNN leverages its dialog structure representation through a graph and an attention mechanism. It recognizes that the conversation’s beginning and end are critical segments since they mark

the initiation and conclusion of inquiries. While other models may overlook these subtle cues, DialGNN focuses on the dialog's core intent concentration areas and accurately identifies it as an 'Invoice Reissuance' issue. This underscores DialGNN's strength in capturing contextual nuances within dialogues.

Finally, Example 3 showcases a dialogue containing vital contextual semantics. While other models tend to reinforce associations between strongly characterized words like 'move', 'area', and 'location' and tagged words such as 'install' and 'moving', DialGNN takes a distinctive approach. It capitalizes on the powerful contextual semantic features embedded within pre-trained models, seamlessly integrating information across sentences and words through interactive updates within its graph structure. Unlike other models that may categorize based on individual word characteristics, DialGNN excels in understanding words within the context and dialogue background. This deep comprehension is evident in this case, where it discerns the customer's dissatisfaction with the business office from the event description and interactions with customer service.

5 Conclusion

In summary, our work introduces the innovative DialGNN framework, which leverages heterogeneous graph neural networks to gain a deeper understanding of multi-turn dialogues. Our framework offers versatile compatibility with various encoders and demonstrates the potential to enhance their performance, even when used in conjunction with pre-trained language models. The extensive array of experiments conducted showcases the efficacy of DialGNN in the context of dialogue understanding. The ability of DialGNN to capture nuanced linguistic patterns, contextual dependencies, and evolving user intents across dialogues sets it apart as a robust and adaptable framework with the potential to advance a myriad of real-world applications. Our study serves as a pivotal step in the ongoing evolution of dialogue systems, paving the way for enhanced dialogue comprehension, customer satisfaction analysis, service quality assurance, and dialogue topic categorization.

Author Contributions Yan Yan and Bo-wen Zhang conceived the idea and designed the DialGNN framework. Peng-hao Min and Guan-wen Ding implemented the experiments and analyzed the results. Jun-yuan Liu prepared the figures and tables. All authors wrote and revised the manuscript.

Funding This study was funded by Fundamental Research Funds for the Central Universities (2023ZKPYJD06) and National Natural Science Foundation of China (61901476).

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Firdaus M, Golchha H, Ekbal A, Bhattacharyya P (2021) A deep multi-task model for dialogue act classification, intent detection and slot filling. Cogn Comput 13(2):626–645
2. Colombo P, Chapuis E, Manica M, Vignon E, Varni G, Clavel C (2020) Guiding attention in sequence-to-sequence models for dialogue act prediction. In: Conitzer V, Sha F (eds) Proceedings of the AAAI conference on artificial intelligence, vol 34. AAAI Press, New York, NY, pp 7594–7601
3. Ahmadvand A, Choi JI, Agichtein E (2019) Contextual dialogue act classification for open-domain conversational agents. In: Carterette B, Kanoulas E, Sanderson M, Croft WB (eds) Proceedings of the 42nd international Acm Sigir Conference on Research and Development in Information Retrieval, vol 19. ACM, Paris, France, pp 1273–1276
4. Schuurmans J, Frasincar F (2019) Intent classification for dialogue utterances. IEEE Intell Syst 35(1):82–88
5. Qin L, Che W, Li Y, Ni M, Liu T (2020) Dcr-net: a deep co-interactive relation network for joint dialog act recognition and sentiment classification. In: Conitzer V, Sha F (eds) Proceedings of the AAAI conference on artificial intelligence, vol 34. AAAI Press, New York, NY, pp 8665–8672
6. Qin L, Li Z, Che W, Ni M, Liu T (2021) Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In: Yang Q, Leyton-Brown K (eds) Proceedings of the AAAI conference on artificial intelligence, vol 35. AAAI Press, New York, NY, pp 3210–3218
7. Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E (2016) Hierarchical attention networks for document classification. In: Walker MA, Ji H, Stent A (eds) Proceedings of the 2016 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1. Association for Computational Linguistics, San Diego, CA, pp 1480–1489
8. Tetreault VRJ (2019) Dialogue act classification with context-aware self-attention. In: Burstein J, Doran C, Solorio T (eds) Proceedings of NAACL-HLT, vol 1. Association for Computational Linguistics, Minneapolis, Minnesota, pp 3727–3733
9. Li C, Li L, Qi J (2018) A self-attentive model with gate mechanism for spoken language understanding. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J (eds) Proceedings of the 2018 conference on empirical methods in Natural Language Processing, vol 1. Association for Computational Linguistics, Brussels, Belgium, pp 3824–3833
10. Duran N, Battle S, Smith J (2023) Sentence encoding for dialogue act classification. Nat Lang Eng 29(3):794–823
11. Kumar H, Agarwal A, Dasgupta R, Joshi S (2018) Dialogue act sequence labeling using hierarchical encoder with crf. In: McIlraith SA, Weinberger KQ (eds) Proceedings of the Aaai conference on artificial intelligence, vol 32. AAAI Press, New Orleans, Louisiana, pp 3440–3447
12. He Z, Tavabi L, Lerman K, Soleymani M (2021) Speaker turn modeling for dialogue act classification. In: Walker MA, Ji H, Stent A (eds) Findings of the Association for Computational Linguistics: EMNLP 2021, vol 1. Punta Cana, Dominican Republic, Association for Computational Linguistics, pp 2150–2157
13. Yan M, Lou X, Chan CA, Wang Y, Jiang W (2023) A semantic and emotion-based dual latent variable generation model for a dialogue system. CAAI transactions on intelligence technology 8(1):1–15
14. Kumar A, Di Eugenio B, Aurisano J, Johnson A (2020) Augmenting small data to classify contextualized dialogue acts for exploratory visualization. In: Calzolari N, Béchet F (eds) Proceedings of the Twelfth Language Resources and Evaluation Conference, vol 1. European Language Resources Association, Marseille, France, pp 590–599
15. Guo M, Zhang Y, Yang Q, Shen G (2022) An intelligent multi-turn dialogue classification method based on resampling. J Phys: Conf Ser 2218(1):12–34
16. Kang Z, Liu Y, Bi G, Fang F, Yin P (2021) No news is an island: Joint heterogeneous graph network for news classification. 2021 IEEE Symposium on Computers and Communications (ISCC) 2218(1):1–7
17. Niu H, Xiong Y, Gao J, Miao Z, Wang X, Ren H, Zhang Y, Zhu Y (2022) Composition-based heterogeneous graph multi-channel attention network for multi-aspect multi-sentiment classification. In: Burstein J, Doran C, Solorio T (eds) International Conference on Computational Linguistics, vol 1. Gyeongju, Republic of Korea, International Committee on Computational Linguistics, pp 6827–6836
18. Yuan P, Jiang L, Liu J, Zhou D, Li P, Gao Y (2021) Dual-level attention based on a heterogeneous graph convolution network for aspect-based sentiment classification. Wirel Commun Mob Comput 2021(6):1–13
19. Yang S, Liu Y, Zhang Y, Zhu J (2022) A word-concept heterogeneous graph convolutional network for short text classification. Neural Process Lett 55(2):735–750
20. Kong D, Ji Z, Sang Y, Dong W, Yang Y (2022) Ua-hgat: Uncertainty-aware heterogeneous graph attention network for short text classification. 2022 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Com-

puting (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics) 2022(1):495–500
21. Zhang B, He Q, Zhang D (2022) Heterogeneous graph neural network for short text classification. Appl Sci 12(17):8–11
22. Shen W, Wu S, Yang Y, Quan X (2021) Directed acyclic graph network for conversational emotion recognition. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, vol 1. ACL, Bangkok, Thailand, pp 1551–1560
23. Pang J, Xu H, Song S, Zou B, He X (2022) Mfdg: A multi-factor dialogue graph model for dialogue intent classification. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, vol 13714. Springer, Cham, pp 691–706
24. Ghosal D, Majumder N (2019) DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In: Proceedings of the 2019 conference on empirical methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP, (eds) Inui K, Jiang J, Ng V, Wan X, vol 1. Association for Computational Linguistics, Hong Kong, China, pp 154–164
25. Devlin J, Chang M-W, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T (eds) Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, vol 1. Association for Computational Linguistics, Minneapolis, Minnesota, pp 4171–4186
26. Veličković P, Cucurull G, Casanova A, Romero A, Lió P, Bengio Y (2018) Graph attention networks. In: Murray I, Larochelle H (eds) International Conference on Learning Representations, vol 1050. OpenReview, Vancouver, Canada, pp 1–12
27. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst 30(1):5998–6008
28. Kingma DP, Ba J (2015) Adam: A method for stochastic optimization. In: Bengio Y, LeCun Y (eds) 3rd International Conference on Learning Representations, vol 14. OpenReview, San Diego, CA, USA, pp 1–15
29. Liu P, Qiu X, Huang X (2016) Recurrent neural network for text classification with multi-task learning. In: Kambhampati S (ed) Proceedings of IJCAI 2016, vol 5. AAAI Press, New York, pp 2873–2879
30. Zhou P, Shi W, Tian J, Qi Z, Li B, Hao H, Xu B (2016) Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of ACL 2016, vol 2. ACL, Berlin, pp 207–212
31. Kim Y (2014) Convolutional neural networks for sentence classification. In: Moschitti A, Pang B, Daelemans W (eds) Proceedings of EMNLP 2014, vol 1. ACL, Doha, pp 1746–1751
32. Wang J, Yu L-C, Lai KR, Zhang X (2016) Dimensional sentiment analysis using a regional CNN-LSTM model. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers, (ed) Ammar W, Erk K, Smith NA, vol 2. Association for Computational Linguistics, Berlin, Germany, pp 225–230
33. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019) Roberta: a robustly optimized bert pretraining approach. arXiv:1907.11692
34. Sun Y, Wang S, Li Y, Feng S, Tian H, Wu H, Wang H (2020) Ernie 2.0: A continual pre-training framework for language understanding. In: Conitzer V, Sha F (eds) Proceedings of the AAAI conference on artificial intelligence, vol 34. AAAI Press, New York, NY, pp 8968–8975
35. Wang K, Han SC, Poon J (2022) Induct-gcn: Inductive graph convolutional networks for text classification. In: Yang J, Zhou J, Sun Z (eds) 2022 26th International Conference on Pattern Recognition (ICPR), vol 1. IEEE, New York, USA, pp 1243–1249
36. Yao L, Mao C, Luo Y (2019) Graph convolutional networks for text classification. In: McIlraith S, Weinberger K (eds) Proceedings of the AAAI Conference on Artificial Intelligence, vol 33. AAAI, San Francisco, pp 7370–7377
37. You R, Zhang Z, Wang Z, Dai S, Mamitsuka H, Zhu S (2019) Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. Adv Neural Inf Process Syst 32(32):9699–9710