Enhanced BERT with Graph and Topic Information for Short Text Classification

Tong Zhang, Ailing Tang, Rong Yan*
College of Computer Science, Inner Mongolia University
Inner Mongolia Key Laboratory of Mongolian Information Processing Technology
National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian
Hohhot 010021, China
Email: csyanr@imu.edu.cn

***Abstract*—Short text classification is an important natural language processing task due to the prevalence of short text on the internet and social media platforms. In this paper, we propose a novel graph-based short text classification method named GBBM (Graph-BERT-BTM Model) that leverages the powerful representation ability of graph data to capture the structural features of short text. In this work, we incorporate topic information to enrich and expand the feature space for the short text and compare our proposed method on five publicly available short text datasets with five existing models. Experimental results indicate the superiority of our proposed method.**

***Index Terms*—short text classification, graph data, topic model**

I. Introduction

With the rapid development of the Internet, a large number of short texts are generated every second on social media platforms [1], and short text classification has become a pressing concern in the field of natural language processing (NLP). It is crucial for various applications due to short texts with valuable information, such as sentiment analysis [2], recommendation systems [3], and topic identification [4]. However, the restricted number of words in short texts limits their semantic and contextual information, rendering the short text classification a more formidable challenge than the long text counterpart. Recently, deep learning models have been widely used to address short text classification tasks, such as convolutional neural networks (CNN) [5] and recurrent neural networks (RNN) [6]. These deep learning models can capture semantic and syntactic information in local consecutive word sequences well. However they may ignore global word cooccurrence in a corpus which carries non-consecutive and long-distance semantics [7].

With the development of deep learning technology, Graph Neural Networks (GNNs) [8] have been the subject of extensive research and implementation. TextGCN [9] constructed a co-occurrence graph of words and documents, which is then transformed into a node classification task using text graph convolutional networks for text classification. However, this model heavily relies on convolutional feature learning and aggregation operations for graph representation learning. As the number of layers increases, it loses its capacity to learn features and fails to attain parallel training. Chen et al. [10] approximated the integral of the embedding function through Monte Carlo methods and implemented batch training to significantly enhance the efficiency of model training. However, their models do not fully incorporate the global structural features of short texts.

In this paper, we present a novel classification method named GBBM (Graph-BERT-BTM Model) that blends graph neural networks and topic models for feature fusion. Specifically, the method involves three key steps. Firstly, we construct a document-topic-word heterogeneous graph on the corpus, which harnesses the representation capacity of graph data to capture the structural characteristics of short texts and supplements them with topic information to expand the feature space. Next, we partition the graph data into subgraphs to enable parallel processing. We input the feature representation obtained from the batch graph data into BERT [11] and utilize an attention mechanism to learn the semantic features of the short text in order to reduce sparsity. Finally, we merge the highly correlated topic information obtained from the topic model to improve the accuracy and effectiveness of short text classification.

II. Methodology

In this section, we present the construction of a heterogeneous graph and subgraph sampling method based on the graph model. To address the limited content of short texts and the lack of semantic and contextual information, we propose to build a document-topic-word heterogeneous graph structure on the corpus, where documents, topics, and words are jointly learned to expand the features of short texts.

A. Building a document-topic-word heterogeneous graph

In the heterogeneous graph of short texts, there are mainly three types of nodes: documents, topics, and words. The relationship edge between document-word is ascertained using calculating the term frequency-inverse document frequency (TF-IDF) value of word nodes contained within the document, while the edge between words is fashioned based on Pointwise Mutual Information (PMI). Furthermore, this paper introduces a topic model to represent the relationship strength among

DOI reference number: 10.18293/SEKE2023-158

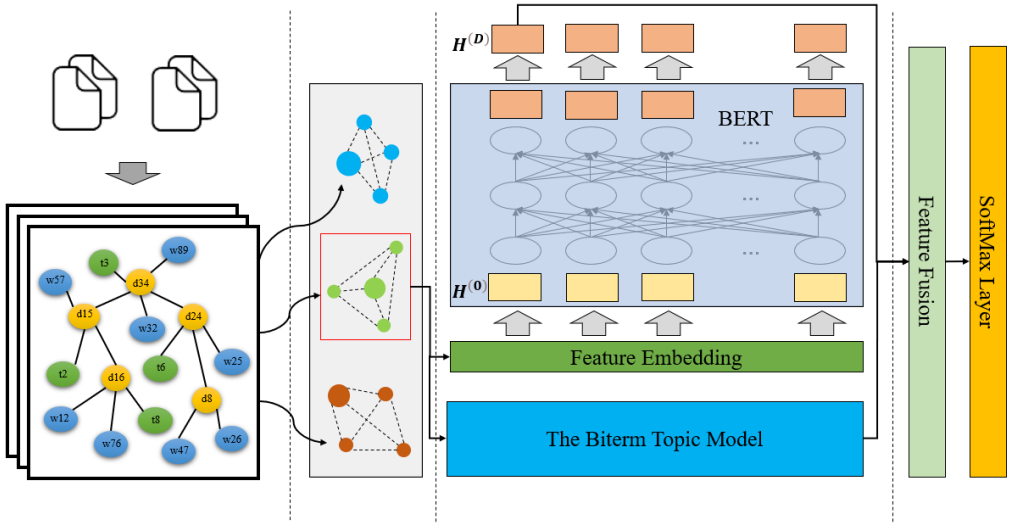


Fig. 1. The architecture of GBBM

documents, topics, and words. The formal definition of the weight value A_{ij} of the edge between i_{th} node and j_{th} node is shown below:

$$A_{ij}\begin{cases} PMI(i,j) & n_i, n_j \ are \ word \ nodes. \\ TF-IDF_{ij} & n_i \ is \ doc \ node, \ n_j \ is \ word \ node. \\ doc-topic_{ij} & n_i \ is \ doc \ node, \ n_j \ is \ topic \ node. \\ topic-word_{ij} & n_i \ is \ topic \ node, \ n_j \ is \ word \ node. \\ 1 & i=j. \\ 0 & other. \end{cases}$$

B. Subgraph Sampling Based on Text Graph

In this paper, we denote the input data as a graph G=(V,E), where V and E represent the node and edge sets in the graph, respectively. For each node v_i, we learn the features of itself and neighboring nodes within a specific range to form a subgraph set without edges. Inspired by Zhang et al. [12], we first calculate the intimacy matrix $\mathbf{S}(i,j)$ between node pair (v_i,v_j) using PageRank [13]. Then, we adopt the top-k intimacy sorting method to select the k-nearest nodes to v_i as its contextual information, constructing a subgraph g_i with k+1 nodes. By repeating this process for all nodes, the complete graph can be represented as G=(g_1, g_2,$\cdots$, g_V). The intimacy matrix $\mathbf{S}$ based on PageRank is defined as Eq.(1):

$$\mathbf{S} = \alpha \cdot (\mathbf{I} - (1-\alpha)\cdot \bar{\mathbf{A}})^{-1} \tag{1}$$

where the factor $\alpha \in [0,1]$, and it is usually set as 0.15. In this work, we introduce the adjacency matrix $\mathbf{A}$ and its corresponding diagonal matrix $\mathbf{D}$ of graph G, denoted as $\mathbf{D}(i,j)$=$\sum_j \mathbf{A}(i,j)$, $\bar{\mathbf{A}}$=$\mathbf{AD}^{-1}$ represents the column-normalized adjacency matrix. We employ the top-k intimacy sampling method to select the closest neighboring nodes around node v_i, forming the subgraph g_i centered on v_i. Finally, the input graph data is projected into a feature vector representation through network embedding to obtain the original feature representation of the node.

C. Model architecture

GBBM achieves joint training of BERT and BTM on the text graph as shown in fig. 1. GBBM contains four parts: (1) a document-topic-word heterogeneous graph, (2) subgraph sampling, (3) feature fusion based on BERT and BTM, and (4) classifier. GBBM learns text graph features using only the attention mechanism in BERT, without any graph convolution or aggregation operations. Firstly, the original input text sequence is transformed into graph-structured data, allowing for the full exploration of complex semantic information within short texts. Additionally, topic modeling is utilized to extract topic information from short texts, which is then used to assist in constructing a text graph consisting of document, topic, and word nodes. PageRank is then employed to learn node features surrounding specific targets in the graph data, which are subsequently partitioned into small batches of subgraphs. For node v_i, its original feature x_i is represented in one-hot form, and a fully connected layer is used to transform the original feature of v_i into an embedding representation with consistent dimensions d_h. The initial input vectors of all nodes in subgraph g_i can be represented by Eq.(2):

$$\mathbf{H}^{(0)} = [h_i^{(0)}, h_{i,1}^{(0)}, \ldots, h_{i,k}^{(0)}]^\top \in \mathbb{R}^{(k+1)\times d_h} \tag{2}$$

with using multiple layers to iteratively update the representation of nodes, the output of layer l can be expressed as Eq.(3):

$$\mathbf{H}^{(l)} = Transformer(\mathbf{H}^{(l-1)}) \tag{3}$$

where

$$\begin{cases} \mathbf{Q} = \mathbf{H}^{(l-1)}\mathbf{W}_Q^{(l)} \\ \mathbf{K} = \mathbf{H}^{(l-1)}\mathbf{W}_K^{(l)} \\ \mathbf{V} = \mathbf{H}^{(l-1)}\mathbf{W}_V^{(l)} \end{cases} \tag{4}$$

In the above equations, $\mathbf{W}_Q^{(l)}$, $\mathbf{W}_K^{(l)}$, $\mathbf{W}_K^{(l)} \in \mathbb{R}^{d_h \times d_h}$ denote the involved varibales, and $\mathbf{X}_i \in \mathbb{R}^{(k+1)\times d_x}$ is the raw features of all nodes in the subgraph g_i.

All initial input vectors of nodes in the short text subgraph are fed into BERT, where the attention mechanism automatically learns short text features and acquires hidden semantic representations. Additionally, the preprocessed short text sequence is used for training BTM. The document topic distribution (DT) and topic word distribution (TW) of the short text can be represented through Eq.(5) and Eq.(6), respectively.

$$DT = \varphi_{b|z} = \frac{n_{b|z} + \beta}{\sum_w n_{b|z} + M\beta} \tag{5}$$

$$TW = \phi_z = \frac{n_z + \alpha}{|B| + K\alpha} \tag{6}$$

where $n_{b|z}$ represents the number of occurrences of co-occurrence word pair b in topic z, K is the number of topics, $|B|$ is the total number of biterms, and M is the number of non-repeating words in the entire corpus [14]. The parameters α and β are the prior parameters of the topic model.

Finally, the semantic representation generated by BERT and the potential topic information obtained from BTM are integrated to form the final representation vector. A Softmax classifier is then used for training, and the category probability distribution of the fused document-level topic information is calculated through Eq.(7). It allows us to obtain the corresponding category for the input short text.

$$P(L_i|H_i, DT_i) = \frac{\exp(H_i, DT_i)}{\sum_{k=1}^{K} \exp(H_k, DT_k)} \tag{7}$$

III. Experiments

In this section, we evaluate the accuracy, precision, recall, and $F1$ score of each model on the same set of short text datasets to ensure a fair comparison.

A. Datasets

In this paper, we use five datasets to compare and verify our proposed method.

- **Pascal Flickr**[1]**:** It is a collection of 4,834 captions primarily used for evaluating short text clustering or classification tasks.
- **GoogleNews**[2]**:** It is a collection of 11,109 news article titles and snippets.
- **Ohsumed**[3]**:** It is a dataset in the field of topic classification, and contains 7,400 medical abstracts with 23 different labels.
- **TREC**[4]**:** It is one of the mainstream datasets in the question-answering task. This article uses the TREC-6 version dataset, which includes 5,452 training and 500 testing questions.
- **SST**[5]**:** The Stanford Sentiment Treebank (SST) dataset is an extended version of the sentiment analysis dataset MR. In this paper, we select SST-2 that contains 9,613 documents.

In accordance with standard practice, the datasets was split into training set, verification set and test set according to the ratio of 7:1.5:1.5.

[1]https://github.com/qiang2100/STTM.
[2]https://github.com/qiang2100/STTM.
[3]http://davis.wpi.edu/xmdv/datasets/ohsumed.html.
[4]https://cogcomp.seas.upenn.edu/Data/QA/QC/.
[5]https://dl.fbaipublicfiles.com/glue/data/SST-2.zip.

TABLE I
COMPARISON RESULTS OF SHORT TEXT DATASETS (%)

Model	Pascal Flickr	GoogleNews	TREC	SST	Ohsumed
TextCNN	48.14	82.36	90.02	81.13	58.38
TextRNN	46.34	80.47	87.33	80.79	50.27
FastText	45.79	81.46	88.90	82.27	52.34
TextGCN	52.17	84.48	90.06	81.66	64.56
BERT-raw	55.10	87.82	95.33	89.67	68.55
GBBM+DT	**59.09**	**92.64**	97.00	**93.67**	**71.61**
GBBM+TW	58.95	90.67	**97.33**	91.35	70.22

B. Experimental Settings

In this work, the configuration of topic model employs the parameter settings from the original paper, specifically α=50/k, β=0.01. GBBM features a hidden layer size of 32, 2 hidden layers, 2 multi-head attention mechanisms, hidden layer dropout of 0.5, attention dropout of 0.3, and utilizes GULE function as its activation function. We assume that the node residual term is independent and determined solely by the original input feature $R(H(k-1), X, G)$=X. An early-stop strategy is used during model training to avoid overfitting. We use NLTK[6] to preprocess the short text data.

C. Results and Analysis

We compare our proposed method with five different models that have been used in recent years across five different classification tasks as shown in Table I. It is evident from the table that irrespective of the type of topic information fused, the proposed GBBM outperforms the other comparative models in terms of classification on all five datasets. Experiments show that the utilization of topic information from short texts not only serves as additional features to alleviate the issue of data sparsity during graph construction but also aids the model in quickly capturing crucial information from texts during the training phase.

TABLE II
COMPARISON RESULTS ON TREC

	BERT			GBBM			
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)	Support
LOC	100.00	100.00	100.00	100.00	100.00	100.00	196
HUM	84.54	91.15	87.72	**92.54**	**96.88**	**94.66**	192
NUM	97.90	93.96	95.89	**99.32**	**98.66**	**98.99**	149
ABBR	100.00	97.12	98.54	100.00	**99.52**	**99.76**	208
ENTY	**85.38**	84.73	85.06	85.24	**91.60**	**93.39**	131
DESC	100.00	83.23	90.11	100.00	**87.50**	**93.33**	16

To further demonstrate the efficient text classification performance of GBBM, we present a comparison of classification

[6]https://github.com/nltk.

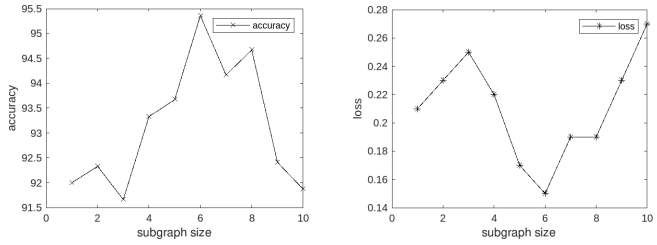
measures for different categories of the TREC dataset in Table II. The experiment shows that the short text graph established based on document-topic-word nodes enables GBBM to have more available contextual information. Even with only 16 texts in the DESC category of the TREC dataset, GBBM can still efficiently complete short text classification tasks.

Table III presents the experimental results of GBBM incorporating multiple perspectives of topic representations, where GBBM+DW indicates GBBM that integrates document-word topic information and GBBM+DT+TW represents GBBM that jointly learns document-topic-word features. As shown in Table III, it can be observed that GBBM+DT+TW and GBBM+DW perform better on most datasets, indicating that the joint learning of document-topic-word features can effectively capture the topic information, thereby improving the classification performance.

TABLE III
CLASSIFICATION PERFORMANCE OF GBBM WITH MULTI-ANGLE TOPIC REPRESENTATIONS

	PascalFlickr	GoogleNews	TREC	SST	Ohsumed
BERT	55.10	87.82	95.33	89.67	68.55
GBBM+DW	57.13	**91.12**	**97.18**	**91.52**	70.36
GBBM+DT+TW	**57.37**	90.47	95.89	91.38	**71.48**

Furthermore, we analyze the learning performance of GBBM with different subgraph sizes (parameter s) on the TREC dataset as shown in fig. 2. The learning performance of GBBM reaches its best when s increases from 1 to 6. However, with a further increase in s, the performance sharply declines. Similar results are observed for other datasets, albeit with different optimal values of s.



(a) accuracy curve (b) loss curve

Fig. 2. The performance with different subgraph sizes

IV. CONCLUSION

In this paper, we propose a novel approach to short text classification that integrates graph neural networks and topic models for feature fusion. We conducted experiments on five public real-world datasets and compared the performance of our method with the current state-of-the-art text classification models. The results showed that our model outperformed the other five baselines, demonstrating the effectiveness of our approach. In future research, we plan to enhance the filtering of weakly related node information based on topic features, while retaining deep semantic information, in order to mitigate the issue of limited memory due to a large number of nodes and improve the accuracy of short text classification.

ACKNOWLEDGMENT

This research is supported by the National Natural Science Foundation of China (Grant No. 61866029).

REFERENCES

[1] N. T. Issa, S. W. Byers, and S. Dakshanamurthy, "Big data: the next frontier for innovation in therapeutics and healthcare," *Expert review of clinical pharmacology*, vol. 7, no. 3, pp. 293–298, 2014.

[2] H. Chen, M. Sun, C. Tu, Y. Lin, and Z. Liu, "Neural sentiment classification with user and product attention," in *Proceedings of the 2016 conference on empirical methods in natural language processing*, 2016, pp. 1650–1659.

[3] W. Zhang, D. Wang, G.-R. Xue, and H. Zha, "Advertising keywords recommendation for short-text web pages using wikipedia," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 2, pp. 1–25, 2012.

[4] J. Zeng, J. Li, Y. Song, C. Gao, M. R. Lyu, and I. King, "Topic memory networks for short text classification," *arXiv preprint arXiv:1809.03664*, 2018.

[5] Y. Chen, "Convolutional neural network for sentence classification," Master's thesis, University of Waterloo, 2015.

[6] X. Zhu, P. Sobihani, and H. Guo, "Long short-term memory over recursive structures," in *International conference on machine learning*. PMLR, 2015, pp. 1604–1612.

[7] P. Hao, J. Li, H. Yu, Y. Liu, and Y. Qiang, "Large-scale hierarchical text classification with recursively regularized deep graph-cnn," 2018.

[8] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner *et al.*, "Relational inductive biases, deep learning, and graph networks," *arXiv preprint arXiv:1806.01261*, 2018.

[9] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, no. 01, 2019, pp. 7370–7377.

[10] J. Chen, T. Ma, and C. Xiao, "Fastgcn: fast learning with graph convolutional networks via importance sampling," *arXiv preprint arXiv:1801.10247*, 2018.

[11] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of naacL-HLT*, 2019, pp. 4171–4186.

[12] J. Zhang, H. Zhang, C. Xia, and L. Sun, "Graph-bert: Only attention is needed for learning graph representations," *arXiv preprint arXiv:2001.05140*, 2020.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." Stanford InfoLab, Tech. Rep., 1999.

[14] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 1445–1456.