Enhancing BERT for Short Text Classification with Latent Information

Ailing Tang[1,2,3], Yufan Hu[1,2,3], and Rong Yan[1,2,3](✉)

[1] College of Computer Science, Inner Mongolia University, Hohhot 010021, China
csyanr@imu.edu.cn

[2] Inner Mongolia Key Laboratory of Mongolian, Information Processing Technology, Hohhot 010021, China

[3] National & Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, Hohhot 010021, China

Abstract. With explosive growth of short text, short text categorization has been attracted increasing attention. How to alleviate the sparsity of short texts is a research hotspot, and takes a enormous challenge for classical text categorization technique. In this paper, we focus on short text expansion based on multi-granularity and explore to construct an EBLI (Enhancing BERT with Latent Information) model by combining BERT and latent information for addressing short text classification task. Additionally, we establish a memory bank to store the whole document topic information that assists in the joint training of deep semantic features and topic features. Experimental results with five widely datasets show that our proposed model achieves better performance of short text classification as well as promote the generalization ability and strong competition ability for the classifier.

Keywords: Short Text Classification · BERT Model · Topic Model · Text Expansion · Memory Bank

1 Introduction

Text classification is one of the core tasks in natural language processing (NLP) and has been used in many real-world applications such as opinion mining [1], sentiment analysis [4], and news classification [21]. Different from the standard text classification, short text classification has to face with a series of difficulties and problems, such as sparsity, shortness, lack of contextual information and semantic inadequacy, which brings an enormous challenge for traditional text classification methods. In recent years, many scholars have put forward some ingenious strategies to solve these problems in short texts. Some way try to solve the sparsity and shortness of short texts by using internal, external resources and deep learning methods to expand text [3,10,11]. Li et al. [10] identified the concept of the text based on Wikipedia as a knowledge base and attempted to add the corresponding information into short text. Nevertheless, these methods strongly depend on the quality of external resources, and these

M. Tanveer et al. (Eds.): ICONIP 2022, LNCS 13625, pp. 122–132, 2023.
https://doi.org/10.1007/978-3-031-30111-7_11

resources are really scarce in fact. Another way explores to construct the classification model for short text. Recently, combining pre-training models [6,9,14,19] are popular for addressing short text classification task. However, these models are lack of generalization ability. Then, Peinelt et al. [16] proposed a novel topic-based architecture to enhance performance of pre-training model.

Inspired by their work, in this paper, we focus on enhancing the classification effect of single label short text by improving text expansion techniques. Firstly, we explore to carry out short text expansion based on multi-granularity to make up for the sparsity of short texts. Based on this, we propose a significantly more lightweight model named EBLI (Enhancing BERT with Latent Information). It can increase the interpretability of the model by latent topic information and improve the efficiency of semantic extraction for short texts. Furthermore, we adopt a memory bank mechanism to achieve the joint training of these features. In summary, EBLI model is able to leverage the advantages of both pre-training model and topic model, not only it can extract the dependencies between words from the input word sequence, but also capture the global semantic information of multiple documents by using the underlying topic. Experimental results show that our framework based on expanded text outperforms the state-of-the-art baselines on five public datasets.

2 Related Work

Our work is most relevant to two bodies of research efforts: text expansion and pre-training model. In this section, we describe a brief related work about these two respects for the specific scenario of short text classification task.

Text expansion has always been fundamental research in short text classification tasks. It is common that using topic model [15] to extract the additional information from short text as expanded words. For instance, Gao et al. [7] proposed a regularized model based on conditional random fields and expanded the content by extracting appropriate words from topic model. These methods can effectively alleviate the impact of sparsity and shortness on the classification effect. However, they ignored the correlation between the words, which makes the topic model unable to clearly express the semantic information of short text. In addition, the emerging keywords extraction techniques [2,17] are also favored, which rests on the most relevant words from single documents based on ‘word’ granularity. Sharma et al. [17] proposed a novel self-supervised approach using contextual and semantic features to extract the keywords. However, they had to face an awkward situation of these information merely reflected the semantic information from ‘word’ granularity, and unable to consider multi-granularity information.

Besides above text expansion techniques, some researches tried to improve pre-training models [9,14] for short text classification, which are typically trained on large-scale corpora unrelated to a specific NLP task. And they are convenient to fine-tune for specific NLP tasks. Compared to other known pre-training models, BERT model [6] captured deep semantic representation and achieved prominent performance on lots of NLP tasks [5,16,18]. Nevertheless, sparsity, shortness, lack of contextual information and semantic inadequacy of short texts are still challenges for BERT model, which will limit the ability of this model during classifying procedure. In the work of [13], it indicated that integrating structured information in the knowledge base into the pre-training model would improving the representation ability. However, their models still needed to mine related information from external knowledge and they did not take ambiguity of the noise of external information. Therefore, how to extract more reliable knowledge and explore appropriate combination method for short text classification is still a dominant problem.

3 Methodology

In this section, we introduce the overall architecture based on text expansion and classification model in detail. Obviously, the improvement of text extension focuses on the construction of extended word set based on multi-granularity resulting in short text found strong related words. Despite of this, single text expanding is not always useful for short text expressions of the flexibility and diversity. Thus, EBLI model is absorbed to accommodate more complex short texts.

3.1 Text Expansion Scheme

The goal of this scheme is to expand strong relevant words from the internal resources. Taking the Biomedical dataset as an example, Fig. 1 shows the text expansion method based on keywords and topic model.

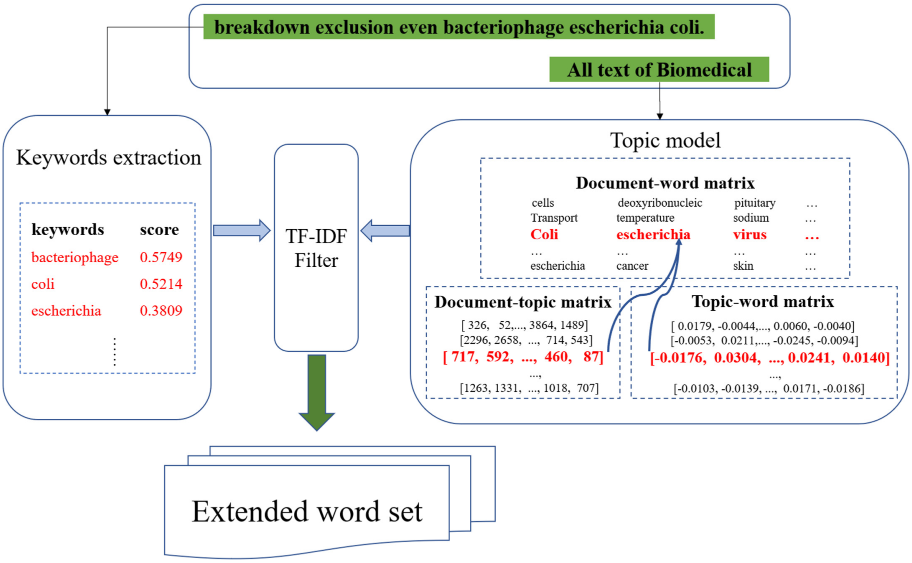

Fig. 1. Text expansion based on multi-granularity information.

Short text is generally of shortness and sparsity which makes it difficult to model. Accordingly, we use the topic model [15] and keyBERT model [17] to retrieve multi-granularity information and obtain an extended candidate word set Y. According to the statistics, the part of speech (POS) of the feature words are generally nouns, verbs and adjectives in short texts. Therefore, in order to ensure the quality of the extended words, we calculate the Term Frequency Inverse Documentation Frequency [8] (TF-IDF) values of Y. And then the minimum value of TF-IDF is regarded as the threshold according to the POS. To be specific, if the calculated TF-IDF value is no less than the threshold, the sample word will be appended to expanded word set Y. Otherwise, the sample word will delete. With high-quality extended word set, we can quickly classify data into a certain category as shown in Algorithm 1. The text expansion technique can be divided into three steps, including construct candidate word set, calculate threshold and determine expanded word set in Algorithm 1.

Algorithm 1. Text expansion

```
Input: $D$={$d_1$,$\cdots$,$d_N$}
Output: extended word set $Y$
#(Step 1)construct candidate word set
for $i = 1$ to $N$ do
    $EK$=$keyBERT(d_i)$
    Append $EK$ to $ekList[i]$
$etList$=$LDA(D)$
$Y$=($etList$; $ekList$)
#(Step 2) calculate threshold
Compute $w$ in ($Y_1$,$\cdots$,$Y_N$)$\in Y$ $TF$-$IDF$ value
$threshold$=$min(w.tfidf)$ while $w.pos$ in ($n$,$v$,$adj$)
#(Step 3) determine expanded word set
for $w$ in ($Y_1$,$\cdots$,$Y_N$)$\in Y$ do
    if $w.tfidf$¡$threshold$ then
        Remove $w$ from $Y$
```

3.2 EBLI Model

Figure 2 shows the architecture of our proposed EBLI model based on expanded text. In our model, we encode two aspects original text $(X_1,\cdots,X_n)$ and the extended word set $(Y_1,\cdots,Y_m)$ with BERT model. Afterwards, the expanded set is mapped into input embedding layer, which outputs by adding token, segment and position embedding. Then feeding input embedding into bi-directional transformer. We use the final hidden state $[CLS]$ token from Eq.(1-2) as the semantic representation of BERT model. Indeed, $C_x \in R^{b\times k}$ and $C_y \in R^{b\times k}$ vectors are from original text and extended word set respectively, which are corresponding to the CLS after a linear classification layer, where b and k denote the batch-size and category of BERT model respectively.

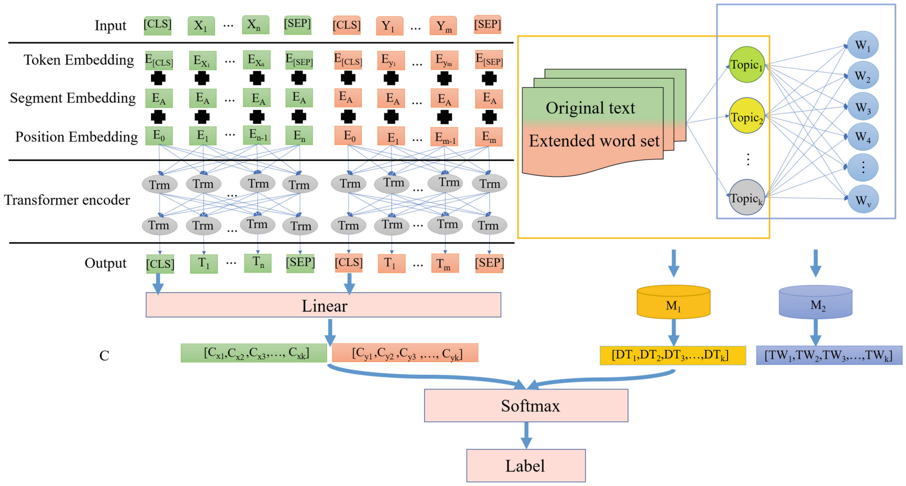

Fig. 2. Architecture of EBLI with DT.

$$G = LayerNorm(T^{i-1} + MultiAttention(T^{i-1})) \tag{1}$$

$$T^i = LayerNorm(G + FeedForward(G)) \tag{2}$$

where, T^i denotes the output of the i-th layer, it can learn and store the semantic relationships and syntactic structure information of document d_i. LayerNorm is layer normalization, MultiAttention is a multiple attention mechanism and FeedForward is a feedforward network with RELU. Notably, it is assumed that the dataset $D = \{d_1, \cdots, d_N\}$ contains N documents, we represent each document as $d_i = [DT, TW]$, which are given by the following Eq. (3). The document-topic distribution $DT \in R^{n\times k}$ represents the overall topic of the document, while the topic-word distribution $TW \in R^{k\times v}$ of word w_i essentially captures how topical the word is in itself. Among them, α and β are the hyperparameters of topic model, $n_i^{(k)}$ and $n_k^{(w)}$ represent the number of times that d_i is sampled as topic k and word w is sampled as topic k.

$$\boldsymbol{DT}_{\mathrm{i,k}} = \frac{\mathrm{n}_{\mathrm{i}}^{(\mathrm{k})} + \alpha}{\sum_{\mathrm{k}=1}^{\mathrm{k}} \left(\mathrm{n}_{\mathrm{i}}^{(\mathrm{k})} + \alpha\right)}, \boldsymbol{TW}_{\mathrm{k,w}} = \frac{\mathrm{n}_{\mathrm{k}}^{(\mathrm{w})} + \beta}{\sum_{\mathrm{v}=1}^{\mathbf{v}} \left(\mathrm{n}_{\mathrm{k}}^{(\mathrm{w})} + \beta\right)} \tag{3}$$

For the sake of subsequent calculations, we compress the TW distribution to a fixed-dimension feature representation. However, the topic model uses the full-batch datasets for training, which is intractable for BERT model due to the memory limitation. Inspired by techniques in [12] which decouples training batch size from the total number of nodes in the graph, we maintain two memory banks M_1 and M_2 that track document-topic features and topic-word features for all documents. During each iteration, we sample a mini-batch b from two memory banks. Correspondingly, the M_1 and M_2 banks are injected into BERT model.

As shown in Eq. 4, we adopt full connection layer and use softmax as activation function to calculate a category probability distribution $P_{DT} \in R^{b\times k}$. Similarly, we can get $P_{TW} \in R^{b\times k}$ according to this function. In this way, BERT model stores the semantic relationships among words and we integrate deep semantic features with topic features resulting in words found in close proximity to one another semantically disambiguated.

$$p\left(L_k \mid C_{Xi}, C_{Yi}, DT_i\right) = \frac{\exp\left(C_{Xi} + C_{Yi} + DT_i\right)}{\sum_{k=1}^{k} \exp\left(C_{Xk} + C_{Yk} + DT_k\right)} \tag{4}$$

4 Experiments

Five popular short text classification datasets (Biomedical[1], Movie Review (MR)[2], Tweet[3], SearchSnippets[4] and StackOverflow[5]) with multiple domains are applied here to run the experiments. To evaluate the robustness and effectiveness of our proposed EBLI model, we compare EBLI model with five baselines.

4.1 Dataset

In the experiment, we randomly divide each dataset into train set, test set and the validation set according to the proportion of 7:1.5:1.5. After preprocessing, we summarize the detailed information of each dataset in Table 1, where category_num denotes the category number of each dataset, and len_avg and len_max denote the average length and maximum length of each document respectively. Moreover, vocabulary and doc_num denote the size of the vocabulary and the number of documents in each dataset respectively.

Table 1. Statistic information of the datasets.

Dataset	category_num	len_avg	len_max	vocabulary	doc_num
Biomedical	20	7.44	28	4,498	19,448
MR	2	11.17	27	4,081	10,662
SearchSnippets	8	14.40	37	5,547	12,295
Tweet	3	17.15	26	5,574	20,735
StackOverflow	20	5.03	17	2,638	16,407

[1] https://github.com/rashadulrakib/short-text-clustering-enhancement/tree/master/data/biomedical/.

[2] http://disi.unitn.it/moschitti/corpora.html.

[3] https://github.com/haroonshakeel/multisenti/.

[4] http://jwebpro.sourceforge.net/data-web-snippets.tar.gz.

[5] https://github.com/jacoxu/StackOverflow/.

4.2 Baselines

For comparison, five baselines are used to illustrate the effectiveness of EBLI model, including BERT[6], ERNIE[7], Roberta[8], Albert[9] and SHINE[10] with heterogeneous information networks in our experiment [20]. For EBLI model, we not only use BERT model to represent over the text documents, but also yield more interpretable results due to the involvement of topic modeling.

4.3 Experiment Settings

In the experiment, we set hyperparameters $\alpha = 50/k$, $\beta = 0.01$, and the number of topics k is set according to the specific dataset. In particular, the k is also equal to the number of categories in short texts. During encoding, we use a batch size of 16, the maximum sequence length is set to 128 for all datasets. For EBLI model, the training epochs are set to 3. We set the learning rate = 5e−5 when updating BERT model. It is worth mentioning that the hidden size of Albert model is set to 312 and ERNIE model with a learning rate of 2e−5. We train our model for a dropout of 0.1 and optimize cross entropy loss using Adam[11] optimizer. Once the performance of model has not improved after more than 1,000 batches, we will stop train model early.

4.4 Results and Discussion

Table 2 presents the comparisons of our model with baselines, where line 6 (EBLI+TW) and line 7 (EBLI+DT) denote the different combination latent information in EBLI model from the word level and the collection level respectively.

Table 2. Results of different models on all datasets.

	Biomedical	MR	SearchSnippets	Tweet	StackOverflow
ERNIE	60.62	71.11	90.95	61.76	81.91
Albert	63.08	69.11	88.29	63.66	79.27
Roberta	61.71	74.73	94.15	64.29	83.21
SHINE	65.25	77.50	94.37	63.27	81.00
BERT	62.62	72.61	92.14	64.91	81.54
EBLI+TW	**79.02**	79.55	93.98	67.28	83.90
EBLI+DT	78.09	**83.68**	**95.61**	**68.00**	**84.11**

[6] https://huggingface.co/bert-base-uncased/.
[7] https://huggingface.co/nghuyong/ernie-2.0-en/.
[8] https://huggingface.co/roberta-base/.
[9] https://huggingface.co/albert-base-v1/.
[10] https://github.com/tata1661/shine-emnlp21/.
[11] https://arxiv.org/pdf/1412.6980.pdf.

As shown in Table 2, it shows that our proposed EBLI model is remarkable and stable no matter how the domain information changed. EBLI model can acquire more robust performance than other models by capturing interpretable information both on word and collection point of view. To be specific, the accuracy of EBLI model reaches 78.09 in EBLI+DT from the document level and 79.02 in EBLI+TW from the word level, but BERT model reaches 62.62 in Biomedical. We analysis that it is due to the text characters of Biomedical with lots of biomedical terms. However, for BERT model, the performance of classification is heavily depend on the integrating degree of data to the model. Apparently, it is difficult for BERT model to learn specifical words, but topic model serves as a simple and efficient way, which can efficient to extract strong interpretable information and alleviate this awkwardness. Simultaneously, the same scene is happened on MR. Besides, we find that a slight improvement for StackOverflow. Its average sequences length ranges from a few to more than 10 and its vocabulary is not enough to capture rich topic information even after expanding. The result validates the effectiveness and importance of the proposed joint optimization framework in leveraging the strengths of both BERT and topic model to complement each other.

Furthermore, in order to measure the impact of text expansion methods on the performance of EBLI model, we exhibit *Precision* (P), *Recall* (R), and $F1$ metrics comparisons on MR in Table 3.

Table 3. Results of EBLI on MR with/without expanding.

	Original text			Expanded text		
	P(%)	R(%)	$F1$(%)	P(%)	R(%)	$F1$(%)
Positive	73.71	62.47	67.63	**74.01**	**72.40**	**73.19**
Negative	65.52	**76.20**	70.45	**71.18**	72.83	**71.99**

From Table 3, it is obvious find that EBLI model achieve better on MR when using extension scheme. It further illuminates that our extension scheme improves the amount of information contained in the short text itself and text expansion based on multi-granularity provides rich extension features, which makes our models with joint training have more available contextual information, and further extract the deep semantic information of short texts.

Figure 3 further shows the comparisons on all datasets. As shown in Fig. 3, it further illustrates that the effectiveness of our expending strategy for the extended text has longer length and stronger semantic correlation.

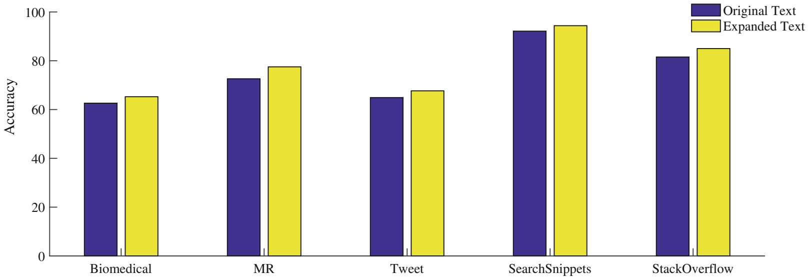

Fig. 3. Comparisons on all datasets with and without expansion.

Next, we exhibit the classification performance of EBLI model with different representations of topic information in Table 4. Compared with traditional BERT model, it is obvious that the EBLI model combined different topic representations (DW and DT+TW) can significantly reduces the lack of contextual information and semantic inadequacy. Especially for Biomedical, the EBLI model achieves the best effect in EBLI+DT+TW. We analysis that it is because topic representation from the multi-angle can better integrating the background constraints. Meanwhile, it is further verified that launching a smaller and simpler model to BERT model will capture richer information for short text classification.

Table 4. Classification performance of EBLI model with different representations.

	Biomedical	MR	SearchSnippets	Tweet	StackOverflow
BERT	62.62	72.61	92.14	64.91	81.54
EBLI+DW	78.54	81.99	**94.36**	**67.64**	**83.90**
EBLI+DT+TW	**79.84**	**82.55**	92.36	66.10	83.66

Furthermore, we attempt to add different topic representations for our model, and obtain $F1$ score under 20 categories of Biomedical as shown in Fig. 4, due to the large number of categories, we use the number to represent the categories to show the experimental results. Results from experiments demonstrate that $F1$ score of BERT model is worse than EBLI model on different categories. The reason is that the 20 categories of Biomedical have a strong correlation, and EBLI model can efficiently boost their representative ability using lexicon information. On the whole, based on the comparison results of the above experiments, the effectiveness of EBLI model with latent information in solving sparsity, shortness,lack of semantic and contextual information has also been fully illustrated.

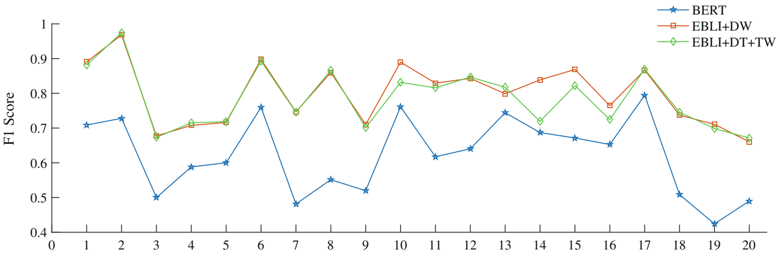

Fig. 4. The $F1$ score of models with different topic represents on Biomedical.

5 Conclusion and Future Work

In this paper, we propose a flexible framework named EBLI for enhancing BERT with latent information for addressing short text classification task. What makes it reasonable is that semantic knowledge and topic knowledge can provide enough diversity for short texts. Experimental results indicate that the effectiveness and generalization of our proposed approach. In future work, we try to explore how to directly reduce the loss of hierarchical semantics and preserve rich and complex semantic information of the document in training procedure in order to better apply it to the multi-classification task for short text.

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

References

1. Bakshi, R.K., Kaur, N., Kaur, R., Kaur, G.: Opinion mining and sentiment analysis. In: Proceedings the 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 452–455. IEEE (2016)
2. Campos, R., Mangaravite, V., Pasquali, A., Nunes, C., Jatowt, A.: YAKE! keyword extraction from single documents using multiple local features. Inf. Sci. **509**, 257–289 (2020)
3. Chen, C., Ren, J.: An improved PLDA model for short text. In: Frasincar, F., Ittoo, A., Nguyen, L.M., Métais, E. (eds.) NLDB 2017. LNCS, vol. 10260, pp. 58–70. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-59569-6_7
4. Chen, J., Hu, Y., Liu, J., Xiao, Y., Jiang, H.: Deep short text classification with knowledge powered attention. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6252–6259. AAAI, Honolulu, Hawaii, USA (2019)
5. Chi, S., Huang, L., Qiu, X.: Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 380–385. Minneapolis, MN, USA (2019)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HL), pp. 4171–4186 (2019)

7. Gao, W., Peng, M., Wang, H., Zhang, Y., Xie, Q., Tian, G.: Incorporating word embeddings into topic modeling of short text. Knowl. Inf. Syst. **61**(2), 1123–1145 (2018). https://doi.org/10.1007/s10115-018-1314-7
8. Guo, A., Yang, T.: Research and improvement of feature words weight based on TFIDF algorithm. In: 2016 IEEE Information Technology. Networking, Electronic and Automation Control Conference, pp. 415–419. IEEE, Chongqing, China (2016)
9. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite BERT for self-supervised learning of language representations. In: International Conference on Learning Representations, pp. 26–30. OpenReview.net, Addis Ababa, Ethiopia (2020)
10. Li, J., Cai, Y., Cai, Z., Leung, H., Yang, K.: Wikipedia based short text classification method. In: Bao, Z., Trajcevski, G., Chang, L., Hua, W. (eds.) DASFAA 2017. LNCS, vol. 10179, pp. 275–286. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-55705-2_22
11. Li, Y.: Short text classification improved by feature space extension. In: IOP Conference Series: Materials Science and Engineering, vol. 533, p. 012046 (2019)
12. Lin, Y., et al.: BERTGCN: Transductive text classification by combining GNN and BERT. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 1456–1462. Association for Computational Linguistics, Bangkok, Thailand (2021)
13. Liu, W., et al.: K-BERT: enabling language representation with knowledge graph. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 2901–2908 (2020)
14. Liu, Y., et al.: Roberta a robustly optimized BERT pretraining approach. CoRR abs/1907.11692 (2019)
15. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
16. Peinelt, N., Nguyen, D., Liakata, M.: tBERT: topic models and BERT joining forces for semantic similarity detection. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7047–7055 (2020)
17. Sharma, P., Li, Y.: Self-supervised contextual keyword and keyphrase retrieval with self-labelling. arXiv e-prints (2019)
18. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification? In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) CCL 2019. LNCS (LNAI), vol. 11856, pp. 194–206. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32381-3_16
19. Sun, Y., et al.: Ernie 2.0: a continual pre-training framework for language understanding. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8968–8975. AAAI, New York, NY, USA (2020)
20. Wang, Y., Wang, S., Yao, Q., Dou, D.: Hierarchical heterogeneous graph representation learning for short text classification. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 3091–3101. Association for Computational Linguistics, Cana, Dominican Republic (2021)
21. Yao, L., Mao, C., Luo, Y.: Graph convolutional networks for text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 7370–7377. AAAI, Honolulu, Hawaii, USA (2019)