Effective Seed-Guided Topic Labeling for Dataless Hierarchical Short Text Classification

Yi Yang[1,3], Hongan Wang[1,3], Jiaqi Zhu[1,2,3](✉), Wandong Shi[1,3],
Wenli Guo[1], and Jiawen Zhang[1,3]

[1] SKLCS, Institute of Software, Chinese Academy of Sciences, Beijing, China
{yangyi2012,hongan}@iscas.ac.cn, zhujq@ios.ac.cn
[2] Zhejiang Lab, Hangzhou, Zhejiang, China
[3] University of Chinese Academy of Sciences, Beijing, China
{shiwandong18,guowenli17,zhangjiawen181}@mails.ucas.edu.cn

Abstract. Hierarchical text classification has a wide application prospect on the Internet, which aims to classify texts into a given hierarchy. Supervised methods require a large amount of labeled data and are thus costly. For this purpose, the task of dataless hierarchical text classification has attracted more and more attention of researchers in recent years, which only requires a few relevant seed words for given categories. However, existing approaches mainly focus on long texts without considering the characteristics of short texts, so are not suitable in many scenarios. In this paper, we tackle dataless hierarchical short text classification for the first time, and propose an innovative model named Hierarchical Seeded Biterm Topic Model (HierSeedBTM), which effectively leverages seed words in Biterm Topic Model (BTM) to guide the hierarchical topic labeling. Specifically, our model introduces iterative distribution propagation mechanism among topic models in different levels to incorporate the hierarchical structure information. Experiments on two public datasets show that the proposed model is more effective than the state-of-the-art methods of dataless hierarchical text classification designed for long texts.

Keywords: Hierarchical text classification · Topic model · Seed word

1 Introduction

With the rapid development of social media, short texts are increasing and widespread on the internet, when people obtain and exchange information through tweets, reviews, and queries. It is important to acquire interesting information from these huge number of short texts with text classification. In many

This work is supported by National Key Research and Development Program of China (2018YFC0116703), Strategic Priority Research Program of Chinese Academy of Sciences (XDC02060500), and Zhejiang Lab (2020NF0AC02).

M. Brambilla et al. (Eds.): ICWE 2021, LNCS 12706, pp. 271–285, 2021.
https://doi.org/10.1007/978-3-030-74296-6_21

scenarios, the category labels of short texts are often organized in a hierarchical structure. Hierarchical text classification (HTC) aim to classify text into a given hierarchy, has a wide variety of applications such as search result classification [2], review classification [20] and sentiment classification [24]. Comparing with flat text classification, HTC leverages the interrelationships among hierarchical structure, and acquires more accurate classification results.

Some models [7,9] adopt a greedy strategy: a local classifier is trained for each category node, and then the classification results are propagated to the next level in a top-down manner. However, the greedy strategy may lead to classification error propagation along the hierarchy. Other models [1] employ the down-up backpropagation strategy: the leaf level is classified at first, and the results are propagated to the top level. Although the models above are widespread used, the lack of plentiful labeled data for training the classifiers limits the application scenario of these models, since carefully-labeled documents require domain expertise and are thus costly.

For this purpose, many researchers focused on dataless hierarchical text classification task as it can successfully reduce the effort in labeling documents. Xiao et al. [26] proposed a generative framework to leverage the hierarchical structural information and compute path-generated probability to classify documents. Meng et al. [18] utilizes a class distribution to generate pseudo documents for training local classifiers and then iteratively refine the global hierarchical model. However, these seed-guided dataless hierarchical text classification methods are designed for long texts without considering the characters of short texts, which are extremely sparse so that only limited features are available to train a classifier. Hence, these models get unsatisfactory performance for classifying short texts.

Recently, some dataless short text classification approaches [11,28] are proposed in a generative framework and achieve significant improvement, which guides the topic labeling process based on a short text topic model to alleviate the data sparsity. Inspired by these studies, we tackle the dataless hierarchical short text classification task for the first time and propose a model named HierSeedBTM. Specifically, we at first calculate the semantic similarity between document words (biterms) and categories as prior knowledge in each level separately, through integrating the seed words along the path from the current category node to the root node in the category hierarchy. That directly guides the generative process of BTM-like topic model for hierarchical topic labeling and inference. Then, an iterative distribution propagation mechanism among topic models in different levels is introduced to incorporate the hierarchical and structural information to make up for the limited data.

In summary, the main contributions of this paper include:

(1) A model HierSeedBTM is presented to solve the task of dataless hierarchical short text classification with seed words, by combining word co-occurrence information and category-word semantic similarity based on word embeddings. To the best of our knowledge, it is the first successful work to tackle the task of dataless hierarchical short text classification.

(2) To effectively utilize the hierarchical structure, a novel iterative propagation mechanism is put forward during the topic sampling process of the topic model. Topic distribution and topic-word distribution are propagated in a top-down manner respectively in each sampling iteration.
(3) Informative experiments are conducted on two hierarchical short text datasets to show that our model outperforms the state-of-the-art baseline methods designed for long texts, especially when the documents are very short.

The remainder of the paper is organized as follows. In Sect. 2, we review recent related work. In Sect. 3, we formalize the problem to be tackled. Section 4 introduces our model in detail. In Sect. 5, the experimental results on hierarchical short text datasets are shown. Section 6 concludes this paper and discusses future work.

2 Related Work

2.1 Dataless Text Classification

Dataless text classification attracts much attention as it only needs a few user-provided seed words for classification, which can successfully reduce the effort in labeling documents. Some researchers studied the methods to generate pseudo-labels or pseudo-documents utilizing seed words for constructing training dataset, and then classify texts with a supervised classification model [6,15]. In particular, Mekala et al. [16] proposed a contextualized weak supervision framework (ConWea) for text classification. The model generates the pseudo-label of documents based on the frequency of seed words, and leverages contextualized representations of words to iteratively train the classifier and expand seed words. Meng et al. [17] proposed a novel weakly-supervised text classification model (WeSTClass). It constructs a semantic space to generate pseudo-documents to train a neural classifier, then fits unlabeled data through bootstrapping.

Another group of researchers studied topic-based models [11,12], in which the generative process is guided by seed words to form category-aware topics. Li et al. [13] proposed Seed-Guided Topic Model (STM), assuming that each document is associated with a single category-aware topic and a mixture of general topics. Yang et al. [28] proposed Seeded Biterm Topic Model (SeedBTM) for dataless short text classification, which leverages both word co-occurrence information from BTM and category-word semantic similarity from word embeddings to classify short texts. All the methods above are designed for flat text classification.

2.2 Dataless Hierarchical Text Classification

As hierarchical text classification can obtain more accurate classification results by leveraging the interrelated structure of categories in different levels, some dataless hierarchical text classification models are studied. Song et al. [23] proposed a dataless hierarchical text classification framework, which firstly represents the document semantics with three kinds of methods, then calculates the semantic similarity between documents and categories as local classification results, and finally adopts a standard hierarchical classification strategy to classify documents. Meng et al. [18] leveraged seed words to model the category semantics as a mixture of von Mises Fisher distributions, and generated meaningful pseudo-documents with LSTM-based language model. Then, the local classifiers are trained and the global hierarchical model is refined iteratively. Xiao et al. [26] proposed a generative framework for weakly hierarchical text classification. It puts a path-dependent score to the cost-sensitive learning algorithm and makes the classification consistent with the category hierarchy during the inference process. However, None of the above models consider the characteristics of short texts.

2.3 Topic Models for Short Texts

Many researchers proposed short text topic models to alleviate the problem of data sparsity. Some studies adopt aggregation strategy [22,25] to generate long documents by aggregating short texts. Others are based on the assumption that each document has only a single latent topic [29,30]. In another direction, Yan et al. [27] proposed BTM, which models the word co-occurrence explicitly and aggregates patterns in the whole corpus for learning topics. Many researches extended BTM with additional information [3,14] or aggregating strategy [8] to make up for the limited data. Our approach selects BTM as the base model, since the model is more flexible for different scenarios and can easily be extended.

3 Preliminaries

In dataless hierarchical short text classification, the class categories constitute a hierarchy τ. It is a tree structure of depth H, and the node in depth 0 is defined as ROOT. The categories of τ are distributed from depth 1 to H. Following the definition in [26], all leaf nodes are in depth H, which can always be satisfied by giving the shallower leaf node a child node until the node reaches depth H.

The categories at each level are denoted as $C_1, C_2, ..., C_H$, with sizes M_1, M_2, ..., M_H, respectively. The category $c_{h,k} \in C_h$ means the category k in the level h, and its representative seed word set is $S_{h,k} = \{s_{h,k,1}, s_{h,k,2}, ..., s_{h,k,l}\}$.

Given an unlabeled document set $D = \{D_1, D_2, ..., D_N\}$, a category hierarchy tree τ and the corresponding seed word sets S, the task of dataless hierarchical short text classification is to assign the most likely category $c_{h,k}$ for each level h to each document D_i. Table 1 summarizes the main notations used in this paper.

4 Proposed Models

In this section, we at first present the overview of our approach and then elaborate the key steps. The first step is estimating category-path-biterm similarity and the second is inferring modified biterm-topic distributions based on topic model for calculating document categories (Fig. 1).

Table 1. Notation in this paper.

Symbol	Description
w	Word
b	Word pair
h	The level in the hierarchy
N	The number of biterm set
B	Biterm set
H	The number of levels in the hierarchy
z_h	The topic in level h
M_h	The number of topics in level h
θ'_h	The modified prior topic distribution in level h
θ^b_h	The modified topic distribution from the dot product of θ_h and $\zeta_{b,h}$
$\zeta^p_{b,h}$	The similarity vector between b and topics in level h
$\phi_{h,z}$	The topic-word distribution of topic z in level h
α_h, β	Hyper-parameter

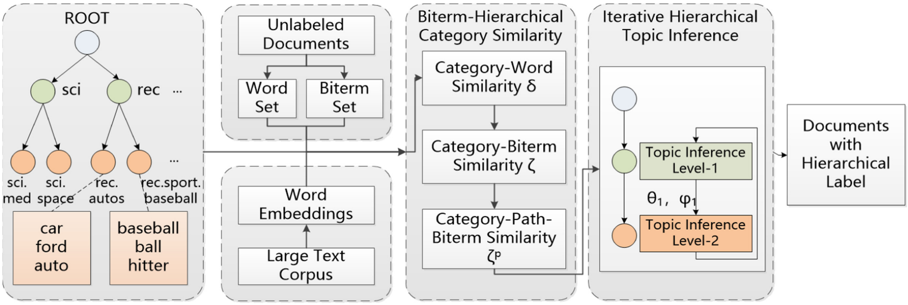

Fig. 1. Overview of the dataless hierarchical short text classification approach.

4.1 Estimating Category-Path-Biterm Similarity

In BTM, a biterm $b_{i,j}$ consists of two words w_i and w_j, which are co-occurring in the same short text regardless of the order. Given a biterm set B, a category hierarchy τ as well as the corresponding seed word set S, our purpose is to calculate semantic similarity between each biterm and each category for guiding the topic labeling.

We at first calculate the semantic similarity between each corpus word w and each seed word s in $c_{h,k}$ through external word embeddings, as it is easy accessible, less time consuming, and can provide external similarity information to alleviate the data sparsity [28]. Obviously, the similarity calculation method can be easily replaced by other methods [18,23] based on different scenarios. We get the word vectors v_s and v_w of a corpus word w and a seed word s respectively, and then calculate the semantic similarity $\text{sim}(s, w)$ as follows:

$$\mathrm{sim}(s, w) = \max(\cos(v_s, v_w), \epsilon) \tag{1}$$

The threshold $\epsilon > 0$ is the lower bound of $\mathrm{sim}(s, w)$ to make it positive. Then, we calculate the category-word similarity $\delta_{h,k,w}$, which is the maximal similarity between each seed word in $S_{h,k} = \{s_{h,k,1}, s_{h,k,2}, ..., s_{h,k,n}\}$ and the document word w:

$$\delta_{h,k,w} = \max_i \left(\mathrm{sim}(s_{h,k,i}, w)\right) \tag{2}$$

Next, we calculate the category-biterm similarity for each biterm b_{w_1,w_2} as the mean value of δ_{h,k,w_1} and δ_{h,k,w_2}:

$$\zeta_{b,h,k} = (\delta_{h,k,w_1} + \delta_{h,k,w_2})/2 \tag{3}$$

To leverage the hierarchical and structural information, the path from ROOT to the leaf node $c_{h,k}$ is denoted as $p_{h,k}$ (abbreviated as p when there is no ambiguity for the leaf node), and our model propagates the semantic similarity along p to obtain the category-path-biterm similarity $\zeta^p_{b,h,k}$ as follows:

$$\zeta^p_{b,h,k} = \sum_{i=1}^{h} \zeta_{b,i,u(k,p,i)} \tag{4}$$

where $u(k, p, i)$ is denoted as the upper category of category k in level i of path p, and $\zeta^p_{b,h,k}$ is the sum similarity of path p from $c_{1,u(k,p,1)}$ to $c_{h,u(k,p,h)}$.

4.2 Hierarchical Seeded Biterm Topic Model

Obviously, the semantic similarity ζ^p can be utilized as biterm-category distribution directly to calculate the document label, but the similarity only contains the external word embedding information without considering the co-occurrence information. Similar to [26], we propose a generative model to modify the semantic similarity based on BTM, named HierSeedBTM, which leverages the prior similarity information and hierarchical structure information to guide the topic labeling.

In HierSeedBTM, the generative process of a topic z_h is influenced by both prior category-path-biterm similarity ζ^p and the topic-word distribution ϕ_z, which guides the model to induce category-aware topics. Moreover, in order to leverage the hierarchical and structural information, the topic distribution θ_h and topic-word distribution ϕ_h are propagated along the path as prior parameters to influence the generative process of the lower level. Notice that the topic number in each level of τ is different, so we define α_h with different dimensions to represent the hyper-parameter of prior topic distribution in each level.

The graphical representation of HierSeedBTM is described in Fig. 2, and the generative process of HierSeedBTM is as follows:

1. Modify the prior topic probability $\theta'_{h,k} \leftarrow \theta_{h-1,u(k,p,h-1)}$
2. Draw a topic-word distribution $\theta_h \sim Dir(\alpha_h + \theta'_h)$
3. For each topic $k = 1, ..., M_h$ in level h
   (a) Draw a topic-word distribution $\phi_h \sim Dir(\beta + \phi_{h-1})$

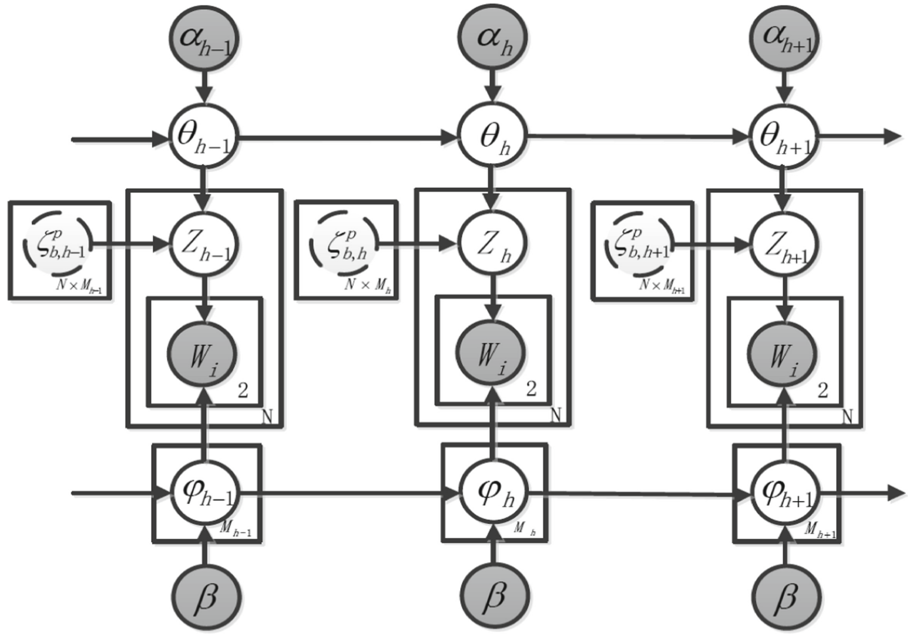

Fig. 2. Graphical representation of HierSeedBTM.

4. For each biterm b in the biterm set B
   (a) Modify the topic distribution $\theta^b_h \propto \zeta^p_{b,h} \cdot \theta_h$
   (b) Draw a topic $z_h \sim Multi(\theta^b_h)$
   (c) Draw two words to form b: $w_i, w_j \sim Multi(\phi_{h,z_h})$

As the category number of level h is different from that of level $h-1$, we assign the topic probability $\theta_{h-1,u(k,p,h-1)}$ in level $h-1$ to the prior topic probability $\theta_{h,k}$ in level h along all paths.

Inference via Gibbs Sampling. Similar to BTM, as the model is intractable, we utilize the Gibbs Sampling to perform the approximate inference. After random initialization on the Markov chain, we iteratively calculate the conditional distribution $P(z_b = z|\mathbf{z}_{\neg b}, B, \boldsymbol{\zeta}, h, \theta_{h-1}, \phi_{h-1})$ for each biterm in each level.

$$\begin{aligned} P(z_b = z|\mathbf{z}_{\neg b}, B, \boldsymbol{\zeta}, h, \theta_{h-1}, \phi_{h-1}) &\propto \zeta^p_{b,z} \cdot (n_z + \theta'_{h,z} + \alpha) \\ \cdot \frac{(n_{w_j|z} + \phi_{h-1,u(k,p,h-1),w_j} + \beta)(n_{w_i|z} + \phi_{h-1,u(k,p,h-1),w_i} + \beta)}{\sum_w(\sum_w n_{w|z} + \phi_{h-1,u(k,p,h-1),w_j} + M\beta)(n_{w|z} + 1 + \phi_{h-1,u(k,p,h-1),w_i} + M\beta)} \end{aligned} \quad (5)$$

where $z_{\neg b}$ is denoted the topic assignments for all biterms except b, n_z is the number of biterms assigned to the topic z, and $n_{w|z}$ is the number of times when the word w is assigned to the topic z. When sampling on level 1, the parent node is ROOT, so the $\theta_{h-1,u(k,p,h-1)}$ and $\phi_{h-1,u(k,p,h-1)}$ are all zero.

With the sampling results, the topic distribution θ_h and topic-word distribution ϕ_h can be calculated as:

$$\theta_{h,z} = \frac{(n_z + \theta'_{h,z} + \alpha)}{|B| + \sum_{k=1}^{M_h}(\theta'_{h,k}) + M_h\alpha} \tag{6}$$

$$\phi_{h,w|z} = \frac{(n_{w|z} + \phi_{h-1,u(k,p,h-1),w} + \beta)}{\sum_{w'}(n_{w'|z} + \phi_{h-1,u(k,p,h-1),w'}) + N\beta} \tag{7}$$

The whole Gibbs Sampling process is shown in Algorithm 1:

Algorithm 1. Gibbs Sampling Process of HierSeedBTM

```
INPUT:B, $M_h$, $\zeta$, $\alpha_h$, $\beta$, $\theta_{h-1}$, $\phi_{h-1}$
OUTPUT: $\theta_h$, $\phi_h$
for $b \in B$ do
    Initialize the topic assignment of $b$;
for $iter = 1$ to $N_{iter}$ do
    for $h = 1$ to $H$ do
        for $b_{w_i,w_j} \in B$ do
            Sample the topic $z$ of $b$ with Equation 5 ;
            Update $n_z, n_{w_i|z}, n_{w_j|z}$;
            Calculate $\theta_h$ with Equation 6 and $\phi_h$ with Equation 7;
```

Predicting Document Category. The classification results of our model are obtained from the leaf level, and the probabilities of inner categories are calculated by summing up the probabilities of their child categories. Specifically, we treat the expectation of the topic proportions of biterms in a document as the topic proportions of the document [27], as shown below:

$$P(z_{h,k}|d,h) = \sum_b P(z_{h,k,b}|b,h)P(b|d) \tag{8}$$

where $P(z_{h,k,b}|b,h)$ can be calculated by Eq. 5 and $P(b|d)$ is estimated based on the relative frequency of b in d. Finally, for document d, the category label z_d can be predicted as the topic with the highest probability:

$$z_{d,h} = \arg\max_k P(z_{h,k}|d,h) \tag{9}$$

5 Experiments

5.1 Datasets

We use two hierarchical short text datasets to evaluate the effectiveness of our model. For both datasets, we at first lower and lemmatize all corpus words and remove stop words, and then filter out the documents that contain only one word or more than 50 words to obtain the short text datasets. The statistics of the two datasets are shown in Table 2.

Table 2. Statistics of datasets.

Dataset	Categories (Level 1 + Level 2)	Documents	Average document length
20NG	7 + 20	10493	23
HuffPost	3 + 9	37054	3.8

- The 20 Newsgroups(20NG)[1] [10] : 20 Newsgoups is a widely used dataset for the text classification task, including flat text classification and hierarchical text classification. It contains about 20,000 newsgroups messages from 20 newsgroups. After filtering long documents, we retain about 10,000 messages as short text datasets. As the categories are close to each other, 20NG can be categorized into a hierarchical structure with two levels. The upper level contains 7 inner categories, and the leaf level contains 20 categories [26].
- HuffPost[2] [19] : It is obtained from HuffPost and contains around 200K news headlines from 2012 to 2018. There are totally 42 categories, and many of them are overlapping, from which we select 9 leaf categories with 3 upper categories. The categories and their statistics are shown in Table 3.

Table 3. Categories and their statistics of the HuffPost dataset.

Category in level 1	Categories in level 2
Family (5)	WEDDINGS (1488), HOME & LIVING (3271), PARENTING(5437),PARENTS (3443), DIVORCE (1418)
Healthy (2)	HOME & LIVING (3271), WELLNESS (10655)
Food (2)	FOOD & DRINK (3890), TASTE (1862)

The number in the first column indicates the respective child category number, and the number in the second column indicates the document number of each category.

5.2 Baselines

We evaluate our model with six baseline models, all of which can leverage seed words to deal with the task of dataless text classification. The first four methods aim at dataless hierarchical classification for general texts, and the last two are dataless short text classification models.

[1] http://qwone.com/jason/20Newsgroups/.
[2] https://www.kaggle.com/rmisra/news-category-dataset.

- Hier-Dataless[3] [23]: it introduces three ways to calculate the semantic similarity between document and category with seed words as document-category probabilities. To be fair, we choose the same word embedding as the semantic representation to calculate semantic similarity.
- WeSHClass[4] [18]: it is a successful weakly-supervised hierarchical text classification model, which leverages seed words to generate pseudo-labeled documents, and then iteratively refines the global hierarchical classifier with supervised methods.
- PCNB[5] [26]: it is a state-of-the-art weakly-supervised hierarchical text classification model within a generative framework. The model adopts the initial semantic similarity calculation method of Hier-Dataless, and constructs a path-cost sensitive Bayes classifier.
- PCEM [26]: it improves the path-cost sensitive classifier and adopts EM technique for semi-supervised hierarchical text classification.
- SeedBTM [28]: it is a state-of-the-art dataless short text classification approach. This model leverages both word co-occurrence information and prior category-word similarity to classify short texts, utilizing seed words to calculate category-word similarity to guide the topic-word distributions of BTM.
- SeedBTM* : it is a variant of SeedBTM with category-biterm similarity ζ to guide the generative process rather than category-word similarity.

For all baseline models, we adopt their implementation codes and parameter settings directly.

5.3 Experiment Settings

For seed words of 20NG, we adopt the setting of Hier-Dataless [23], and for HuffPost, we use descriptive LDA (DescLDA)[4] to select 3∼9 representative words for each category as seed words, as shown in Table 4.

For all datasets, we set the topic number $K'_h = M_h$ (category number in each level), $\alpha = 50/K'_h$, $\beta = 0.3$ and $\epsilon = 0.0001$. We set the number of iterations to 8 as our model achieves competitive performance since then. For word embeddings, we employ the widely used GloVe Common Crawl[6] [21], which contains 840B tokens, 2.2M vocab and 300d vectors. It is also used in baseline models Hier-Dataless, PCNB, PCEM, SeedBTM, and SeedBTM*. WeSHClass adopts the self-training word embeddings based on the unlabeled documents.

[3] https://github.com/CogComp/cogcomp-nlp/tree/master/dataless-classifier.

[4] https://github.com/yumeng5/WeSHClass.

[5] https://github.com/HKUST-KnowComp/PathPredictionForTextClassification.

[6] http://nlp.stanford.edu/data/glove.840B.300d.zip.

Table 4. Seed words of the HuffPost dataset.

Category	Seed words
FAMILY	Family adorable divorce daughter parent
HEALTH	Health weight yoga mental drug cancer disease doctor
FOOD	Food sweet cake chocolate cheese
HOME & LIVING	Home house room
DIVORCE	Divorce child parent relationship kid couple split
PARENTS	Mom kid parent baby dad girl
WEDDINGS	Wedding marriage couple bride love bridal
PARENTING	Kid child parent mom teach study learn life
HEALTHY LIVING	Health life cancer mental care
WELLNESS	Wellness cancer drug heart weight stress
FOOD & DRINK	Food cook taste wine cake chocolate
TASTE	Taste cook dinner ice breakfast wine meal delicious sweet

5.4 Experimental Results

We evaluate the classification performances of HierSeedBTM using Macro-F1 and Micro-F1. For each model, we calculate the F1 scores for Level-1 and Level-2 based on the hierarchy, and then get the F1 scores for all categories of Level-All in general. We run 10 times on each dataset to get the average values, as shown in Table 5 and Table 6.

Table 5. Macro-F1 (%) of HierSeedBTM and baselines on all datasets.

Dataset	20NG			HuffPost		
	Level-l	Level-2	Level-all	Level-l	Level-2	Level-all
HierSeedBTM	53.5	**48.1**	**49.5**	**71.2**	**41**	**48.5**
WeSHClass	**54.7**	37.7	41.2	61.7	30.7	38.9
PCNB	35.8	22	25.5	66.2	34.8	42.7
PCEM	47.9	30.9	35.3	67.6	40.3	47.1
Hier-Dataless	46.2	34.4	37.5	65.3	34.7	42.4
SeedBTM	38.9	25.8	29.2	65	38.6	45.2
SeedBTM*	42	34	36.1	65.1	38.4	45.1

The best results in the table are highlighted in bold, and we can observe that HierSeedBTM performs better than almost all baseline models in both datasets.

For Macro-F1 in Table 5, on the 20NG dataset, HierSeedBTM increases 8.3 than the second WeSHClass and increases 12 than PCEM in level-all.

That certifies our model can make better use of hierarchical structures to improve the classification accuracy of leaf nodes. The small deficiency against WeSHClass in level-1 can be attributed to the reduced difficulty of classifying texts with moderate length in an abstract level. On the HuffPost dataset, HierSeedBTM improves 1.4 than PCEM and 9.6 than WeSHClass in level-all. For Micro-F1 in Table 6, HierSeedBTM is in line with WeSHClass on the 20NG dataset, but increases 5.8 than WeSHClass on the HuffPost dataset in level-all.

Table 6. Micro-F1 (%) of HierSeedBTM and baselines on all datasets.

Dataset	20NG			HuffPost		
	Level-1	Level-2	Level-all	Level-1	Level-2	Level-all
HierSeedBTM	65.6	**49.4**	**57.5**	**70.8**	**37.6**	**54.2**
WeSHClass	**73.8**	41.1	**57.5**	62.8	32.7	48.4
PCNB	52.5	27.8	40.1	66.7	35.4	51
PCEM	63.2	39.7	51.5	67.7	38.4	53.1
Hier-Dataless	55.4	36.0	45.8	65.2	34.0	49.6
SeedBTM	54	24.7	32.9	66.1	35.7	51.1
SeedBTM*	57.4	31.2	36.2	66	35.3	50.8

For WeSHClass, the pseudo-labeled documents are relatively noisy and the trained classifier cannot well distinguish short texts due to the sparse feature, so WeSHClass gets poorer performance in HuffPost than the generative models. For PCEM, only category-word similarities are propagated along paths, but our model further propagates the distributions among each iteration and gets better performance. The results indicate that the propagation mechanism and generative framework for short texts make our model more effective to leverage the hierarchical structure information and alleviate the data sparsity of short texts. Moreover, the performances of baseline models fluctuate among datasets with different lengths, while HierSeedBTM behaves more stable.

Compared with SeedBTM and the variant model SeedBTM*, HierSeedBTM increases 3.3∼12 in Macro-F1 and 3.1∼21.4 in Micro-F1 respectively, which explains that our propagation strategy based on the hierarchical structure can integrate more evidences from different abstraction levels to classify short texts accurately.

5.5 Parameter Study

In this section, we study the impact of different parameter settings on the classification performance of HierSeedBTM. When paying attention to one parameter, other parameters are fixed to the default values given in Sect. 5.3.

The Impact of Iteration Number. Iteration number is an important parameter to our model, because the smaller of this number means the more efficient of our model. We change the value in the range of [1,20], and its Macro-F1 and Micro-F1 results are shown in Fig. 3.

When the number of iterations is 6, HierSeedBTM can achieve good classification performance, and F1-score is almost stable after this point. Therefore, we set the iteration number to 8 in our model. The fast convergence should give credit to the regulating effect of information propagation mechanism among the hierarchical structure and the prior knowledge from word embeddings.

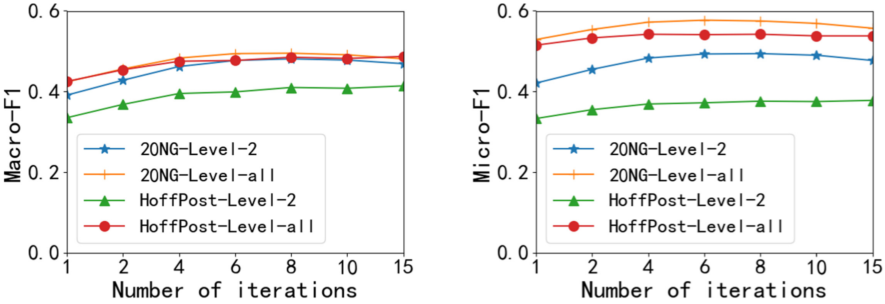

Fig. 3. F1 values of level-2 and level-all with different iterations.

The Impact of β. β is the hyper-parameter of the topic model, which affects the topic word distribution ϕ. We vary the value in the range of [0.01, 0.6] to evaluate its impact, and the results are shown in Fig. 4.

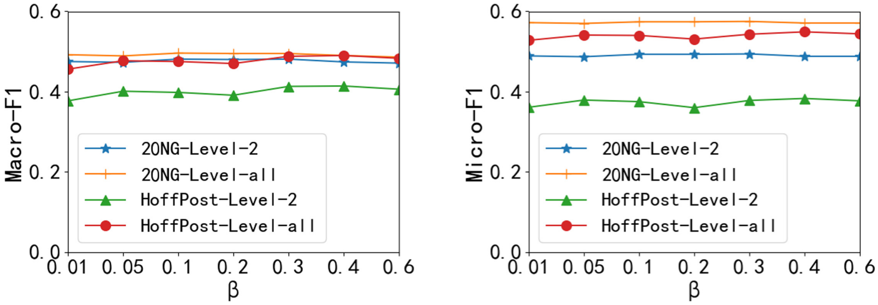

Fig. 4. F1 values of level-2 and level-all with different β values.

The flat lines of the 20NG dataset indicates that the dataset with a longer text length is relatively insensitive to β. But for the shorter HuffPost, the results fluctuate and arrive at the best performance in the range of [0.3, 0.4], so we set β to 0.3 in our model.

For the other hyper-parameter α, $\alpha = 50/K'_h$ is a widely used parameter setting of topic models [13,27], and the experimental results also indicate that the classification accuracy of our model is insensitive to α. Hence, the detailed comparison is omitted here due to the page limit.

6 Conclusion

In this paper, we propose an effective model for dataless hierarchical short text classification. Our model leverages seed words to guide the generative process of BTM for topic labeling and introduces an iterative distribution propagation mechanism to incorporate the hierarchical and structural information. Moreover, the propagation mechanism brings efficient performance because of the fast convergence. Experiments on both hierarchical short text datasets show that our model performs better than other baseline methods, especially when the length of the document is extremely short and more sparse.

In the future, we plan to study how to incorporate contextualized vector representation [5] to better tackle this task. In addition, it is important to study the impact factors of datasets for classification accuracy, such as category imbalance and bias as well as different hierarchical structures.

References

1. Bennett, P., Nguyen, N.: Refined experts: improving classification in large taxonomies. In: SIGIR, pp. 11–18. ACM (2009)
2. Chen, H., Dumais, S.T.: Bringing order to the web: automatically categorizing search results. In: CHI, pp. 145–152. ACM (2000)
3. Chen, W., Wang, J., Zhang, Y., Yan, H., Li, X.: User based aggregation for biterm topic model. In: ACL, vol. 2 (Short Papers), pp. 489–494 (2015)
4. Chen, X., Xia, Y., Jin, P., Carroll, J.: Dataless text classification with descriptive LDA. In: AAAI, pp. 2224–2231 (2015)
5. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT, pp. 4171–4186. Association for Computational Linguistics (2019)
6. Druck, G., Mann, G., McCallum, A.: Learning from labeled features using generalized expectation criteria. In: SIGIR, pp. 595–602. ACM (2008)
7. Dumais, S.T., Chen, H.: Hierarchical classification of web content. In: SIGIR, pp. 256–263. ACM (2000)
8. Jiang, L., Lu, H., Xu, M., Wang, C.: Biterm pseudo document topic model for short text. In: ICTAI, pp. 865–872. IEEE (2016)
9. Koller, D., Sahami, M.: Hierarchically classifying documents using very few words. In: ICML, pp. 170–178. Morgan Kaufmann (1997)
10. Lang, K.: Newsweeder: learning to filter netnews. In: ICML, pp. 331–339. Morgan Kaufmann (1995)
11. Li, C., Chen, S., Qi, Y.: Filtering and classifying relevant short text with a few seed words. Data Inf. Manag. **3**(3), 165–186 (2019)
12. Li, C., Chen, S., Xing, J., Sun, A., Ma, Z.: Seed-guided topic model for document filtering and classification. ACM Trans. Inf. Syst. **37**(1), 9:1–9:37 (2019)

13. Li, C., Xing, J., Sun, A., Ma, Z.: Effective document labeling with very few seed words: a topic model approach. In: CIKM, pp. 85–94. ACM (2016)
14. Li, X., Zhang, A., Li, C., Guo, L., Wang, W., Ouyang, J.: Relational biterm topic model: short-text topic modeling using word embeddings. Comput. J. **62**(3), 359–372 (2018)
15. Liu, B., Li, X., Lee, W.S., Yu, P.S.: Text classification by labeling words. In: AAAI, vol. 4, pp. 425–430 (2004)
16. Mekala, D., Shang, J.: Contextualized weak supervision for text classification. In: ACL, pp. 323–333. Association for Computational Linguistics (2020)
17. Meng, Y., Shen, J., Zhang, C., Han, J.: Weakly-supervised neural text classification. In: CIKM, pp. 983–992. ACM (2018)
18. Meng, Y., Shen, J., Zhang, C., Han, J.: Weakly-supervised hierarchical text classification, vol. 33, no. 01, pp. 6826–6833 (2019)
19. Misra, R.: News category dataset (2018). https://doi.org/10.13140/RG.2.2.20331.18729
20. Pang, B., Lee, L., Vaithyanathan, S.: Thumbs up? sentiment classification using machine learning techniques. In: EMNLP, pp. 79–86 (2002)
21. Pennington, J., Socher, R., Manning, C.: Glove: global vectors for word representation. In: EMNLP, pp. 1532–1543. Association for Computational Linguistics (2014)
22. Quan, X., Kit, C., Ge, Y., Pan, S.J.: Short and sparse text topic modeling via self-aggregation. In: IJCAI, pp. 2270–2276 (2015)
23. Song, Y., Roth, D.: On dataless hierarchical text classification. In: AAAI, pp. 1579–1585. AAAI Press (2014)
24. Tang, D., Qin, B., Liu, T.: EMNLP, pp. 1422–1432. The Association for Computational Linguistics (2015)
25. Weng, J., Lim, E.P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential Twitterers. In: WSDM, pp. 261–270. ACM (2010)
26. Xiao, H., Liu, X., Song, Y.: Efficient path prediction for semi-supervised and weakly supervised hierarchical text classification. In: WWW, pp. 3370–3376. ACM (2019)
27. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: WWW, pp. 1445–1456. ACM (2013)
28. Yang, Y., et al.: Dataless short text classification based on biterm topic model and word embeddings. In: Bessiere, C. (ed.) International Joint Conferences on Artificial Intelligence Organization, IJCAI, pp. 3969–3975 (2020)
29. Yin, J., Wang, J.: A Dirichlet multinomial mixture model-based approach for short text clustering. In: SIGKDD, pp. 233–242. ACM (2014)
30. Zhao, W.X., et al.: Comparing twitter and traditional media using topic models. In: Clough, P., et al. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 338–349. Springer, Heidelberg (2011). https://doi.org/10.1007/978-3-642-20161-5_34