Dataless Short Text Classification Based on Biterm Topic Model and Word Embeddings

Yi Yang[1,3], **Hongan Wang**[1,2,3], **Jiaqi Zhu**[1,2,3*], **Yunkun Wu**[3], **Kailong Jiang**[3], **Wenli Guo**[3] and **Wandong Shi**[3]
[1]SKLCS, Institute of Software, Chinese Academy of Sciences
[2]Zhejiang Lab
[3]University of Chinese Academy of Sciences
{yangyi2012, hongan}@iscas.ac.cn, zhujq@ios.ac.cn,
{wuyunkun13, jiangkailong17, guowenli17, shiwandong18}@mails.ucas.edu.cn

Abstract

Dataless text classification has attracted increasing attentions recently. It only needs very few seed words of each category to classify documents, which is much cheaper than supervised text classification that requires massive labeling efforts. However, most of existing models pay attention to long texts, but get unsatisfactory performance on short texts, which have become increasingly popular on the Internet. In this paper, we at first propose a novel model named Seeded Biterm Topic Model (SeedBTM) extending BTM to solve the problem of dataless short text classification with seed words. It takes advantage of both word co-occurrence information in the topic model and category-word similarity from widely used word embeddings as the prior topic-in-set knowledge. Moreover, with the same approach, we also propose Seeded Twitter Biterm Topic Model (SeedTBTM), which extends Twitter-BTM and utilizes additional user information to achieve higher classification accuracy. Experimental results on five real short-text datasets show that our models outperform the state-of-the-art methods, and especially perform well when the categories are overlapping and interrelated.

1 Introduction

It is very popular for people to obtain and exchange information in different applications and websites on the Internet, such as instant messages, social media and news media. A huge number of short texts are generated every day, and it is crucial to acquire important and interesting information from them. Short text classification plays a fundamental role in short text processing. Many recent studies on this task were based on supervised deep learning methods [Wang *et al.*, 2017; Zeng *et al.*, 2018], which achieved significant improvement over traditional classification models. However, the lack of plentiful training data limits the application of these models, since labeling documents is very expensive and time consuming for domain experts.

Dataless text classification has attracted more and more attentions, as it only requires a small set of seed words for each category [Liu *et al.*, 2004; Druck *et al.*, 2008], which are much cheaper than labeling documents. Many existing approaches [Chen *et al.*, 2015b; Li *et al.*, 2016b; Li *et al.*, 2018a] based on topic models took advantage of seed words and word co-occurrence information in a weakly-supervised manner, and performed significantly well in many scenarios. However, these models aim at normal documents, but are not suitable for irregular short texts on the Internet.

Different from long documents, short texts are extremely sparse and only limited word co-occurrence information can be utilized to form a topic and further correspond to a category, so the models above get unsatisfactory results for classifying short texts. [Li *et al.*, 2019] proposed a seed-guided topic model for dataless short text classification and filtering (SSCF), but it costs much time and neglects other types of information to enhance the correlation between categories and words. [Shalaby and Zadrozny, 2019] proposed a concept raw context model (CRX), which employs raw concept mentions from the knowledge base to learn concept embeddings for dataless text classification, but an appropriate even domain-specific knowledge base is indispensable.

Comparatively, word embeddings are widely used nowadays and relatively easy to access. Recently, many researches [Li *et al.*, 2016a; Xun *et al.*, 2016; Li *et al.*, 2018b] made use of word embeddings in topic models for short texts, and demonstrated that word embeddings can effectively help to identify similar words in both syntactic and semantic levels, and this kind of external information can successfully alleviate the data sparsity. Based on this idea, we propose a novel model named Seeded Biterm Topic Model (SeedBTM), which incorporates pre-trained word embeddings into the classical short-text topic model BTM [Yan *et al.*, 2013]. Specifically, we calculate the maximum similarity between a corpus word and seed words of a category to get the category-word similarity score, served as the prior topic-in-set knowledge for topic-word distributions [Andrzejewski and Zhu, 2009]. Moreover, the key idea in this approach can also be extended to other biterm-based topic models. We further

*Corresponding author

proposed Seeded Twitter Biterm Topic Model (SeedTBTM) based on Twitter-BTM [Chen *et al.*, 2015a] when the user (author) information of short texts is available.

In summary, the contributions of this paper include:

(1) An approach is presented to solve the task of dataless short text classification with seed words, by combining word co-occurrence information and category-word similarity based on word embeddings. To the best of our knowledge, it is the first successful work to utilize both short-text topic model and word embeddings to classify short texts in a dataless manner.

(2) Two models SeedBTM and SeedTBTM are respectively proposed through applying our approach on short-text topic models BTM and Twitter-BTM. That indicates our approach is applicable on different topic models by effectively integrating meta information of short texts, such as users (authors).

(3) Informative Experiments are conducted on five real-world datasets to show that our models significantly outperform the state-of-the-art baseline methods, especially when the categories are overlapping and interrelated.

The remainder of the paper is organized as follows. In Section 2, we review recent related work. Section 3 introduces the two models with the uniform approach. In Section 4, the experimental results on real short-text datasets are shown. Section 5 concludes this work and discusses future work.

2 Related Work

2.1 Dataless Text Classification

Many researchers focused on dataless text classification because the proposed models can successfully reduce the effort in labeling documents for text classification. Some work tried to exploit auxiliary knowledge bases to accomplish dataless classification [Chang *et al.*, 2008; Türker, 2019; Shalaby and Zadrozny, 2019]. However, external knowledge bases are hard to obtain in many scenarios and these models are not suitable for classifying documents into specific and fine-grained categories.

Another group of researchers proposed dataless classifiers by utilizing category seed words. Some of these studies were based on pseudo-labels obtained approximately from seed words [Liu *et al.*, 2004; Druck *et al.*, 2008], but these pseudo-labels often result in noisy training data. [Meng *et al.*, 2018] proposed a novel weakly-supervised text classification model (WeSTClass). It at first constructs a semantic space by learning vector representations, then extends category seed words in the semantic space to generate pseudo-documents for training a neural classifier, and finally fits unlabeled data through bootstrapping. However, this model would fail when the categories are interrelated, as that means the semantic sub-spaces of these categories are overlapping and confusing.

Recently, several topic model based approaches have been proposed for dataless text classification with seed words. TLC++ [Hingmire and Chakraborti, 2014] labels topics clustered by Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003] with seed words based on information gain, and transforms topic-word distributions into category-word distributions for classifying documents. [Chen *et al.*, 2015b] proposed descriptive LDA (DescLDA), which constructs some descriptive documents with seed words to guide topic-word distributions, and then labels documents via clustering document-topic distributions. [Li *et al.*, 2016b] proposed Seed-Guided Topic Model (STM). It assumes that each document is associated with a single category topic and a mixture of general topics. The former consisting of seed words is used to determine the document category, and the latter help to capture more word co-occurrence information. As an extension of STM, [Li *et al.*, 2018a] proposed DFC for dataless text filtering and classification. However, these models focus on long documents and cannot accommodate the data sparsity of short texts. Furthermore, [Li *et al.*, 2019] proposed SSCF to handle short texts, which estimates the co-occurrence correlation of seed words and corpus words with a word network topic model (WNTM) [Zuo *et al.*, 2016], but this approach is time consuming and does not make use of word embeddings to supplement word similarity information.

2.2 Topic Models for Short Texts

For short texts, applying conventional topic models directly is less effective due to document-level data sparsity, so many researches aggregate short texts to pseudo-documents through additional metadata [Weng *et al.*, 2010] or heuristic strategies as pre-processing [Quan *et al.*, 2015]. Some other models are based on the assumption that each document has only a single latent topic, like Twitter-LDA [Zhao *et al.*, 2011] and Dirichlet Multinomial Mixture (DMM) [Yin and Wang, 2014]. This assumption is proved to be effective in many scenarios, but sometimes it is too strong to cluster well.

As another branch, [Yan *et al.*, 2013] proposed Biterm Topic Model (BTM), which explicitly models word co-occurrence patterns in the generative process of the whole corpus. [Chen *et al.*, 2015a] further proposed Twitter-BTM by combining BTM and Twitter-LDA, which introduces user information and the background topic into BTM. In addition, [Jiang *et al.*, 2016] proposed biterm pseudo-document topic model (BPDTM) using the word co-occurrence network to construct biterm-based pseudo-documents. [Li *et al.*, 2018b] proposed relational BTM (R-BTM) to link short texts via a similarity matrix of words computed by word embeddings. Our approach selects BTM as the base model, since it is flexible for different scenarios and can easily be extended to combine various aggregation strategies with metadata.

3 Proposed Models

In this section, we at first present the overview of our approach and then elaborate the key step of estimating the category-word similarity. After that, we propose two models SeedBTM and SeedTBTM based on this approach in detail.

3.1 Approach Overview

Given a corpus of unlabeled documents $D = \{d_1, ..., d_N\}$ and a set of target categories $Z = \{z_1, ..., z_K\}$ with respective seed words, our approach aims to assign a category label to each document, taking advantage of external word embeddings. It consists of two steps as Figure 1 illustrates. The

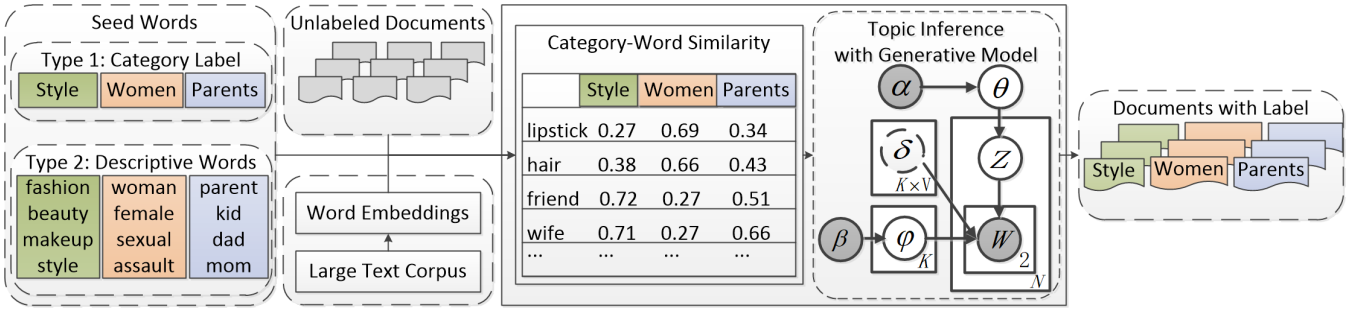


Figure 1: Overview of the dataless short text classification approach

first is to estimate the category-word similarity matrix, which provides prior knowledge about the relation between topics (categories) and words. The second is to infer document-topic distributions, incorporating the information above into the generative process of a biterm-based topic model to adjust and optimize topic-word distributions.

3.2 Estimating Category-Word Similarity

To effectively utilize seed words to guild the classification of short texts, we follow the idea that word similarity scores based on word embeddings could represent the semantic correlations between words [Li *et al.*, 2016a; Li *et al.*, 2018b]. In our models, with a vocabulary $W = \{w_1, ..., w_V\}$ including V words, each category corresponds to a topic z, and has a category-word similarity vector δ_z as the prior topic-in-set knowledge. Next, we will introduce how to calculate the value of each dimension $\delta_{z,w}(w \in W)$ with seed words.

At first, for a category seed word s and a corpus word w, we get the word vectors v_s and v_w through word embeddings, and then calculate the word similarity $\text{sim}(s, w)$ as follows:

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

Notice that the range of the cosine function is $[-1, 1]$, but we hope the similarity score $\text{sim}(s, w)$ can be regarded as a positive weight and lies in $(0, 1]$, so here we set a threshold $\epsilon > 0$ to give a lower bound on $\text{sim}(s, w)$.

Then, for a category z with multiple seed words, $s_{z,1}, ..., s_{z,n_z}$, we compute the category-word similarity $\delta_{z,w}$ as the maximum similarity between each seed word and w:

$$\delta_{z,w} = \max_i(\text{sim}(s_{z,i}, w)) \tag{2}$$

3.3 Seeded Biterm Topic Model

With the similarity scores, we extend the base model BTM and proposed SeedBTM. In BTM, A biterm $b_{i,j}$ contains two words w_i and w_j co-occurring in a short text, regardless of the order. Given a sampled topic z, we think the generation of a word w in SeedBTM is influenced by both prior category-word similarity δ_z and the topic-word distribution ϕ_z. That drives the model to induce category-aware topics in the inference process. As Figure 1 illustrates, the generative process of SeedBTM is as follows:

1. Draw a distribution over topics: $\theta \sim Dir(\alpha)$
2. For each topic $k = 1, ..., T$
   (a) Draw a topic-word distribution $\phi_z \sim Dir(\beta)$
   (b) Modify the topic-word distribution $\phi'_z \propto \delta_z \cdot \phi_z$
3. For each biterm b in the biterm set B
   (a) Draw a topic $z_b \sim Multi(\theta)$
   (b) Draw two words to form b: $w_i, w_j \sim Multi(\phi'_z)$

Compared to BTM, our model transforms the topic-word distribution ϕ_z to ϕ'_z by multiplying the category-word similarity vector δ_z (2b), and samples both words of a biterm following the new distribution ϕ'_z (3b).

Inference via Gibbs sampling. Similar to BTM, after random initialization on the Markov chain, we iteratively calculate the conditional distribution $P(z_b|z_{\neg b}, B, \boldsymbol{\delta})$ for each biterm $b = (w_i, w_j)$, where $z_{\neg b}$ denotes the topic assignments for all biterms except b, and B is the set of biterms in the whole corpus. The formula can be easily obtained by applying the chain rule on the joint probability for all biterms.

$$P(z_b|\mathbf{z}_{\neg b}, B, \boldsymbol{\delta}) \propto \delta_{z,w_i} \cdot \delta_{z,w_j} \cdot (n_z + \alpha) \cdot \frac{(n_{w_i|z+\beta})(n_{w_j|z} + \beta)}{\sum_w (n_{w|z} + 1 + M\beta)(\sum_w n_{w|z} + M\beta)} \tag{3}$$

where 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. Note that δ_{z,w_i} and δ_{z,w_j} are the added terms to introduce the prior knowledge for the two words respectively.

Predicting document category. Like BTM, SeedBTM treats the expectation of the topic proportions of biterms in a document as the topic proportions of the document:

$$P(z|d) = \sum_b P(z|b)P(b|d) \tag{4}$$

where $P(z|b)$ can be calculated by Equation 3 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:

$$z_d = \arg\max_i P(z_i|d) \tag{5}$$

3.4 Seeded Twitter Biterm Topic Model

The key idea above can be extended to other biterm-based topic models. We change the base model BTM to Twitter-BTM, which incorporates user information and the background topic into BTM, and propose SeedTBTM. In this

Dataset	Category Names (Number of Documents in Each Category)	N_u	N_d
SearchSnip	business(1500), software(1500), culture(2210), education(2660), engineering(370), health(1180), politics(1500), sport(1420)	–	17.9
Reuters10T	earn(2476), acq(acquision)(2192), money-supply(153), sugar(143), trade(354), ship(152), crude(395), grainwheat(148), interest(281), money-fx(foreign exchange)(291)	–	3.4
AGnews	business(14101), science(12068), sport(7707), world(14222)	–	4.6
TwitterTrends10	avengers(15762), amtrak188(9436), baltimoreriots(7983), bb17(4001), build2015(3649), homekit(8856), mh370(6322), nepalearthquake(9045), sxsw(3307), wwdc(3165)	1.7	5.6
HuffN10	comedy(2151), sports(2461), business(2794), healthy living(2869), culture & arts(272), education(526), parenting(2543), food & drink(1804), style & beauty(4495), travel(3384)	5.0	5.1
HuffN4-SWPB	business(2794), parents(1973), style(979), women(1517)	3.5	5.1
HuffN4-ECMC	college(628), comedy(2151), education(526), media(1575)	4.8	5.2

Table 1: Categories and statistics of datasets

model, each user u is associated with a multinomial distribution θ^u over K topics drawn from a Dirichlet prior $Dir(\alpha)$, and the background topic $\mathcal{B}$ is associated with a multinomial distribution $\phi^{\mathcal{B}}$ following a Dirichlet prior $Dir(\beta)$. Each word w in any biterms has an indicator y. $y = 0$ means the word w is a background word, while $y = 1$ indicates w is a topic (category) word. y is associated with a uniform Bernoulli distribution π for all users drawn from a Beta prior $Beta(\gamma)$. When generating a topic word, the modified distribution ϕ'_z is adopted similar to SeedBTM. The generative process of SeedTBTM is omitted due to the page limit.

Inference via Gibbs sampling. We also perform Gibbs sampling to calculate the conditional topic distribution of a biterm for a user as follows:

$$P(z_{u,b}|\mathbf{z}_{\neg(u,b)}, B, \mathbf{y}, \boldsymbol{\delta}) \propto \delta_{z,w_{u,b,1}} \cdot \delta_{z,w_{u,b,2}} \cdot (n_z^u + \alpha) \cdot (\frac{(n_{w_{u,b,1}|z} + \beta)}{\sum_w (n_{w|z} + 1 + M\beta)})^{y_{u,b,1}} (\frac{(n_{w_{u,b,2}|z} + \beta)}{(\sum_w n_{w|z} + M\beta)})^{y_{u,b,2}} \tag{6}$$

where n_z^u is the number of biterms assigned to topic z for user u. $w_{u,b,m}$ and $y_{u,b,m}$ are respectively the word and the indicator of the m-th word of biterm b for user u. The distribution of the indicator y_i can be calculated as follow:

$$P(y_{u,b,m} = 0 \mid \mathbf{y}_{\neg(u,b,m)}, \mathbf{z}, B, \boldsymbol{\delta}) \propto (n_{(0)} + \gamma) \cdot \frac{(n_{w_{u,b,m}|\mathcal{B}} + \beta)}{(n_{(.)|\mathcal{B}} + V\beta)} \tag{7}$$

$$P(y_{u,b,m} = 1 \mid \mathbf{y}_{\neg(u,b,m)}, \mathbf{z}, B, \boldsymbol{\delta}) \propto (n_{(1)} + \gamma) \cdot \frac{\delta_{z,w_{u,b,m}}(n_{w_{u,b,m}|z} + \beta)}{\sum_w \delta_{z,w}(n_{w|z} + V\beta)} \tag{8}$$

where $n_{(0)}$ and $n_{(1)}$ are respectively the number of words assigned to the background topic $\mathcal{B}$ and category topics. Note that $\mathcal{B}$ does not have seed words as well as the prior similarity scores with corpus words, so for a fair comparison, we need to normalize the modified term for category-word proportions in the calculation of $P(y_{u,b,m} = 1)$. The prediction for the category label is same as SeedBTM and omitted here.

4 Experiments

4.1 Datasets

We show the model effectiveness on five real short-text datasets, the last two of which contain user information.

- SearchSnip [Li *et al.*, 2019]: Search Snippets is a widely used dataset for short text classification. These texts were selected from results of web search transactions using predefined phrases in different domains. This dataset contains 8 categories and 12340 web search snippets.
- Reuters10T: Reuters-21578[1] is a dataset including news on Reuters newswire in 1987, and we just use the news titles as short texts. We choose 10 largest categories and remove the documents belonging to multiple categories.
- AGnews [Zhang *et al.*, 2015]: It is another news article dataset, and the titles of 4 largest categories are selected.
- TwitterTrends10[2]: It contains 140K tweets of 10 trending events from 2015 to 2017, and can verify the capability of our models in event classification on social media.
- News Category Dataset[3] [Misra, 2018]: 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, like college and education. We choose one 10-category sub-dataset HuffN10 and two 4-category sub-datasets, HuffN4-SWPB and HuffN4-ECMC, focusing on overlapping categories.

For all datasets, we at first lower and lemmatize all corpus words, remove stop words and then filter out the documents containing only one word. The categories and statistics of these datasets after pre-processing are shown in Table 1. N_u indicates the average document number of users and N_d indicates the average word number of documents.

4.2 Baselines

We evaluate our models against five baseline methods. They all deal with dataless text classification with seed words. The first two of which make use of word embeddings to compute category-word similarity, while the other three are based

[1] http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html
[2] https://www.kaggle.com/laiyin/trendtweet10
[3] https://www.kaggle.com/rmisra/news-category-dataset

Dataset	SeedBTM		WeSTClass		WeSTClass*		STM		SSCF	DescLDA
	S^L	S^D	S^L	S^D	S^L	S^D	S^L	S^D	S^D	S^D
SearchSnip	67.6	**83.4**	13.7	15.0	65.7	80.3	64.9	80.0	80.2	70.0
Reuters10T	39.2	**61.5**	22.5	16.9	33.4	38.1	37.6	54.9	57.9	59.1
AGnews	61.8	73.1	76.6	**76.7**	75.8	74.7	66.3	69.5	71.4	58.9
TwitterTrends10	43.0	**71.5**	9.3	68.5	40.6	49.8	9.1	62.9	67.8	57.3
HuffN10	54.4	**60.8**	37.1	58.0	39.4	38.3	10.1	53.9	53.3	37.6
HuffN4-SWPB	70.8	**71.5**	32.6	34.9	65.2	70.3	26.8	14.1	60.6	46.8
HuffN4-ECMC	51.7	**61.9**	30.0	35.4	55.6	54.9	41.3	53.2	54.5	39.0

Table 2: Macro-F1 (%) of SeedBTM and baselines on all datasets (The best results are highlighted in bold)

Dataset	SeedTBTM		WeSTClass		WeSTClass*		STM		SSCF	DescLDA
	S^L	S^D	S^L	S^D	S^L	S^D	S^L	S^D	S^D	S^D
TwitterTrends10	50.5	**71.9**	65.2	70.4	42.0	49.9	16.2	67.1	68.1	57.9
HuffN10	74.0	**75.3**	49.9	57.9	54.7	54.6	10.3	55.0	56.8	50.6
HuffN4-SWPB	81.9	**82.3**	77.4	76.3	65.9	71.9	13.0	13.3	70.9	62.8
HuffN4-ECMC	**79.1**	78.2	66.5	65.5	59.7	54.6	41.6	56.1	56.7	46.7

Table 3: Macro-F1 (%) of SeedTBTM and baselines on the datasets with user information (The best results are highlighted in bold)

on topic models to capture word co-occurrences, but none of them combine the effects of these two kinds of information.

- WeSTClass[4] [Meng *et al.*, 2018]: It is a state-of-the-art weakly-supervised text classification model. It utilizes seed words and word embeddings learnt from the corpus to generate pseudo-labeled documents for training a classifier, and then bootstraps it with unlabeled data.
- WeSTClass*: In order to make explicit evaluations for our models, we implement an intermediate method replacing the self-trained embeddings in WeSTClass with external and pre-trained word embeddings, Glove Common Crawler[5] [Pennington *et al.*, 2014], which is also used in our models.
- STM[6] [Li *et al.*, 2016b]: It is a successful approach to dataless text classification for long texts based on topic model, utilizing the co-occurrence information among seed words and corpus words.
- SSCF [Li *et al.*, 2019]: It is a state-of-the-art topic-based model for this task, specially aiming at short texts.
- DescLDA [Chen *et al.*, 2015b]: It is an extended model of LDA by constructing descriptive documents with seed words to guide the classification results.

For WeSTClass(*) and STM, we adopt their implementation codes directly. For SSCF and DescLDA, all parameters are tuned according to the suggestions provided by the authors. As to user information, SeedTBTM integrates it in the generative process, while for baseline models, the user name is regarded as a part of text contents (words).

4.3 Experiment Settings

Seed word settings. There are two ways (sources) to get seed words as problem input, the category label itself consisting of several words (usually only one word, denoted by S^L), or multiple descriptive words derived from them (denoted by S^D). The former type of seed words can be directly obtained from the corresponding dataset, while for the latter type, the selection strategy we adopt is similar to other work. We firstly run SeedBTM with S^L, and then manually choose for each category 3$\sim$10 suitable and representative words from the top 30 topic words. For the sake of fairness, our models share the same seed words (both S^L and S^D) for each dataset with all baseline methods, except that SSCF and DescLDA have only S^D type, which is consistent with their papers.

Parameter settings. For all datasets, we set the topic number $K' = K$ (category number), $\alpha = 50/K + 1$, $\beta = 0.1$, $\gamma = 1$ and $\epsilon = 0.0001$. We set the number of iterations to 50 as our models achieve competitive performance since then. For word embeddings, we employ the widely used GloVe Common Crawl as mentioned before. It contains 840B tokens, 2.2M vocab and 300d vectors.

4.4 Experimental Results

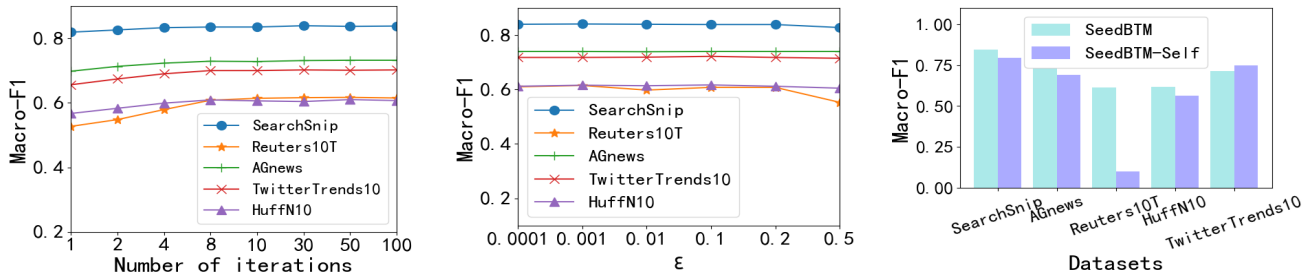
We at first evaluate the classification performances of SeedBTM using Macro-F1. We run every model 10 times on each dataset to get the average value shown in Table 2.

We can observe that SeedBTM performs significantly better than baseline models on 6 of 7 datasets except AGnews. Specifically, the Macro-F1 value increases about 1.2$\sim$6.3 percent compared to the best baseline result in these datasets. That indicates our model is able to enlarge the guiding role of seed words in short text classification with the help of easily accessible word embeddings. In addition, the improvement over WeSTClass* confirms that our model can utilize word embeddings more effectively through fusing word co-occurrence information in short texts. As an exception, our model underperforms WeSTClass and WeSTClass* on AGnews, since the categories in this dataset can be easily distinguished, and the word co-occurrence information is thus not so helpful. That explains our approach is more suitable for datasets with overlapping and interrelated categories. Moreover, The performances of baseline models fluctuate among

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

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

[6]https://github.com/WHUIR/STM



(a) (b) (c)

Figure 2: Macro-F1 of SeedBTM with varied parameters: (a) iteration numbers; (b) thresholds ϵ; (c) word embeddings

different datasets, but our model can get more stable and effective classification results, so behaves domain insensitive.

Similarly, we evaluate SeedTBTM with user information on TwitterTrends10 and the three sub-datasets from HuffPost. The average Macro-F1 results are shown in Table 3.

Among all methods, SeedTBTM stably achieves the best performance on all datasets. Also, SeedTBTM significantly increases Macro-F1 scores about 0.4∼16.3 percent compared to SeedBTM on the four common datasets. That indicates our approach can effectively utilize metadata like user information to improve the accuracy of dataless short text classification, which is impossible to realize for supervised methods without sufficient labeled data for each user.

For different types of seed words, although the results of our models with S^L is generally worse than those with S^D except the last dataset, the gap is distinctly reduced compared to baseline methods. That implies with fewer seed words, our approach can achieve comparable performances via combining the information of word co-occurrence and category-word similarity, which simplifies the subtle process of selecting descriptive words automatically or manually.

4.5 Parameter Study

We now study the impact of different parameter settings on the classification performance of SeedBTM by using S^D as seed words. When paying attention to one parameter, other parameters are fixed to the default values given in Section 4.3.

The impact of iteration number. This is an important factor for our models, because a smaller value means costing less time to obtain classification results. We vary the number in the range of [1,100], and the results are shown in Figure 2(a). We can see that SeedBTM can achieve good classification performances when the number of iterations is 10, and the F1-score is almost unchanged after that point. The fast convergence should give credit to the regulating effect of the prior knowledge from word embeddings.

The impact of threshold ϵ. We vary the lower bound for the word similarity in Equation 1 from 0.0001 to 0.5 and observe the influence on model performances shown in Figure 2(b). The flat lines certify that our approach is insensitive to this threshold, especially for smaller values less than 0.2.

The impact of word embeddings. In Table 2, WeSTClass outperforms WeSTClass* with S^D on TwitterTrends10 and HuffN10 datasets, but not for others. That indicates word embeddings trained from the corpus may be more suitable in some scenarios. Thus, we replace the external general word embeddings in SeedBTM with self-trained word embeddings and name the new model as SeedBTM-Self for comparison. From Figure 2(c), we find that SeedBTM-Self performs worse than SeedBTM on SearchSnip, AGnews and HuffN10, and is even significantly poor on Reuters10T, while on the larger dataset TwitterTrends10 with event-related documents, SeedBTM-Self is better than SeedBTM. The results demonstrate that when the categories are domain-specific, external word embeddings are probably insufficient to characterize ad-hoc word correlations, but self-trained word embeddings from plentiful documents can lead to better performances. Therefore, our models can choose suitable word embeddings depending on the scope and granularity of categories to be predicted.

5 Conclusion

Aiming at the task of dataless short text classification with seed words, we propose a novel approach integrating both word co-occurrence information and category-word similarity information, which are respectively brought by biterm-based topic model and widely used word embeddings. The proposed models significantly outperform other baseline methods, especially for difficult and confusing classification problems. Moreover, the flexibility on the base topic model can effectively incorporate metadata like users to further improve the classification accuracy, which is hard to realize for supervised methods due to the lack of specific labels.

In the future, we plan to extend this approach to deal with the hierarchical short text classification task. In addition, it is interesting to study how to optimize the embedding-based similarity computation and its integration with topic models.

Acknowledgements

This work was supported by the National Key R&D Program of China (2017YFC0803805, 2018YFC0116703), and the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDC02060500.

References

[Andrzejewski and Zhu, 2009] David Andrzejewski and Xiaojin Zhu. Latent Dirichlet allocation with topic-in-set knowledge. In *NAACL HLT Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 43–48. Association for Computational Linguistics, 2009.

[Blei *et al.*, 2003] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[Chang *et al.*, 2008] Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. Importance of semantic representation: Dataless classification. In *AAAI*, volume 2, pages 830–835, 2008.

[Chen *et al.*, 2015a] Weizheng Chen, Jinpeng Wang, Yan Zhang, Hongfei Yan, and Xiaoming Li. User based aggregation for biterm topic model. In *ACL*, volume 2 (Short Papers), pages 489–494, 2015.

[Chen *et al.*, 2015b] Xingyuan Chen, Yunqing Xia, Peng Jin, and John Carroll. Dataless text classification with descriptive LDA. In *AAAI*, pages 2224–2231, 2015.

[Druck *et al.*, 2008] Gregory Druck, Gideon Mann, and Andrew McCallum. Learning from labeled features using generalized expectation criteria. In *SIGIR*, pages 595–602. ACM, 2008.

[Hingmire and Chakraborti, 2014] Swapnil Hingmire and Sutanu Chakraborti. Topic labeled text classification: a weakly supervised approach. In *SIGIR*, pages 385–394. ACM, 2014.

[Jiang *et al.*, 2016] Lan Jiang, Hengyang Lu, Ming Xu, and Chongjun Wang. Biterm pseudo document topic model for short text. In *ICTAI*, pages 865–872. IEEE, 2016.

[Li *et al.*, 2016a] Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Topic modeling for short texts with auxiliary word embeddings. In *SIGIR*, pages 165–174. ACM, 2016.

[Li *et al.*, 2016b] Chenliang Li, Jian Xing, Aixin Sun, and Zongyang Ma. Effective document labeling with very few seed words: A topic model approach. In *CIKM*, pages 85–94. ACM, 2016.

[Li *et al.*, 2018a] Chenliang Li, Shiqian Chen, Jian Xing, Aixin Sun, and Zongyang Ma. Seed-guided topic model for document filtering and classification. *ACM Transactions on Information Systems (TOIS)*, 37(1):9:1–37, 2018.

[Li *et al.*, 2018b] Ximing Li, Ang Zhang, Changchun Li, Lantian Guo, Wenting Wang, and Jihong Ouyang. Relational biterm topic model: Short-text topic modeling using word embeddings. *The Computer Journal*, 62(3):359–372, 2018.

[Li *et al.*, 2019] Chenliang Li, Shiqian Chen, and Yan Qi. Filtering and classifying relevant short text with a few seed words. *Data and Information Management*, 3(3):165–186, 2019.

[Liu *et al.*, 2004] Bing Liu, Wee Sun Li, Xiaoli and, and Philip S Yu. Text classification by labeling words. In *AAAI*, volume 4, pages 425–430, 2004.

[Meng *et al.*, 2018] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised neural text classification. In *CIKM*, pages 983–992. ACM, 2018.

[Misra, 2018] Rishabh Misra. News category dataset, June 2018.

[Pennington *et al.*, 2014] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543. Association for Computational Linguistics, 2014.

[Quan *et al.*, 2015] Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. Short and sparse text topic modeling via self-aggregation. In *IJCAI*, pages 2270–2276, 2015.

[Shalaby and Zadrozny, 2019] Walid Shalaby and Wlodek Zadrozny. Learning concept embeddings for dataless classification via efficient bag-of-concepts densification. *Knowledge and Information Systems*, 61:1047–1070, 2019.

[Türker, 2019] Rima Türker. Knowledge-based dataless text categorization. In *ESWC*, pages 231–241. Springer, 2019.

[Wang *et al.*, 2017] Jin Wang, Zhongyuan Wang, Dawei Zhang, and Jun Yan. Combining knowledge with deep convolutional neural networks for short text classification. In *IJCAI*, pages 2915–2921, 2017.

[Weng *et al.*, 2010] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential Twitterers. In *WSDM*, pages 261–270. ACM, 2010.

[Xun *et al.*, 2016] Guangxu Xun, Vishrawas Gopalakrishnan, Fenglong Ma, Yaliang Li, Jing Gao, and Aidong Zhang. Topic discovery for short texts using word embeddings. In *ICDM*, pages 1299–1304. IEEE, 2016.

[Yan *et al.*, 2013] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *WWW*, pages 1445–1456. ACM, 2013.

[Yin and Wang, 2014] Jianhua Yin and Jianyong Wang. A Dirichlet multinomial mixture model-based approach for short text clustering. In *SIGKDD*, pages 233–242. ACM, 2014.

[Zeng *et al.*, 2018] Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R Lyu, and Irwin King. Topic memory networks for short text classification. In *EMNLP*, pages 3120–3131. Association for Computational Linguistics, 2018.

[Zhang *et al.*, 2015] Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NIPS*, pages 649–657, 2015.

[Zhao *et al.*, 2011] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing Twitter and traditional media using topic models. In *ECIR*, pages 338–349. Springer, 2011.

[Zuo *et al.*, 2016] Yuan Zuo, Jichang Zhao, and Ke Xu. Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, 48(2):379–398, 2016.