Dataless Text Classification: A Topic Modeling Approach with Document Manifold

Ximing Li
College of Computer Science and Technology, Jilin University, China
liximing86@gmail.com

Changchun Li
College of Computer Science and Technology, Jilin University, China
changchunli93@gmail.com

Jinjin Chi
College of Computer Science and Technology, Jilin University, China
chijinjin616@gmail.com

Jihong Ouyang
College of Computer Science and Technology, Jilin University, China
ouyj@jlu.edu.cn

Chenliang Li*
School of Cyber Science and Engineering, Wuhan University, China
cllee@whu.edu.cn

ABSTRACT

Recently, dataless text classification has attracted increasing attention. It trains a classifier using seed words of categories, rather than labeled documents that are expensive to obtain. However, a small set of seed words may provide very limited and noisy supervision information, because many documents contain no seed words or only irrelevant seed words. In this paper, we address these issues using document manifold, assuming that neighboring documents tend to be assigned to a same category label. Following this idea, we propose a novel Laplacian seed word topic model (LapSWTM). In LapSWTM, we model each document as a mixture of hidden category topics, each of which corresponds to a distinctive category. Also, we assume that neighboring documents tend to have similar category topic distributions. This is achieved by incorporating a manifold regularizer into the log-likelihood function of the model, and then maximizing this regularized objective. Experimental results show that our LapSWTM significantly outperforms the existing dataless text classification algorithms and is even competitive with supervised algorithms to some extent. More importantly, it performs extremely well when the seed words are scarce.

CCS CONCEPTS

• **Computing methodologies** → **Learning paradigms**; **Machine learning algorithms**; Supervised learning by classification; Classification and regression trees;

KEYWORDS

Topic Modeling, Dataless Text Classification, Seed Word, Document Manifold

*corresponding author

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3271671

ACM Reference Format:
Ximing Li, Changchun Li, Jinjin Chi, Jihong Ouyang, and Chenliang Li. 2018. Dataless Text Classification: A Topic Modeling Approach with Document Manifold. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3269206.3271671

1 INTRODUCTION

Text classification aims to train a classifier that can automatically assign test documents with pre-defined category labels. Traditional text classification algorithms are mainly developed in the paradigm of supervised learning, which requires a significant number of labeled documents. However, manually labeling documents is expensive and time-consuming, and worse of all, it hinders further analysis in a real-time manner where the fast response is a demanding need. This brings a significant challenge to automatic text classification.

Recently, dataless text classification with seed words, a popular solution to eliminate the human effort in collecting labeled documents, has delivered promising classification performance. This kind of algorithms can be trained on unlabeled documents with representative words (i.e., seed words) of categories. As discussed in [7], manually selecting seed words for categories is much cheaper than assigning category labels to documents. This process of seed word selection can be further enhanced with existing unsupervised techniques such as unsupervised topic modeling [5, 16]. In this sense, the dataless text classification with seed words has become a promising complement to supervised learning.

Previous studies have shown that the existing dataless text classifiers empirically perform well, and even achieve very close performance to supervised algorithms [4–9, 14–16, 19]. However, the effectiveness of these dataless classifiers is largely affected by the seed words. When only few seed words are available, the classification performance often deteriorates by a large margin. For example, the descriptive latent Dirichlet allocation model (DescLDA) [5] learns discriminative topic distributions using descriptive documents constructed by repeating seed words. It performs close to the support vector machines (SVMs) on the *Reuters* dataset, but its classification accuracy sharply drops from 0.8 to 0.3 when the seed words are scarce.

The underlying causes of this problem are that fewer seed words bring much less and noisy supervision information. That is, given a small set of seed words many documents contain no seed words or only irrelevant ones. Taking the popular *Reuters* dataset as an example, if we use the seed words selected only from the short category labels (i.e., about 1.1 seed words per label), there are about 62% documents containing no seed words and about 12% documents containing only irrelevant ones. In this situation, we could only capture very limited discriminative signals for training text classifiers.

In this paper, we aim to address these issues by exploiting local neighborhood structure, i.e., document manifold [2, 3, 12, 18]. We assume that highly similar documents tend to belong to a same category. By preserving the local neighborhood structure, the training documents can be linked, so that it spreads the supervision even if many of them contain no seed words and simultaneously modifies the noisy labeling information for the documents containing irrelevant seed words only. Following this idea, we propose a novel topic model-based dataless classification algorithm, namely Laplacian seed word topic model (LapSWTM). In LapSWTM, each document is modeled as a distribution of category topics, where each one corresponds to a category label. We describe the supervision by constructing document-specific Dirichlet priors for these category topic distributions using seed word occurrences. Specifically, we incorporate a manifold regularization term with respect to category topic distributions into the log-likelihood function of the model, so that neighboring documents tend to have similar category topic distributions. The generalized expectation maximization algorithm is used to optimize the regularized objective of LapSWTM. Experimental results on two real-world datasets demonstrate that our LapSWTM can significantly outperform the state-of-the-art dataless classification algorithms. More importantly, LapSWTM can achieve competitive classification performance with very few seed words.

For clarity, the contributions of this paper are summarized as follows:

- We develop a novel LapSWTM algorithm for dataless text classification with cheaper seed words, rather than labeled documents. In LapSWTM, we define two types of topics, i.e., category topic and background topic. Each category topic corresponds to a distinctive category label, and the background topic is used to capture the global semantic information over the whole corpus. To incorporate the supervision provided by the seed words, we construct document-specific Dirichlet priors for category topic distributions using seed word occurrences.
- Specifically, we exploit local neighborhood structure to simultaneously enrich discriminative classification signals and correct the noise. This is achieved by incorporating a manifold regularization term with respect to category topic distributions into the log-likelihood function of the model.
- We conduct a number of experiments to evaluate LapSWTM on two popular datasets, where one is balanced and the other is imbalanced. Both datasets have two pre-defined seed word sets, including a small set and a relatively larger one. Evaluation results show that our LapSWTM significantly outperforms the state-of-the-art dataless classification algorithms, especially when the seed words are scarce

Table 1: A summary of important notations

Notation	Description
D	number of documents
K	number of category topics
G	number of background topics
ϕ	category topic-word distribution
β	Dirichlet prior of ϕ
$\widehat{\phi}$	background topic-word distribution
$\widehat{\beta}$	Dirichlet prior of $\widehat{\phi}$
θ_d	category topic distribution for document d
α_d	document-specific Dirichlet prior of θ_d
$\widehat{\theta}_d$	background topic distribution for document d
$\widehat{\alpha}_d$	Dirichlet prior of $\widehat{\theta}_d$
δ	Bernoulli distribution: a topic type indicator

The rest of this paper is organized as follows: In Section 2, we introduce the proposed LapSWTM algorithm in detail. Section 3 shows the experimental results. In Section 4, we review recent related works. In Section 5, we conclude this work.

2 MODEL

In this section, we present the proposed Laplacian seed word topic model (LapSWTM) for dataless text classification with seed words.

2.1 LapSWTM Overall

We first describe the model structure of LapSWTM. The LapSWTM consists of two kinds of topics, i.e., category topic and background topic. Each category topic represents a specific category label[1], describing the semantics of the category; each background topic is used to describe the background semantics over the whole corpus. Besides, each document is associated with a distribution over category topics and a distribution over background topics. Each word token can be either generated by a category topic or a background topic, following the same process as latent Dirichlet allocation (LDA) [1].

Formally, LapSWTM consists of K category topic distributions ϕ over words, drawn from the Dirichlet prior β, and G background topic distributions $\widehat{\phi}$ over words, drawn from the Dirichlet prior $\widehat{\beta}$. For each document d, it draws a distribution θ_d over category topics from the document-specific Dirichlet prior α_d, and a distribution $\widehat{\theta}_d$ over background topics from the Dirichlet prior $\widehat{\alpha}$. For each word token, it draws a topic type indicator c_{dn} from a Bernoulli distribution δ_{dn}. If $c_{dn} = 1$, LapSWTM draws a category topic z_{dn} from θ_d, and then w_{dn} from $\phi_{z_{dn}}$; if $c_{dn} = 0$, it generates

[1] In LapSWTM, the number of category topics is equivalent to the number of categories.

w_{dn} from the distributions with respect to background topics. For clarity, some important notations are outlined in Table 1, and the generative process of LapSWTM is summarized as follows:

- For each category topic $k \in \{1, \cdots, K\}$
  - Draw a category topic-word distribution $\phi_k \sim \text{Dirichlet}(\beta)$
- For each background topic $g \in \{1, \cdots, G\}$
  - Draw a background topic-word distribution $\widehat{\phi}_g \sim \text{Dirichlet}(\widehat{\beta})$
- For each document $d \in \{1, \cdots, D\}$
  - Draw a category topic distribution $\theta_d \sim \text{Dirichlet}(\alpha_{\text{d}})$
  - Draw a background topic distribution $\widehat{\theta}_d \sim \text{Dirichlet}(\widehat{\alpha})$
  - For each word $w_{dn} \in \{1, \cdots, N_d\}$
    - Draw $c_{dn} \sim \text{Bernoulli}(\delta_{dn})$
    - If $c_{dn} = 1$
      - Draw $z_{dn} \sim \text{Multinomial}(\theta_d)$
      - Draw $w_{dn} \sim \text{Multinomial}(\phi_{z_{dn}})$
    - If $c_{dn} = 0$
      - Draw $\widehat{z}_{dn} \sim \text{Multinomial}(\widehat{\theta}_d)$
      - Draw $w_{dn} \sim \text{Multinomial}(\widehat{\phi}_{\widehat{z}_{dn}})$

Document-specific Dirichlet prior α_d. In LapSWTM, the hyper-parameter α_d is used to describe the membership degree prior of category topics. That is, we use this prior to incorporate the supervision provided by the seed words. Straightforwardly, we compute α_d by seed word occurrences, since seed words are representatives of categories (i.e., category topics). Given any document d, the value of prior α_{dk} is computed by:

$$\alpha_{dk} = \pi \frac{SF_{dk}}{\sum_{i=1}^{K} SF_{di}} + \alpha_0 \tag{1}$$

where SF_{dk} is the number of times that seed words of category k occur in document d; π is a concentration parameter; and α_0 is a smoothing parameter.

Bernoulli parameter δ_{dn}. The Bernoulli distribution makes a decision whether word w_{dn} is drawn from a category topic or a background topic. Here, we compute δ_{dn} by the dot product $\theta_d^T \tau_{w_{dn}}$, where each component $\tau_{w_{dn}k}$ denotes the relevance degree between word w_{dn} and category topic k. Inspired by a previous seed-guided topic model (STM) [14], for each word v we compute its relevance degree for category topic k as follows:

$$u_{vk} = \max\left(\frac{SC_{vk}}{\sum_{i=1}^{K} SC_{vi}} - \frac{1}{K}, 0\right)$$
$$u_{vk} \leftarrow \frac{u_{vk}}{\sum_{i=1}^{V} u_{ik}} \quad \tau_{vk} = \max\left(\frac{u_{vk}}{\sum_{i=1}^{K} u_{vi}}, \epsilon\right) \tag{2}$$

where V is the number of words; SC_{vk} denotes the number of co-occurrences between word v and seed words of category k; and the parameter ϵ is used to avoid zero. Following [14], we empirically set ϵ to 0.01.

2.2 Model Fitting with Document Manifold

In LapSWTM, the hidden variables include two document-topic distributions, i.e., θ and $\widehat{\theta}$, and two topic-word distributions, i.e., ϕ and $\widehat{\phi}$. Given a training dataset W, a straightforward model fitting method is to maximize the log-likelihood function with respect to these hidden variables of interest, given by:

$$\begin{aligned}\log p\left(W, \theta, \widehat{\theta}, \phi, \widehat{\phi} | \alpha, \widehat{\alpha}, \beta, \widehat{\beta}, \tau\right) = \\ + \sum_{k=1}^{K} \log p\left(\phi_k | \beta\right) + \sum_{g=1}^{G} \log p\left(\widehat{\phi}_g | \widehat{\beta}\right) + \sum_{d=1}^{D} \log p\left(\theta_d | \alpha_d\right) \\ + \sum_{d=1}^{D} \log p\left(\widehat{\theta}_d | \widehat{\alpha}\right) + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \log p\left(w_{dn} | \theta_d, \widehat{\theta}_d, \phi, \widehat{\phi}, \tau\right)\end{aligned} \tag{3}$$

Regularized log-likelihood objective. The supervision provided by the seed words is often limited and noisy. Many documents may not be supported by the provided seed words. This situation is further exacerbated when very few seed words are available. In LapSWTM, note that we use the document-specific prior α_d to incorporate the supervision information. Reviewing Eq.1, it provides uniform label priors (i.e., no supervision available) for documents containing no seed words, and noisy label priors for documents without relevant seed word occurrences.

To address these issues, we exploit the manifold regularization to preserve the local neighboring structure between documents. In the context of LapSWTM, it means that the category topic distributions of documents should be similar if they are nearest neighbors. To achieve this constraint, we use the following manifold regularization term:

$$\mathcal{R}(\theta) = \frac{1}{2} \sum_{k=1}^{K} \sum_{i,j=1}^{D} \left(\theta_{ik} - \theta_{jk}\right)^2 W_{ij} \tag{4}$$

The nearest neighboring weight is defined by:

$$W_{ij} = \begin{cases} 1 & \text{if } d_i \in \Pi(d_j) \text{ or } d_j \in \Pi(d_i) \\ 0 & \text{otherwise} \end{cases}$$

where $\Pi(d)$ denotes the set of R nearest neighbors of document d.

We combine the log-likelihood function of Eq.3 with the manifold regularizer of Eq.4, so that obtaining the final objective of LapSWTM:

$$\mathcal{L}(\theta, \widehat{\theta}, \phi, \widehat{\phi}) = \log p\left(W, \theta, \widehat{\theta}, \phi, \widehat{\phi} | \alpha, \widehat{\alpha}, \beta, \widehat{\beta}, \tau\right) - \lambda \mathcal{R}(\theta) \tag{5}$$

where $\lambda \geq 0$ is a regularization parameter.

2.2.1 Optimization. We learn the hidden variables $\{\theta, \widehat{\theta}, \phi, \widehat{\phi}\}$ by maximizing the objective of Eq.5, however, the maximization is intractable due to the regularization term. For efficient computation, the generalized expectation maximization (GEM) algorithm is used by following [3]. Overall, we first initialize $\{\theta, \widehat{\theta}, \phi, \widehat{\phi}\}$ randomly, and then iteratively perform the E-step (i.e., estimating topic assignments of word tokens) and M-step (i.e., updating $\{\theta, \widehat{\theta}, \phi, \widehat{\phi}\}$) until convergence. We now introduce optimization details.

E-step. The goal of the E-step is to estimate the probabilities, i.e., the posterior, of topic assignments of word tokens. Given the current hidden distributions, we can directly estimate the topic assignments by applying the Bayes rule:

$$p\left(z_{dn}=k|\theta_d,\phi,\tau_{w_{dn}}\right) = \theta_d^T\tau_{w_{dn}}\frac{\theta_{dk}\phi_{kw_{dn}}}{\sum_{i=1}^{K}\theta_{di}\phi_{iw_{dn}}} \triangleq N_{dnk} \quad (6)$$

$$p\left(\widehat{z}_{dn}=g|\theta_d,\widehat{\theta}_d,\widehat{\phi},\tau_{w_{dn}}\right) = \left(1-\theta_d^T\tau_{w_{dn}}\right)\frac{\widehat{\theta}_{dg}\widehat{\phi}_{gw_{dn}}}{\sum_{i=1}^{G}\widehat{\theta}_{di}\phi_{iw_{dn}}} \triangleq \widehat{N}_{dng} \quad (7)$$

For convenience, we consider the posteriors as the soft numbers of topic assignments. That is, N_{dnk} and $\widehat{N}_{dng}$ are used to denote the soft numbers of w_{dn} assigned to category topic k and background topic g, respectively.

M-step. The goal of the M-step is to update $\{\theta,\widehat{\theta},\phi,\widehat{\phi}\}$ given soft numbers of topic assignments N_{dnk} and $\widehat{N}_{dng}$ obtained in the previous E-step. The distributions $\widehat{\theta}$, ϕ and $\widehat{\phi}$ without manifold constraints can be individually updated by using Bayesian estimation, i.e., the expectations of their posteriors:

$$\widehat{\theta}_{dg} = \frac{\widehat{N}_{dg}+\widehat{\alpha}}{\sum_{i=1}^{G}\widehat{N}_{di}+G\widehat{\alpha}} \quad (8)$$

$$\phi_{kv} = \frac{N_{kv}+\beta}{\sum_{i=1}^{V}N_{ki}+V\beta} \quad (9)$$

$$\widehat{\phi}_{gv} = \frac{\widehat{N}_{gv}+\widehat{\beta}}{\sum_{i=1}^{V}\widehat{N}_{gi}+V\widehat{\beta}} \quad (10)$$

where $\widehat{N}_{dg}=\sum_{n=1}^{n=N_d}\widehat{N}_{dng}$ is the soft number of words assigned to background topic g in document d; $N_{kv}=\sum_{d=1}^{d=D}\sum_{w_{dn}=v}N_{dnk}$ and $\widehat{N}_{gv}=\sum_{d=1}^{d=D}\sum_{w_{dn}=v}\widehat{N}_{dng}$ are the soft numbers of word v assigned to category topic k and background topic g, respectively.

To update θ, we perform a Newton-Raphson iteration over the manifold regularization term $\mathcal{R}(\theta)$ until the overall objective of Eq.5 decreases. After some simple derivations, the final update equation of θ is given by:

$$\theta_{dk} \leftarrow (1-\gamma)\theta_{dk}+\gamma\frac{\sum_{i=1}^{D}\theta_{ik}W_{di}}{\sum_{i=1}^{D}W_{di}} \quad (11)$$

where γ is the learning rate of the Newton-Raphson iteration, such that $0\le\gamma\le 1$. In this Newton-Raphson iteration, the θ is initialised by the expectation of the current posterior:

$$\theta_{dk} = \frac{N_{dk}+\alpha_{dk}}{\sum_{i=1}^{K}(N_{di}+\alpha_{di})} \quad (12)$$

where $N_{dk}=\sum_{n=1}^{n=N_d}N_{dnk}$ is the soft number of words assigned to category topic k in document d.

For clarity, we summarize the GEM optimization of LapSWTM in the following *Algorithm 1*.

Algorithm 1 GEM for LapSWTM

```
Initialize parameters α̂, β, β̂, π, λ, R and γ
Initialize hidden variables θ, θ̂, phi and φ̂ randomly
Compute α and τ using Eqs.1 and 2
Find the nearest neighboring documents
For t = 1, 2, ..., MaxIter
    E-step: Update all N_dnk and N̂_dng using Eqs.6 and 7
    M-step:
        Update θ̂, φ and φ̂ using Eqs.8, 9 and 10
        Update θ using Eq.12
        Copy θ to θ'
        While L(θ) ≤ L(θ') Do
            Copy θ' to θ
            Update θ' using Eq.11
        End While
End for
```

Remark. Note that γ can be considered as a tuning parameter to control the importance of neighboring documents during Newton-Raphson iterations. When $\gamma=0$, the manifold regularization does not work.

2.2.2 Finding the Nearest Neighbors. Before performing LapSWTM, we need to find the R nearest neighbors for all training documents. In this work, for each document pair $\{d_i,d_j\}$ we measure their distance by a mixture of the cosine of their category topic priors and the cosine of their TFIDF (denoted by Ω) representations:

$$\text{Distance}(d_i,d_j) = cos\left(\alpha_{d_i},\alpha_{d_j}\right)+cos\left(\Omega_{d_i},\Omega_{d_j}\right) \quad (13)$$

Besides, note that computing the distances of all document pairs is often very expensive, requiring $O(D^2)$ time. To ameliorate this, for each document we search its nearest neighbors on a randomly selected subset of the training corpus. In this work, the subset size is empirically set to 2000.

2.2.3 Predicting Test Documents. Since the category topic and category label have an one-to-one correspondence in LapSWTM, for test documents we estimate the category topic distributions θ and predict them using the dominating category topic:

$$y_{d_{test}} = \underset{k\in 1,\cdots,K}{\text{argmax}}\ \theta_{d_{test}} \quad (14)$$

3 EXPERIMENT

In this section, we empirically evaluate the effectiveness of the proposed LapSWTM.

3.1 Experimental Setup

Dataset In the experiment, we use two widely used datasets in classification evaluations, including the balanced *Newsgroup*[2] and the imbalanced *Reuters*[3].

[2] http://qwone.com/~jason/20Newsgroups/
[3] http://kdd.ics.uci.edu/database/reuters21578/reuters21578.html

Table 2: Statistics of datasets, classification tasks and seed word sets. #*Train* / #*Test*: number of training/test documents; #*AvgD*: average length of documents; *K*: number of categories; *V*: number of word types; S^L/S^D: average number of seed words in S^L/S^D.

Dataset	Classification task	#*Train*	#*Test*	#*AvgD*	*K*	*V*	S^L	S^D
Newsgroup	med-space	1187	790	153.7	2	21391	1.5	5.5
	pc-mac	1168	777	98.3	2	14798	4	5
	politics-religion	3031	2028	196.4	2	31358	4	12
	politics-sci	3948	2629	174.6	2	37117	4	17
	comp-religion-sci	6765	4502	153.4	3	44048	7.3	19.7
	politics-rec-religion-sci	7793	5187	158.5	4	46187	4.25	15.5
	autos-motorcycles-baseball-hockey	2389	1950	122.2	4	25339	1	4
	All categories	11314	7532	152.1	20	52761	1.5	4.75
Reuters	All categories	5228	2057	70.8	10	7419	1.1	6

The *Newsgroup* dataset contains 18846 documents in 20 categories, 11314 documents for training and 7532 documents for testing. To conduct more evaluations, we further create seven subset datasets of *Newsgroup*, leading to seven supplement classification tasks as previous evaluations in [8, 14].

The *Reuters* dataset consists of 7285 documents in 10 categories. We use the standard train/test split, 5228 documents for training and 2057 documents for testing.

For all datasets, we remove the stopwords, and the words that are shorter than 2 characters or occur in less than 5 documents.

Seed word set. For all datasets, we use two pre-defined seed word sets, namely S^L and S^D. The S^L contains words from category labels by manually removing the unrelated words. The S^D are manually selected from a candidate set, learned by unsupervised algorithms, such as the LDA topic model [5, 14]. Comparing the two sets, the S^L is much smaller, whose average numbers of seed words are almost less than 4.

The statistics of the datasets and seed word sets are outlined in Table 2.

Baseline algorithm. In the experiment, five baseline algorithms are compared, two dataless algorithms and three supervised algorithms. The descriptions and parameter settings of baselines are presented as below.

- **DescLDA** [5] is an extension of LDA, which learns discriminative topic distributions using descriptive documents constructed by repeating seed words. DescLDA requires that each seed word is only associated with a single topic. For each overlapping seed words in S^L and S^D, we thus randomly left them to one of their categories. We run DescLDA with different topic numbers over the set $\{20, 30, \cdots, 80\}$, and report the best results in all tasks.
- **STM** [14] is the state-of-the-art topic model based dataless classification algorithm. We use the code[4] provided by its authors. Following the original paper, we tune its crucial parameters, such as ρ, and then report the best results in all tasks.
- **sLDA** [17] is a supervised extension of LDA by incorporating a response variable. We use the code[5] provided by its authors. We also train sLDA with topic numbers over $\{20, 30, \cdots, 80\}$, and report the best results in all tasks.
- **Naive Bayes (NB)** is a supervised classifier based on the Bayes rule. We use the TFIDF representations of documents and train it using the *sklearn* tool[6].
- **SVMs** is a traditional supervised classification algorithm. We also use the TFIDF representations of documents and employ the *sklearn* tool to train SVMs with the default parameters.

In terms of all algorithms, we report the average results of 10 independent runs. In terms of dataless algorithms, we train them on a mixture of the training and test documents, following the procedure used in [14].

Besides, the parameters of our LapSWTM are empirically set as follows: $\pi = 0.1$, $\alpha_0 = \widehat{\alpha} = 0.1$, $\beta = \widehat{\beta} = 0.1$, $G = 1$, $R = 5$ and $\gamma = 0.1$. We will present empirical evaluation results of three crucial parameters in the following Section 3.4.

Evaluation Metric. We examine the classification performance by Micro-F1 and Macro-F1.

[4] https://github.com/ly233/Seed-Guided-Topic-Model
[5] http://www.cs.cmu.edu/~chongw/slda/
[6] http://scikit-learn.org/stable/

The F1 score is actually a combination of the recall and precision:

$$\text{Recall} = \frac{TP}{\#NP}$$
$$\text{Precision} = \frac{TP}{\#NPP}$$
$$\text{F1} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} \quad (15)$$

where *TP* is the number of true positive predictions; *#NP* is the number of positive documents; and *#NPP* is the number of positive predictions.

Specifically, Micro-F1 and Macro-F1 are the F1 score of the whole dataset and the average F1 score of all categories, respectively.

3.2 Classification Performance

Tables 3 and 4 report the classification performance of 6 methods over all classification tasks. The observations made between baseline algorithms and the proposed LapSWTM are given as below.

Comparing with dataless algorithms. First, we present the evaluation results of *Newsgroup* and *Reuters* with all categories. Overall, we can observe that LapSWTM consistently performs better than the two dataless classification algorithms on Micro-F1 and Macro-F1. Compared with STM, the performance gain of Micro-F1 is about 0.03, and that of Macro-F1 is about 0.02∼0.04. Note that the main difference between them is that LapSWTM uses the manifold regularization. We thus argue that the improvement over STM directly indicates that preserving the local neighborhood structure is beneficial to enhance the classification performance. Compared with DescLDA, the scores of LapSWTM are significantly higher, i.e., almost over 0.1 improvements.

Second, we turn to the classification tasks of *Newsgroup* subsets. We can see that LapSWTM performs the best on 6 out of 7 *Newsgroup* subset tasks. The F1 scores of LapSWTM are about 0.01 higher than those of STM. The performance gain is not very significant, since these *Newsgroup* subsets contain less categories and more seed words. making the classification tasks relatively "easier". Surprisingly, we observe that LapSWTM significantly performs better than DescLDA, and the gain is even about 0.3 in some settings. This result indicates that LapSWTM is much more robust than DescLDA.

From the perspective of the seed word set, we note that the proposed LapSWTM algorithm also delivers superior performance, i.e., LapSWTM+S^L>STM+S^L>DescLDA+S^L and the same for S^D. It is worthwhile to note that DescLDA obtains very poor performance when S^L is used. This observation is consistent with the results reported in [5, 14]. The possible reason is that DescLDA constructs descriptive documents by repeating seed words, however too few seed words in S^L make them less discriminative. In other words, we argue that DescLDA is very sensitive with the number of seed words. In contrast, STM and LapSWTM perform much more stable on the two seed word sets. The performance gap between S^L and S^D is only about 0.04∼0.08 on both datasets with all categories, and even only 0.001∼0.04 on *Newsgroup* subsets. They work well on S^L with very limited seed words. Being different from DescLDA, STM and LapSWTM incorporate the supervision of seed words during model definition. This may be a better choice for topic model-based dataless classifiers. Besides, we can see that the scores of S^D are consistently higher than those of S^L. This indicates that increasing the seed word size can stably improve classification results just like we thought it would.

Comparing with supervised algorithms. LapSWTM is even on a par with the supervised classification algorithms. Specially, LapSWTM+S^D outperforms sLDA in all settings and NB on 7 out of 9 classification tasks, e.g., *Newsgroup* with all categories, and slightly worse than SVMs. In terms of sLDA, the scores of LapSWTM+S^D are significantly higher, e.g., 0.13 higher of Micro-F1 on *Newsgroup* with all categories and 0.13 higher of Macro-F1 on *Reuters*. Besides, even LapSWTM+S^L obtains better scores than sLDA. LapSWTM+S^D is quite competitive with NB, where it achieves 0.03 higher Micro-F1 and Macro-F1 scores than NB on *Newsgroup* with all categories. Compared with these supervised algorithms, LapSWTM is driven by a few seed words for each category only, which are much cheaper to obtain. We thus argue that LapSWTM is a practical complementary choice for real world applications.

3.3 Analysis of LapSWTM

Note that the test documents may also contain seed words. For LapSWTM, this means that seed words directly provide prior guidance to the test documents, indicated by the document-specific category topic prior (ref. Eq.1). We now empirically analyze how the seed word occurrences in the test documents affect the prediction results.

In this evaluation, we employ *Newsgroup* and *Reuters* of all categories, and divide both of test datasets into four disjoint subsets: the test documents (1) containing no seed words (*NonSW*), (2) containing the seed words of the true category only (*OnlyTSW*), (3) containing not only the seed words of the true category but also the ones of false categories (*TFSW*), (4) containing seed words of false categories only (*OnlyFSW*). For both datasets, we train LapSWTM on all documents, and examine the classification results of these subsets individually.

We present the Micro-F1 scores in Table 5. Overall speaking, we can observe that the Micro-F1 performance rank is roughly given by *OnlyTSW*>*NonSW*>*TFSW*>*OnlyFSW*. Some detailed observations are also made. First, the Micro-F1 scores of *OnlyTSW* are significantly higher than those of other subsets in most settings, and even very close to 1 when S^D is used. This result is reasonable, because the test documents in *OnlyTSW* can be considered to be marked with the true category label by seed words. Second, it is interesting that the results of *NonSW* are better than those of *TFSW*. This suggests that noisy information provided based on the seed word occurrences adversely affects the final classification accuracy. Finally, we can see that the Micro-F1 scores of *OnlyFSW* are the worst, but all scores are achieved around 0.6. The result indicates that our LapSWTM can effectively classify the test documents even with false priors only.

3.4 Evaluation of Parameters

We empirically evaluate three crucial parameters of LapSWTM, including π, G and R. Evaluation results on *Newsgroup* and *Reuters* of all categories are shown in this subsection.

Table 3: The results of Micro-F1. The best results of dataless classifiers are highlighted in boldface. "‡" means that the gain of LapSWTM+S^D is statistically significant at 0.01 level.

Dataset	Classification task	LapSWTM		DescLDA		STM		sLDA	NB	SVMs
		S^L	S^D	S^L	S^D	S^L	S^D			
Newsgroup	med-space	0.979	**0.980**	0.872‡	0.969‡	0.971	0.972	0.909‡	0.974	0.979
	pc-mac	0.901	**0.937**	0.732‡	0.884‡	0.889‡	0.933	0.725‡	0.924	0.940
	politics-religion	0.912	**0.952**	0.881‡	0.886‡	0.903‡	0.950	0.905‡	0.952	0.963
	politics-sci	0.962	**0.967**	0.738‡	0.849‡	0.961	0.962	0.919‡	0.960	0.969
	comp-religion-sci	0.923	**0.927**	0.577‡	0.637‡	0.922	0.926	0.899‡	0.928	0.943
	politics-rec-religion-sci	0.915	**0.949**	0.634‡	0.776‡	0.913‡	0.944	0.839‡	0.946	0.948
	autos-motorcycles-baseball-hockey	0.947	0.961	0.576‡	0.782‡	0.918‡	**0.971**	0.882‡	0.960	0.968
	-All categories	0.748	**0.810**	0.647‡	0.694‡	0.703‡	0.775‡	0.684‡	0.775	0.823
Reuters	All categories	0.891	**0.931**	0.352‡	0.821‡	0.862‡	0.905‡	0.775‡	0.932	0.973

Table 4: The results of Macro-F1. The best results of dataless classifiers are highlighted in boldface. "‡" means that the gain of LapSWTM+S^D is statistically significant at 0.01 level.

Dataset	Classification task	LapSWTM		DescLDA		STM		sLDA	NB	SVMs
		S^L	S^D	S^L	S^D	S^L	S^D			
Newsgroup	med-space	0.979	**0.980**	0.863‡	0.971	0.961‡	0.967	0.901‡	0.968	0.976
	pc-mac	0.902	**0.944**	0.711‡	0.723‡	0.898‡	0.942	0.724‡	0.911‡	0.931
	politics-religion	0.911	**0.951**	0.873‡	0.879‡	0.901‡	**0.951**	0.901‡	0.949	0.961
	politics-sci	0.960	**0.967**	0.723‡	0.835‡	0.960	0.962	0.911‡	0.958	0.967
	comp-religion-sci	0.926	**0.927**	0.576‡	0.612‡	0.919	0.926	0.897‡	0.925	0.941
	politics-rec-religion-sci	0.928	**0.944**	0.625‡	0.763‡	0.914‡	0.941	0.833‡	0.946	0.944
	autos-motorcycles-baseball-hockey	0.945	0.962	0.551‡	0.752‡	0.916‡	**0.973**	0.885‡	0.958	0.964
	All categories	0.693	**0.783**	0.625‡	0.671‡	0.674‡	0.741‡	0.641‡	0.757	0.816
Reuters	All categories	0.799	**0.877**	0.332‡	0.809‡	0.751‡	0.832‡	0.743‡	0.903	0.939

Table 5: The Micro-F1 results of test subsets with respect to seed word occurrences

Seed word set	*Newsgroup*				*Reuters*			
	NonSW	*OnlyTSW*	*TFSW*	*OnlyFSW*	*NonSW*	*OnlyTSW*	*TFSW*	*OnlyFSW*
S^L	0.800	0.907	0.578	0.686	0.913	0.892	0.853	0.667
S^D	0.860	0.976	0.776	0.550	0.884	0.997	0.910	0.604

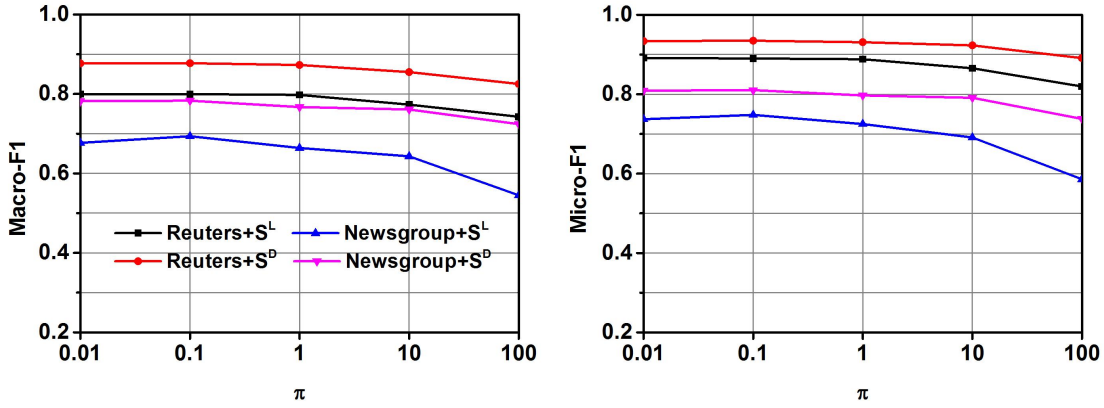

Figure 1: Evaluation results of different π values

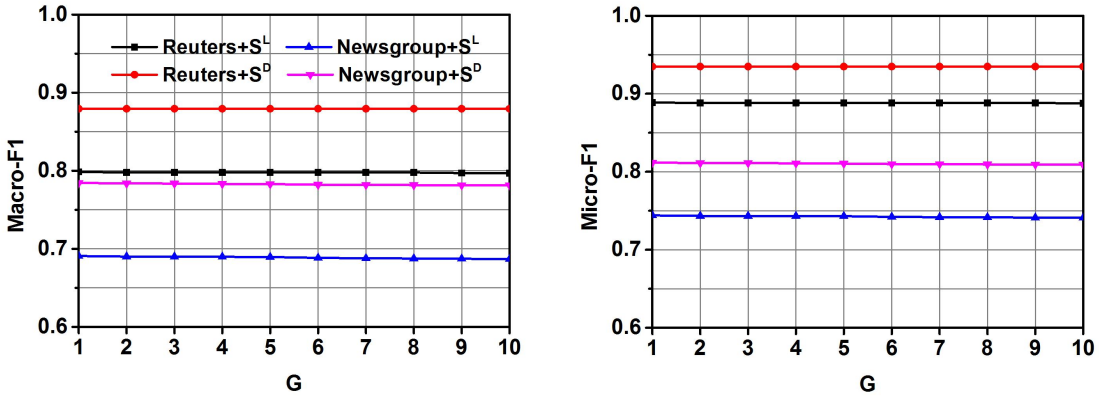

Figure 2: Evaluation results of different G values

Figure 1 presents the evaluation results of different values of π over the set $\{0.01, 0.1, \cdots, 100\}$. We can observe that LapSWTM performs better when π is relatively smaller. In LapSWTM, this concentration parameter π is used to control the order of magnitude of the document-specific category topic prior α_d. The larger α_d will strongly restrict the estimation of category topic distribution θ_d. However, α_d is often quite noisy, since training documents may contain seed words from the "wrong" categories (reviewing Eq.1). We argue that is why LapSWTM performs better with smaller π values. In the experiments, we set π to 0.1.

The parameter G denotes the number of background topics of LapSWTM. Figure 2 plots the evaluation results of G values from 1 to 10. We can observe that the performance gap between different G values is very small in all settings. That is to say, LapSWTM is insensitive to the background topic number, making the model more robust in real-world applications. We set G to 1 in the experiments.

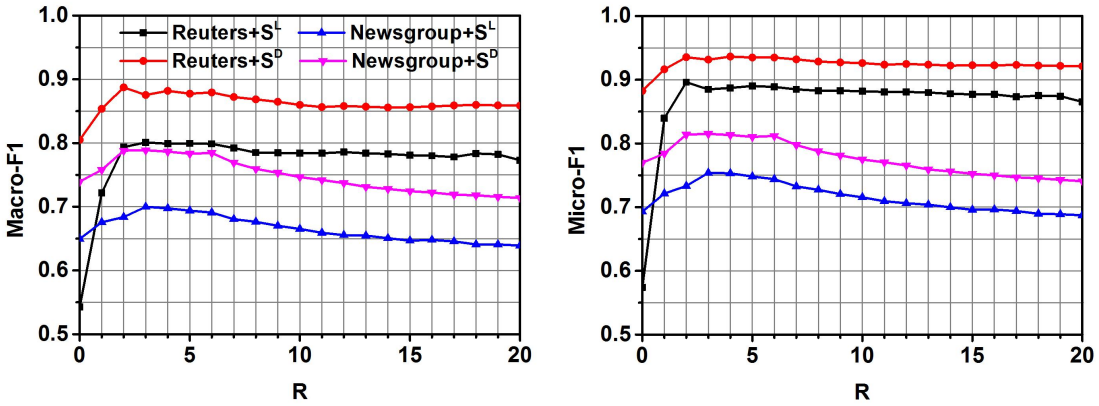

Figure 3: Evaluation results of different R values

The parameter R specifies the number of nearest neighbors used in the document manifold regularization. We examine the classification performance of different R values over the set $\{0, 1, \cdots, 20\}$. As the results shown in Figure 3, we have two observations. First, the scores of R=0 are always lower than those of other R values, especially for *Reuters* with S^L. Note that R=0 implies that the manifold regularization does not utilized for classification. We thus conclude that using manifold regularization can effectively improve the classification performance in dataless learning manner. Second, the best scores are often achieved when R=2,3,4,5,6. This indicates that fewer nearest document neighbors are enough to enrich the supervision information. Besides, the performance roughly goes down as R becomes larger. The possible reason is that more nearest neighbors may link many unrelated documents of different categories, i.e., incurring additional noise to the estimations of θ.

4 RELATED WORK

In this section, we review some recent related works on dataless text classification and topic models with document manifold.

4.1 Dataless Text Classification

Recently, dataless text classification has attracted much more attention and many dataless classifiers have been proposed [4–9, 13–16, 19]

A straightforward methodology of dataless classification is to automatically create labeled datasets for training. The work [13] presents a bootstrapping framework to generate noisy training dataset. It employs category labels and title words as key words, and enriches them by finding the most similar words. These selected key words are used to bootstrap context clusters, which are feed into a naive Bayes classifier as labeled training dataset. The seed word naive Bayes classifier (SNB) [16] finds representative words (i.e., seed words) of categories by words' information gains, which are computed on the k-means clustering results. The seed words constitute pseudo representative documents for categories, and these pseudo ones further constitute a noisy training dataset based on the similarity. A naive Bayes classifier is then trained on the noisy training dataset. SNB repeats these steps until convergence. Another way of dataless classification is to exploit auxiliary knowledge bases. For example, the algorithm in [4] maps the words and documents into a same semantic space of Wikipedia concepts, and measures the similarity between a document and a category (i.e., words occurred in the category label) using explicit semantic analysis. However, such algorithms may not be applicable to many real applications, since the knowledge bases are not always consistent with the current dataset. In contrast, our LapSWTM does not require any auxiliary knowledge base.

Specially, several topic model-based dataless algorithms have been developed in recent years. ClassifyLDA [9] employs annotators to assign labels to topics learned by the standard LDA over the unlabeled dataset. The topics marked with the same label are then aggregated into a single one, and LDA is trained using these aggregated topics for prediction. The topic label classification model [8] extends ClassifyLDA by allowing each hidden LDA topic to be marked with more than one label. STM [14] assumes that each document is drawn from a mixture of a single category-topic and all general-topics. It employs the seed word co-occurrence information to computes category word probability, which is helpful for classification. In the proposed LapSWTM, we also define two types of topics, i.e., category topic and background topic, but each document is allowed to be drawn from all K category topics. We argue that LapSWTM is more flexible and applicable to multi-class learning. Besides, LapSWTM uses the manifold regularization to preserve local neighboring structure, enriching the supervision besides the seed words.

4.2 Topic Models with Document Manifold

To our knowledge, the manifold regularization methods have been widely used to improve the standard topic models in unsupervised and semi-supervised manners [2, 3, 5, 11, 12, 18]. The algorithms in [2, 3] extend probabilistic latent semantic indexing (PLSI) [10] by incorporating manifold regularization terms with Euclidean distance and Kullback-Leibler divergence. They empirically outperform the standard PLSI model, especially for the clustering task. The authors of [11] proposed an semi-supervised extension of the

maximum entropy discrimination LDA [20] with a manifold regularizer. Compared with these models, LapSWTM can be directly applied to classification tasks without using any labeled document.

5 CONCLUSION

In this paper, we develop a topic model based dataless classification algorithm with manifold regularization, named LapSWTM. The motivation is to spread the supervision of seed words by preserving local neighboring structure, leading to better classification performance. To achieve this, we incorporate a manifold regularization term into the log-likelihood function of the model. This regularized objective is optimized by the generalized expectation maximization algorithm.

We conduct a number of experiments on two datasets. Experimental results show that LapSWTM consistently performs better than the existing dataless classification algorithms, and is on a par with supervised algorithms in some settings. More importantly, LapSWTM can achieve competitive performance with small sets of seed words.

ACKNOWLEDGMENT

We would like to acknowledge support for these projects from the National Natural Science Foundation of China (NSFC) [No.61602204, No.61502344, No.61472157].

REFERENCES

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research* 3 (2003), 993–1022.
[2] Deng Cai, Qiaozhu Mei, Jiawei Han, and Chengxiang Zhai. 2009. Probabilistic Dyadic Data Analysis with Local and Global Consistency. In *ACM Conference on Information and Knowledge Management*. 911–920.
[3] Deng Cai, Xuanhui Wang, and Xiaofei He. 2008. Modeling hidden topics on document manifold. In *International Conference on Machine Learning*. 105–112.
[4] Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: dataless classification. In *AAAI Conference on Artificial Intelligence*. 830–835.
[5] Xingyuan Chen, Yunqing Xia, Peng Jin1, and John Carroll. 2015. Dataless text classification with descriptive LDA. In *AAAI Conference on Artificial Intelligence*. 2224–2231.
[6] Doug Downey and Oren Etzioni. 2008. Look ma, no hands: analyzing the monotonic feature abstraction for text classification. In *Neural Information Processing Systems*. 393–400.
[7] Gregory Druck, Gideon Mann, and Andrew McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 595–602.
[8] Swapnil Hingmire and Sutanu Chakraborti. 2014. Topic labeled text classification: A weakly supervised approach. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 385–394.
[9] Swapnil Hingmire, Sandeep Chougule, and Girish K. Palshikar. 2013. Document classification by topic labeling. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 877–880.
[10] Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*. 50–57.
[11] Wenbo Hu, Jun Zhu, Hang Su, Jingwei Zhuo, and Bo Zhang. 2017. Semi-supervised max-margin topic model with manifold posterior regularization. In *International Joint Conference on Artificial Intelligence*. 1865–1871.
[12] Seungil Huh and Stephen E. Fienberg. 2010. Discriminative Topic Modeling based on Manifold Learning. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 653–662.
[13] Youngjoong Ko and Jungyun Seo. 2004. Learning with Unlabeled Data for Text Categorization Using Bootstrapping and Feature Projection Techniques. In *Annual Meeting on Association for Computational Linguistics*.
[14] Chenliang Li, Jian Xing, Aixin Sun, and Zongyang Ma. 2016. Effective document labeling with very few seed words: a topic modeling approach. In *ACM International on Conference on Information and Knowledge Management*. 85–94.
[15] Ximing Li and Bo Yang. 2018. A Pseudo Label based Dataless Naive Bayes Algorithm for Text Classification with Seed Words. In *International Conference on Computational Linguistics*. 1908–1917.
[16] Bing Liu, Xiaoli Li, Wee Sun Lee, , and Philip S. Yu. 2004. Text classification by labeling words. In *AAAI Conference on Artificial Intelligence*. 425–430.
[17] Jon D. Mcauliffe and David M. Blei. 2007. Supervised topic models. In *Neural Information Processing Systems*. 121–128.
[18] Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. 2008. Topic modeling with network regularization. In *international conference on World Wide Web*. 101–110.
[19] Daochen Zha and Chenliang Li. 2017. Multi-label Dataless Text Classification with Topic Modeling. *arXiv:1711.01563* (2017).
[20] Jun Zhu, Amr Ahmed, and Eric P. Xing. 2012. MedLDA: maximum margin supervised topic models. *Journal of Machine Learning Research* 13 (2012), 2237–2278.