Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

Dataless Text Classification with Descriptive LDA

Xingyuan Chen[1], **Yunqing Xia**[2], **Peng Jin**[1*] **and John Carroll**[3]
[1]School of Computer Science, Leshan Normal University, Leshan 614000, China
cxyforpaper@gmail.com, jandp@pku.edu.cn
[2]Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China
yqxia@tsinghua.edu.cn
[3]Department of Informatics, University of Sussex, Brighton BN1 9QJ, UK
j.a.carroll@sussex.ac.uk

Abstract

Manually labeling documents for training a text classifier is expensive and time-consuming. Moreover, a classifier trained on labeled documents may suffer from overfitting and adaptability problems. Dataless text classification (DLTC) has been proposed as a solution to these problems, since it does not require labeled documents. Previous research in DLTC has used explicit semantic analysis of Wikipedia content to measure semantic distance between documents, which is in turn used to classify test documents based on nearest neighbours. The semantic-based DLTC method has a major drawback in that it relies on a large-scale, finely-compiled semantic knowledge base, which is difficult to obtain in many scenarios. In this paper we propose a novel kind of model, descriptive LDA (DescLDA), which performs DLTC with only category description words and unlabeled documents. In DescLDA, the LDA model is assembled with a *describing device* to infer Dirichlet priors from prior descriptive documents created with category description words. The Dirichlet priors are then used by LDA to induce category-aware latent topics from unlabeled documents. Experimental results with the 20*Newsgroups* and *RCV1* datasets show that: (1) our DLTC method is more effective than the semantic-based DLTC baseline method; and (2) the accuracy of our DLTC method is very close to state-of-the-art supervised text classification methods. As neither external knowledge resources nor labeled documents are required, our DLTC method is applicable to a wider range of scenarios.

Introduction

A typical procedure for creating a machine learning-based classifier is: (1) human experts define categories, which are usually represented by category labels and sometimes also category descriptions; (2) human experts manually assign labels to training documents selected from the problem domain; (3) a classifier is automatically trained on the labeled documents; and (4) the classifier is applied to unlabeled documents to predict category labels. Supervision is provided by human experts in steps (1) and (2). In (1), the supervision is represented by the category labels/descriptions. As the human experts understand the classification problem well, it is not difficult for them to perform this step. In step (2), the supervision is represented by the labeled documents, which is labor-intensive. Moreover, a classifier trained on a limited number of labeled documents in a specific domain usually suffers from challenging problems such as overfitting (Cawley and Talbot 2010) and adaptability (Bruzzone and Marconcini 2010).

Research efforts have been made to **reduce** the effort required in step (2). For example, semi-supervised learning (Nigam et al. 2000; Blum and Mitchell 1998) trains on a small number of labeled documents and a larger number of unlabeled documents. Weakly-supervised learning methods (Liu et al. 2004; Hingmire and Chakraborti 2014) use either labeled words or latent topics that can control each class to retrieve relevant documents as initial training data. A drawback is that labeled documents are still required.

Recent research efforts have attempted to **eliminate** the labor in step (2). For example, dataless text classification (DLTC) (Chang et al. 2008) addresses the classification problem using only category label/description as supervision. In one approach (Gabrilovich and Markovitch 2007), a semantic similarity distance between documents is calculated based on Wikipedia. Documents are assigned category labels according to semantic distance using the nearest neighbors algorithm. As no labeled documents are required, human effort is saved, which makes the DLTC method very attractive. However, a drawback of such approaches is that they rely on a large-scale semantic knowledge base, which does not exist for many languages or domains.

In this paper, we propose a dataless text classification model called *descriptive LDA* (DescLDA), which incorporates topic modeling. In DescLDA, a *describing device* (DD), is joined to the standard LDA model to infer descriptive Dirichlet priors (i.e., a topic-word matrix) from a few documents created from descriptive words in category labels/descriptions. These priors can then influence the generation process, making the standard LDA capable of inferring topics for text classification. Compared to existing DLTC models (Chang et al. 2008), DescLDA does not require any external resources, which makes DescLDA suitable for text classification problems in open domains.

DescLDA has a number of advantages over supervised LDA models. Firstly, DescLDA requires only category descriptions as supervision, saving the human labor of producing labeled data required by supervised LDA models. Secondly, there can be no risk of overfitting in model training

since no labeled data is required. Thirdly, DescLDA is applicable in cases where only descriptive words are available; humans can thus concentrate on precisely describing a specific category, rather than building/adapting semantic resources or labeling documents.

DescLDA is the first successful application of a topic modeling-based method to DLTC. Experimental results show that our method outperforms the semantic-based DLTC baseline and performs at a similar level to state-of-the-art supervised text classification methods. The main contributions of this paper are:

1. Proposing DescLDA, which couples a describing device to infer Dirichlet priors (β) with a standard LDA model in order to induce category-aware latent topics.
2. Designing the DescLDA based DLTC algorithm, which requires neither external resources nor labeled data.
3. Evaluating our method against the DLTC baseline method and state-of-the-art supervised text classification methods on the 20Newsgroups (Lang 1995) and RCV1 (Lewis et al. 2004) datasets.

Related Work

Dataless Text Classification

Dataless text classification (DLTC) methods can be divided into two types: classification-based and clustering-based. Classification-based methods employ automatic algorithms to create machine-labeled data. Ko and Seo (2004) use category labels and keywords to bootstrap context clusters based on co-occurrence information. The context clusters are viewed as labeled data to train a Naive Bayes classifier. Unfortunately the quality of the machine-labeled data is hard to control, which may result in unpredictable bias. Liu et al. (2004) annotate a set of descriptive words for each class, which are used to extract a set of unlabeled documents to form the initial training set. The EM algorithm is then applied to build a classifier with a better pseudo training set. However, judging whether a word is representative of a class is a difficult task, and inappropriate annotations may result in biased training data.

In contrast, clustering-based methods first measure the similarity between documents using models built on category labels/descriptions, cluster the test documents, and finally assign the clusters to categories. Gliozzo, Strapparava, and Dagan (2005) use latent semantic space to calculate coarse similarity between documents and labels, and the Gaussian Mixture algorithm to obtain uniform classification probabilities for unlabeled documents. Barak, Dagan, and Shnarch (2009) improve the similarity calculation by identifying concrete terms relating to the meaning of the category labels from WordNet and Wikipedia. Wikipedia is also used by Chang et al. (2008), who propose a nearest-neighbor based method. The drawback is that a large-scale semantic knowledge base is required.

Our work follows the clustering-based approach, but differs from previous work in requiring no external resources.

Supervised LDA Models

LDA (Blei, Ng, and Jordan 2003) is widely used in topic modeling. LDA assumes that each document in a corpus is generated by a mixture of topics. Each topic is a distribution over all words in the vocabulary.

LDA has been successfully revised for a supervised learning setting. Blei and McAuliffe (2007) propose supervised LDA (sLDA) which uses labeled documents to find latent topics that can best predict the categories of unlabeled documents. Lacoste-Julien, Sha, and Jordan (2008) introduce discriminative LDA (DiscLDA) which applies a class-dependent linear transformation on the topic mixture proportions. Ramage et al. (2009) propose labeled LDA (lLDA) which constrains LDA by defining a one-to-one correspondence between topics and document labels. Zhu, Ahmed, and Xing (2012) introduce Maximum Entropy Discrimination LDA (MedLDA) which explores the maximum margin principle to achieve predictive representations of data and more discriminative topic bases. All of these studies require labeled documents to infer category-aware latent topics.

LDA has also been adapted to include external supervision. Lin and He (2009) incorporate a sentiment lexicon in a joint sentiment-topic model for sentiment analysis. Boyd-Graber, Blei, and Zhu (2007) incorporate a WordNet hierarchy when building topic models for word sense disambiguation. Rosen-Zvi et al. (2004) extract author names from articles and use them in an author-topic model for author-oriented document classification. Although these methods do not use labeled documents, they do rely on external resources.

Our work differs from the above in two major respects: we use neither labeled documents nor external resources. The only supervision comes from the descriptive words in category labels/descriptions, which are much easier to obtain.

Model

Below we present standard LDA, our DescLDA model, and an explanation of the describing device and sampling. Then we explain how to create the prior descriptive documents.

LDA

In LDA, the generative process of a corpus D consisting of documents d each of length N_d is as follows:

1. Choose $\beta \sim Dir(\eta)$.
2. Choose $\theta \sim Dir(\alpha)$.
3. For the n-th word w_n in document d:
   (a) Choose a topic $z_n \sim Multi(\theta)$.
   (b) Choose a word w_n from $p(w_n|z_n, \beta)$

Assuming the documents in the corpus are independent of each other, the corpus probability is:

$$p(D|\alpha,\beta) = \prod_{d\in D}\int p(\theta_d|\alpha)\Big(\prod_{n=1}^{N_d}\sum_{z_{d_n}} p(z_{d_n}|\theta_d)p(w_{d_n}|z_{d_n},\beta)\Big)d\theta_d, \quad (1)$$

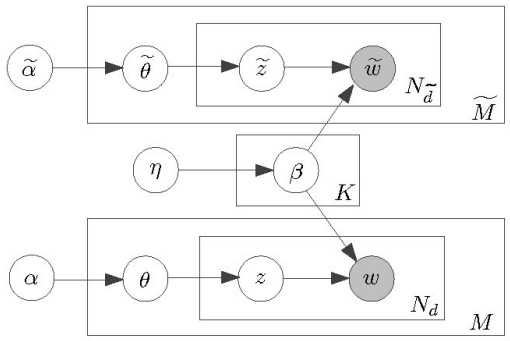


Figure 1: In DescLDA, a describing device (above) is coupled to a standard LDA model (below).

where α and η are hyper-parameters that specify the nature of the priors on θ and β (for smoothing the probability of generating word w_{d_n} in a document d from topic z_{d_n}).

LDA aims to induce the topics $\beta_{1:K}$ that can precisely represent the whole corpus by maximizing $p(D|\alpha, \beta)$ based on word co-occurrences. Previous research (sLDA, etc.) uses labeled documents as supervision. However, our Descriptive LDA model deals with the classification task with supervision coming merely from category labels/descriptions.

DescLDA

Descriptive LDA (DescLDA) is an adaptation of LDA that incorporates a describing device (DD). DD infers Dirichlet priors (i.e., β) from category labels/descriptions, and these priors are shared with LDA. The Dirichlet priors drive LDA to induce the category-aware topics. Figure 1 illustrates this.

In DescLDA, the generative process of an ordinary corpus D consisting of documents d each of length N_d, and a descriptive corpus $\widetilde{D}$ consisting of prior descriptive documents $\widetilde{d}$ each of length $N_{\widetilde{d}}$ is:

1. Choose $\beta \sim Dir(\eta)$.
2. For the prior descriptive document $\widetilde{d}$, choose $\widetilde{\theta} \sim Dir(\widetilde{\alpha})$.
3. For the n-th word $\widetilde{w_n}$ in the descriptive document $\widetilde{d}$:
   (a) Choose a topic $\widetilde{z} \sim Multi(\widetilde{\theta})$
   (b) Choose a word $\widetilde{w}$ from $p(\widetilde{w}|\widetilde{z}, \beta)$.
4. For the ordinary document $\widetilde{d}$, choose $\theta \sim Dir(\alpha)$.
5. For n-th word w_n in ordinary document d:
   (a) Choose a topic $z_n \sim Multi(\theta)$
   (b) Choose a word w_n from $p(w|z, \beta)$.

Let the global corpus $\hat{D}$ be the union of D and $\widetilde{D}$. Assuming the documents are independent, the probability of $\hat{D}$ is:

$$p(\hat{D}|\alpha, \widetilde{\alpha}, \beta) = p(D|\alpha, \beta)p(\widetilde{D}|\widetilde{\alpha}, \beta), \tag{2}$$

where $p(\widetilde{D}|\widetilde{\alpha}, \beta)$ is the probability of descriptive corpus $\widetilde{D}$:

$$p(\widetilde{D}|\widetilde{\alpha}, \beta) = \prod_{\widetilde{d} \in \widetilde{D}} \int p(\widetilde{\theta_d}|\widetilde{\alpha}) \Big(\prod_{n=1}^{N_{\widetilde{d}}} \sum_{\widetilde{z_{d_n}}} p(\widetilde{z_{d_n}}|\widetilde{\theta_d}) p(\widetilde{w_{d_n}}|\widetilde{z_{d_n}}, \beta) \Big) d\widetilde{\theta_d}. \tag{3}$$

Note that in LDA, α in Figure 1 is a vector. But in DescLDA, $\widetilde{\alpha}$ in Eq.3 is a square matrix. By adjusting $\widetilde{\alpha_k}$, we are able to influence the topic β_k. In this paper, for simplicity, we define $\widetilde{\alpha}$ as a unit diagonal matrix to make the i-th prior descriptive document correspond to the i-th topic.

Describing Device

The describing device (DD) is a simple LDA model which generates the prior descriptive documents. DD consists of:

- *Descriptive corpus* ($\widetilde{D}$): contains descriptive documents constructed with category labels/descriptions.
- *Descriptive parameter* (β): a parameter which is generated by DD and shared with LDA.
- *Other LDA parameters*: hyper-parameter $\widetilde{\alpha}$ and the length of each describing document $N_{\widetilde{d}}$.

Note that the approach to classification of Lin and He (2009) uses external resources to generate Dirichlet priors. One could thus ask whether the Dirichlet priors in the DescLDA model could be defined by a human rather than being automatically inferred by the describing device. Our answer is negative. We argue that human-defined priors can be either too general or too arbitrary. Instead, the automatically-inferred priors can make DescLDA adaptable to open domains.

Sampling

Word co-occurrences play a key role in parameter estimation in the probabilistic topic model. We therefore investigate what influences Gibbs sampling in DescLDA and how co-occurrences allow DescLDA to infer categories.

In Gibbs sampling (Griffiths and Steyvers 2004), the probability of $w \in d$ generated by topic β_k is:

$$p(z_j = k|z_{-j}, w) = \frac{\theta_k \beta_{k,w}}{\sum_{m=1}^{K} \theta_m \beta_{m,w}}, \tag{4}$$

where $\theta_m = \frac{n^{(d)}_{-j,m} + \alpha}{n^{(d)}_{-j,\cdot} + K\alpha}$, $\beta_{m,w} = \frac{n^{(w)}_{-j,m} + \eta}{n^{(\cdot)}_{-j,m} + W\eta}$ and W is vocabulary size. The expectation of variable θ_k is:

$$\mathrm{E}(\theta_k) = \frac{\alpha + \sum_{i \neq j} p(z_i = k|z_{-i}, w_i)}{N_d + K\alpha}. \tag{5}$$

Replacing θ_k with $\mathrm{E}(\theta_k)$ in Eq.4, we obtain

$$p(z_j = k|z_{-j}, w) = \frac{1}{N_d + K\alpha} \frac{\beta_{k,w}}{\sum_{m=1}^{K} \theta_m \beta_{m,w}} \sum_{i \neq j} \Big[p(z_i = k|z_{-i}, w) + \alpha \Big]. \tag{6}$$

The probability of word w generated by the k-th topic β_k is determined by the probabilities of the other words in this document generated by the k-th topic β_k.

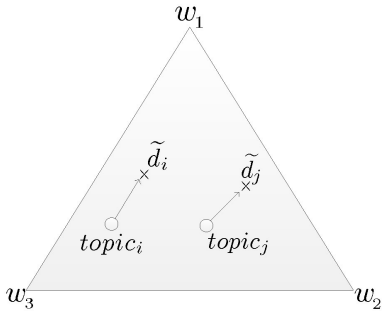


Figure 2: An illustration of DescLDA sampling.

Consider another word v also occurring in d. Let $n_d(v)$ be the number of occurrences of v in d. After one iteration of Gibbs sampling:

$$n_{k,v}(w) = \sum_{d \in D} \frac{n_d(w) n_d(v)}{N_d + K\alpha} \frac{\beta_{k,w}}{\sum_{m=1}^{K} \theta_{d,m} \beta_{m,w}} \Big(p(z_{d,v} = k) + \alpha \Big). \quad (7)$$

Eq.7 shows how word v influences word w during sampling; $n_{k,v}(w)$ is determined by three components:

1. $\frac{n_d(w) n_d(v)}{N_d + K\alpha}$, referred to as the co-occurrence factor,
2. $\frac{\beta_{k,w}}{\sum_{m=1}^{K} \theta_{d,m} \beta_{m,w}}$, and
3. the topic probability $p(z_{d,v} = k)$.

The second component is not adjustable. Thus only the co-occurrence factor and topic probability influence Gibbs sampling.

We now explain how this makes DescLDA capable of inferring categories. First, to form the prior descriptive documents, we select words with a higher co-occurrence factor with the categories, referred to as *descriptive words*. Category labels are the best choices. Often, category descriptions are also available, and the words in these are also good choices. Next, to improve $p(z_{d,v} = k)$ we repeat the descriptive words in the descriptive documents, thus increasing the sample probability of the words which frequently co-occur with word v. Given these descriptive documents, DD finds the optimal descriptive parameter β, which LDA uses to induce topics that correspond to the categories. This is illustrated in Figure 2: topics (denoted by $\circ$) are pulled by the descriptive documents (denoted by $\times$) rather than word co-occurrences alone (see Eq.7). As a result, each test document will be assigned a topic corresponding to a descriptive document from a category.

For example, considering the category labeled with *earnings* in the RCV1 corpus, we view *earning* as the descriptive word. Then we obtain a descriptive document for this category by repeating the descriptive word a few times. The describing device is able to increase the probability of word *earning* in topic $z_{d,\text{earning}}$. Meanwhile, words that have a high co-occurrence factor with *earning* can also be pulled from the documents to induce the topic $z = \text{earning}$.

Table 1: Descriptive words for the RCV1 dataset.

Category	Label	Descriptive words
acq	acquisition	acquisition, merger, cash, takeover, sale, agreement, asset, purchase, buy
coffee	coffee	coffee, export, ico, quota
crude	crude	crude, oil, gas, petroleum, energy, bp, barrel, opec, pipeline
earn	earnings	earning, net, income, loss, cost, profit, gain
money-fx	foreign exchange	foreign exchange, currency exchange, bank rate, monetary, finance, budget, currency
interest	interest	interest, bank rate, money rate, bank, bill, interest rate, debt, loan
gold	gold	gold, mining, ounce, resource
ship	ship	ship, port, cargo, river, seamen, refinery, water, vessel
sugar	sugar	sugar, tonne
trade	trade	trade, foreign agreement, export, goods, import, industry

Descriptive Documents

Definition of the Descriptive Words Descriptive words are ordinary words that can jointly describe a category. For example, *earning*, *profit* and *cost* could be the descriptive words for a category *earnings*. A single descriptive word may not adequately describe a category; for example, the word *earning* may appear in many categories, so to describe the category *earnings* it should be combined with other descriptive words such as *profit* and *cost*.

Choosing the Descriptive Word(s) We extract the descriptive word(s) from the category labels/descriptions. For the 20Newsgroups dataset, we use the category descriptions of Song and Roth (2014). For RCV1, similarly to Xie and Xing (2013), we use the ten largest categories; unfortunately there are no category descriptions available, so we developed the following procedure to compile the descriptive words:

1. Without using category labels, run LDA on the documents to induce 30 latent topics from the documents in RCV1.
2. Manually assign a category label to each latent topic, following Hingmire and Chakraborti (2014) – although in contrast to that work we discard latent topics that cannot be assigned a category label.
3. Manually select the descriptive words from each latent topic assigned a category label.

Table 1 shows the descriptive words for the RCV1 categories. We note that there are other approaches that could be used to mine descriptive words. For example, synonymous words could be extracted from a dictionary, or related entries could be retrieved from Wikipedia. However, we choose not to use external resources, but merely to perform a minimal amount of manual filtering.

Constructing the Descriptive Document(s) We assume that the descriptive words for a category contribute equally, so we just list the words in the descriptive document for the category. However, there are usually many occurrences of

the descriptive words in the corpus. To make these words visible to the category we take a pragmatic approach and repeat them in the descriptive document. To determine how many repetitions, we note that in Eq.6, two factors should be considered in selecting the descriptive words. $p(z_j = k|z_{-j}, w)$ is determined by two components: $\frac{\beta_{k,w}}{\sum_{m=1}^{K}\theta_m\beta_{m,w}}$ and $p(z_i = k|z_{-i}, w)$. With a very low number, the first component will very high – and vice versa. Neither case will produce a useful topic-word probability.

In this work, we simply repeat each descriptive word a number of times which is proportional to the frequency of each descriptive word in the corpus. Note that the descriptive document is category-specific. In other words, a word can only serve as a descriptive word for one category.

Algorithm

Our DescLDA-based DLTC method comprises three steps:

1. Construct the descriptive documents.
2. Induce latent topics with DescLDA.
3. Assign category labels to the test documents.

Steps 1 and 2 are described in the previous section. For step 3, recall that DescLDA induces latent topics from the global corpus $\hat{D}$ (consisting of the ordinary corpus D and the descriptive corpus $\widetilde{D}$). In the end, every descriptive document will be probabilistically assigned to the induced topics. Based on the document-topic distribution, we compute an optimal partition of $\hat{D}$ to obtain document clusters. Given a cluster that contains a descriptive document, we assign the category label of the descriptive document to every document in the cluster. Algorithm 1 presents this more formally.

Algorithm 1: DescLDA-based dataless text classification

Input:
A collection of test documents D
A set of category labels L
A set of category descriptions S
Output:
Category labels $\hat{L}[\,]$ of the test documents D

$i \leftarrow 0$ % Initialize category index
$\hat{D} \leftarrow NULL$ % Corpus
for $i < |L|$ **do**
  $w_i^{\text{label}}[\,] \leftarrow extract\_words(L[i])$
  $w_i^{\text{desc}}[\,] \leftarrow extract\_words(S[i])$
  $w_i^{\text{all}}[\,] \leftarrow combine(w_i^{\text{label}}[\,], w_i^{\text{desc}}[\,])$
  $D^{\text{prior}} \leftarrow generate\_desc\_document(w_i^{\text{all}}[\,])$
  $\hat{D} \leftarrow \hat{D} + D^{\text{prior}}$
$T \leftarrow \mathbf{DescLDA}(D, \widetilde{D}, \alpha, \widetilde{\alpha}, \eta)$ % Inducing topics
$C \leftarrow cluster\_documents(T)$
$j \leftarrow 0$ % Reset cluster index
for $j < |C|$ **do**
  $\widetilde{d} \leftarrow get\_desc\_doc(C[j])$
  $l \leftarrow get\_category\_label(\widetilde{d})$
  $k \leftarrow 0$ % Reset document index
  for $k < number\_of\_doc(T[j])|$ **do**
    $\hat{L}[k] \leftarrow l$
return Category labels $\hat{L}[\,]$ of the test documents

Evaluation

Setup

Datasets We use two datasets:

- 20Newsgroups (20NG): Introduced by Lang (1995), 20Newsgroups is a collection of approximately 20,000 newsgroup documents, partitioned (nearly) evenly across 20 different newsgroups. The dataset is divided into training (60%) and test (40%) sets. We use 20NG in our evaluation of multiclass text classification. In our evaluation of binary classification we use a sub-dataset, 20NG10 (Raina, Ng, and Koller 2006), which involves 10 binary classification tasks.

  The original category labels in 20NG are sometimes not real words (e.g., *sci.crypt*). Chang et al. (2008) propose expanding the 20NG category labels automatically to real words (e.g., *science cryptography*) according to the original data in the category. Song and Roth (2014) further provide a finely-compiled category description for each 20NG category. In this work, we use the category labels of Chang et al. (2008) and category descriptions of Song and Roth (2014).

- RCV1: An archive of multi-labeled newswire stories (Lewis et al. 2004), RCV1 contains 21,578 documents of 135 topics; 13,625 stories are used as the training set and 6,188 stories as the test set. In our text classification evaluation, we use the ten largest categories identified by Xie and Xing (2013), in which there are 5,228 training documents and 2,057 test documents.

  Note there are no category descriptions in RCV1. As described above, we designed a procedure to compile the descriptive words for the RCV1 categories. In the experiments, we use the descriptive words in Table 1 as category descriptions.

In our experiments we use the standard training/test partitions of the two datasets.

Evaluation Metrics We adopt the standard evaluation metric, accuracy, defined as the percentage of correctly classified documents out of all test documents.

Methods We evaluate DescLDA against three baseline methods:

- **SemNN**: the dataless text classification method presented by Chang et al. (2008), which uses category labels as supervision and adopts Wikipedia as an external resource to calculate semantic distance. We select SemNN as our baseline because it is the state-of-the-art dataless text classification model. However, SemNN is difficult to reproduce because it involves Wikipedia, a huge knowledge

base. We therefore cite the publicly-available experimental results (Chang et al. 2008). Unfortunately, these results do not include multiclass text classification. Thus the SemNN method is compared only in the binary classification experiment. We note that Song and Roth (2014) modify SemNN to deal with the dataless hierarchical text classification problem. However, we do not compare DescLDA against Song and Roth (2014) because we deal with the dataless *flat* text classification problem.

- SVM: the support vector machine model. We choose SVM because it is a state-of-the-art supervised text classification model. We follow Wang et al. (2013) and use linear SVM using the package LIBSVM[1] with the regularization parameter $C \in \{1e-4, ..., 1e+4\}$ selected by 5-fold cross-validation. Note that SVM is sensitive to the volume of training data, so we also report the number of training samples at which SVM starts to perform better than our DescLDA model.
- sLDA: the supervised LDA model Blei and McAuliffe (2007). We choose sLDA as our baseline because it is a text classification model aiming to deal with text classification problem via topic modeling in a supervised manner (i.e., requiring some labeled data). In our experiment, we adopt the implementation of Wang, Blei, and Li (2009)[2].

For our DescLDA method, we set $\alpha = 0.1$ and $\eta = 0.2$. We vary K (the number of topics) across the range used in previous work (Blei and McAuliffe 2007). For the number of iterations, in preliminary experiments we observed good accuracy at 30. We run DescLDA 5 times and report the average accuracy.

Binary Text Classification

For SVM and sLDA, we train binary classifiers on the labeled documents and evaluate on the test documents. DescLDA is evaluated on the same test documents. To create the descriptive documents, the descriptive words are repeated for 75 percent of their term frequencies in the corpus (this percentage being determined empirically in a preliminary study). Regarding the source of the descriptive words, we evaluate two settings: (1) DescLDA#1 which uses just category labels, and (2) DescLDA#2 which uses category descriptions. The results are shown in Figure 3.

DescLDA#1 and SemNN are comparable in that they use the same category labels. However, the DescLDA model considerably outperforms SemNN, by 3 percentage points, despite only receiving supervision from category descriptions rather than external semantic resources. DescLDA also slightly outperforms the supervised methods, SVM and sLDA. Although the DescLDA model is not statistically significantly better than sLDA[3], this result is still surprising since DescLDA is a weakly supervised dataless method.

Looking into the dataset (Lang 1995), we notice that the labeled documents in the two categories of each binary classification problem contain very different word co-occurrences. For example, in the *soc.religion.christian* vs. *rec.sport.hockey* classification problem there is little overlap. Thanks to the prior descriptive documents, DescLDA is sensitive to such contextual differences, so better classification predictions are made.

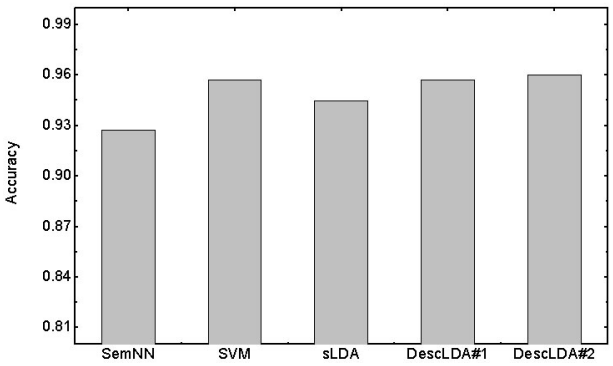


Figure 3: Binary classification applied to 20NG10.

Figure 3 also shows that DescLDA#2 (using category descriptions) performs slightly better than DescLDA#1 (using category labels). This is surprising since category descriptions contain more information than labels. We therefore conclude that labels are sufficiently powerful sources of information for binary classification in the 20NG10 dataset.

Multiclass Text Classification

The multiclass results are shown in Figure 4 (unfortunately, this experiment cannot include SemNN because there are no publicly available multiclass text classification results for that method). The accuracy of DescLDA#2 is close to SVM and sLDA, on both the 20NG and RCV1 datasets. We find that sLDA is not statistically significantly better than DescLDA#2 on either the RCV1 or 20NG datasets[4].

It is noteworthy that DescLDA#1 performs much worse than DescLDA#2. This observation is rather different from that in the binary classification problem above. The reason is that category labels are no longer sufficient for characterizing the categories in the multiclass text classification task. As a comparison, category descriptions contain a few high-quality descriptive words which are representative and discriminative. This is why a significant contribution is made to multiclass text classification accuracy by category descriptions on both datasets. We therefore conclude that high-quality descriptive words are crucial to our DescLDA model.

Descriptive Document Construction

Recall that the descriptive documents are constructed by repeating the descriptive words a number of times proportional to their term frequencies in the corpus. In this experiment, we investigate how the proportion influences the accuracy of DescLDA in the multiclass text classification task. We

[1] http://www.csie.ntu.edu.tw/∼cjlin/libsvm/

[2] http://www.cs.cmu.edu/∼chongw/slda/

[3] One-tailed paired-sample t-test p-value=0.072 on 20NG10.

[4] One-tailed paired-sample t-test p-value=0.079 on RCV1 and 0.243 on 20NG

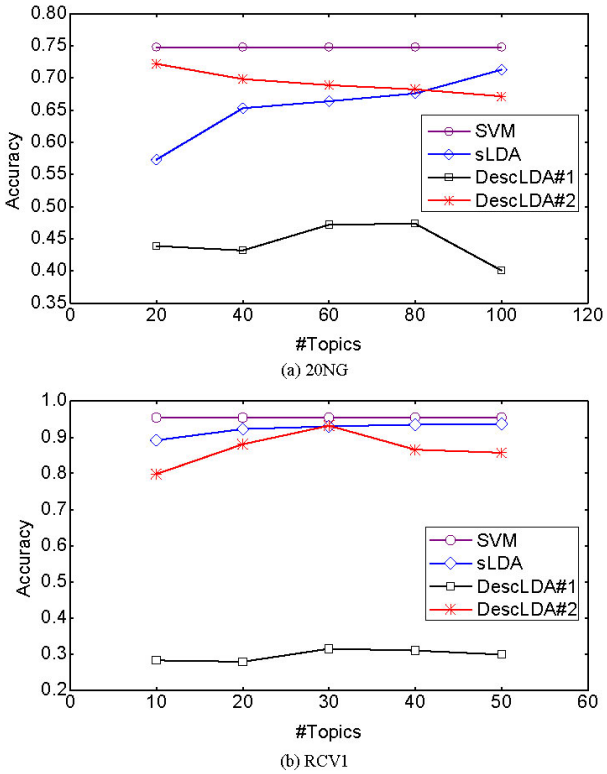


Figure 4: Multiclass text classification applied to (a) 20NG and (b) RCV1.

vary the proportion from 10% to 300%. Experimental results are presented in Figure 5. It can be seen from Figure 5 that DescLDA achieves the best accuracy between 25% and 100%, on both RCV1 and 20NG.

Volume of SVM Training Data

We vary the amount of training data in order to find the number of training samples at which SVM starts to perform better than our DescLDA model. We randomly select samples from the training dataset to create smaller datasets with the proportion of data in each category being identical to the whole training dataset.

Figure 6 shows that our dataless DescLDA model performs better than SVM when there are fewer than 425 (20NG) or 250 (RCV1) training samples in each category. Another interesting finding is that volume of training data for a high-quality SVM classifier varies greatly on two datasets. In practice, one is difficult to foresee how many labeled samples are enough to train a good SVM classifier for a new domain. In some extreme cases, the volume of training data is very big. This justifies the advantage of DescLDA model, which requires no labeled data in addressing the text classification problem.

Conclusions and Future Work

In this paper we proposed *descriptive LDA* (DescLDA) as a way of realizing dataless text classification (DLTC).

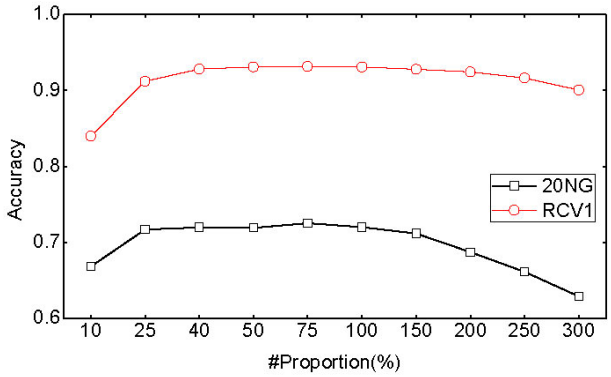


Figure 5: DescLDA using different proportions of descriptive words to construct the descriptive documents.

DescLDA has two major advantages over previous approaches. Firstly, no external resources are required; using only category labels/descriptions, DescLDA is able to induce descriptive topics from the unlabeled documents. Moreover, it achieves better accuracy than semantic-based DLTC methods that use external semantic knowledge. Secondly, no labeled data is required to train a classifier. By incorporating a *describing device*, DescLDA is able to infer Dirichlet priors (β) from descriptive documents created from category description words. The Dirichlet priors are in turn used by LDA to induce category-aware latent topics. In our binary and multiclass text classification experiments, DescLDA achieves accuracies that are comparable to supervised models, i.e., SVM and sLDA.

There are a number of opportunities for further research. Firstly, in this study the descriptive words are explicitly extracted from category descriptions; we intend to investigate techniques for refining and extending these sets of words. Secondly, as a simplifying assumption, we give each descriptive word an equal contribution in the descriptive documents; we will investigate lifting this assumption and allowing them to make different contributions. Thirdly, DescLDA could be well-suited to multi-label classification, since test documents can be probabilistically assigned to different descriptive topics; we will investigate this possibility.

Acknowledgments

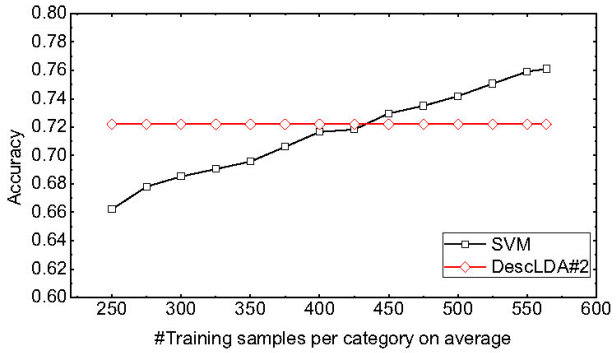
This work is partially supported by the National Science Foundation of China (61373056, 61272233). Peng Jin is the corresponding author. We thank the anonymous reviewers for their insightful comments.

References

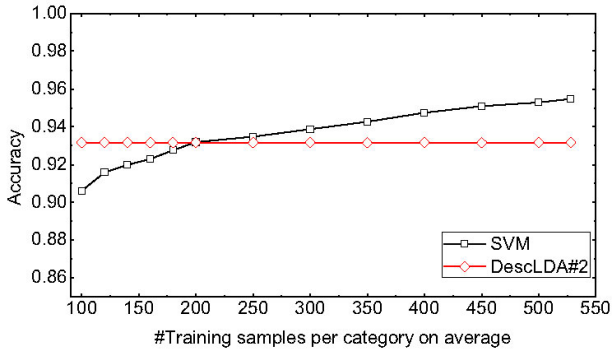
Barak, L.; Dagan, I.; and Shnarch, E. 2009. Text categorization from category name via lexical reference. In *Proceedings of NAACL'09-Short*, 33–36.

Blei, D. M., and McAuliffe, J. D. 2007. Supervised topic models. In *Proceedings of NIPS'07*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.



(a) 20NG



(b) RCV1

Figure 6: SVM using different numbers of training samples.

Blum, A., and Mitchell, T. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of COLT'98*, 92–100.

Boyd-Graber, J. J.; Blei, D. M.; and Zhu, X. 2007. A topic model for word sense disambiguation. In *Proceedings of EMNLP-CoNLL'09*, 1024–1033.

Bruzzone, L., and Marconcini, M. 2010. Domain adaptation problems: A DASVM classification technique and a circular validation strategy. *IEEE Trans. Pattern Anal. Mach. Intell.* 32(5):770–787.

Cawley, G. C., and Talbot, N. L. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *J. Mach. Learn. Res.* 11:2079–2107.

Chang, M.-W.; Ratinov, L.; Roth, D.; and Srikumar, V. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of AAAI'08 - Volume 2*, 830–835.

Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of IJCAI'07*, 1606–1611.

Gliozzo, A.; Strapparava, C.; and Dagan, I. 2005. Investigating unsupervised learning for text categorization bootstrapping. In *Proceedings of HLT'05*, 129–136.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences* 101(Suppl. 1):5228–5235.

Hingmire, S., and Chakraborti, S. 2014. Sprinkling topics for weakly supervised text classification. In *Proceedings of ACL'14-short paper*, 55–60.

Ko, Y., and Seo, J. 2004. Learning with unlabeled data for text categorization using bootstrapping and feature projection techniques. In *Proceedings of ACL'04*, 255–262.

Lacoste-Julien, S.; Sha, F.; and Jordan, M. 2008. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Proceedings of NIPS'08*.

Lang, K. 1995. NewsWeeder: Learning to filter netnews. In *Proceedings of ICML'95*, 331–339.

Lewis, D. D.; Yang, Y.; Rose, T. G.; and Li, F. 2004. RCV1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research* 5(Apr):361–397.

Lin, C., and He, Y. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of CIKM'09*, 375–384.

Liu, B.; Li, X.; Lee, W. S.; and Yu, P. S. 2004. Text classification by labeling words. In *Proceedings of AAAI'04*, 425–430.

Nigam, K.; McCallum, A. K.; Thrun, S.; and Mitchell, T. 2000. Text classification from labeled and unlabeled documents using EM. *Mach. Learn.* 39(2-3):103–134.

Raina, R.; Ng, A. Y.; and Koller, D. 2006. Constructing informative priors using transfer learning. In *Proceedings of ICML'06*, 713–720.

Ramage, D.; Hall, D.; Nallapati, R.; and Manning, C. D. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of EMNLP'09*, 248–256.

Rosen-Zvi, M.; Griffiths, T.; Steyvers, M.; and Smyth, P. 2004. The author-topic model for authors and documents. In *Proceedings of UAI'04*, 487–494.

Song, Y., and Roth, D. 2014. On dataless hierarchical text classification. In *Proceedings of AAAI'14*, 1579–1585.

Wang, Q.; Xu, J.; Li, H.; and Craswell, N. 2013. Regularized latent semantic indexing: A new approach to large-scale topic modeling. *ACM Trans. Inf. Syst.* 31(1):5:1–5:44.

Wang, C.; Blei, D. M.; and Li, F. 2009. Simultaneous image classification and annotation. In *Proceedings of IEEE CVPR'09*, 1903–1910.

Xie, P., and Xing, E. P. 2013. Integrating document clustering and topic modeling. *CoRR* abs/1309.6874.

Zhu, J.; Ahmed, A.; and Xing, E. P. 2012. MedLDA: Maximum margin supervised topic models. *J. Mach. Learn. Res.* 13(1):2237–2278.