A Pseudo Label based Dataless Naive Bayes Algorithm for Text Classification with Seed Words

Ximing Li, Bo Yang*

College of Computer Science and Technology, Jilin University, China
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, China
liximing86@gmail.com; ybo@jlu.edu.cn

Abstract

Traditional supervised text classifiers require a large number of manually labeled documents, which are often expensive to obtain. Recently, dataless text classification has attracted more attention, since it only requires very few seed words of categories that are much cheaper. In this paper, we develop a pseudo-label based dataless Naive Bayes (PL-DNB) classifier with seed words. We initialize pseudo-labels for each document using seed word occurrences, and employ the expectation maximization algorithm to train PL-DNB in a semi-supervised manner. The pseudo-labels are iteratively updated using a mixture of seed word occurrences and estimations of label posteriors. To avoid noisy pseudo-labels, we also consider the information of nearest neighboring documents in the pseudo-label update step, i.e., preserving local neighborhood structure of documents. We empirically show that PL-DNB outperforms traditional dataless text classification algorithms with seed words. Especially, PL-DNB performs well on the imbalanced dataset.

1 Introduction

Automatic text classification is one of the most popular directions in the machine learning community. A typical procedure for creating a classifier in supervised learning consists of two steps: (1) Given a collection of text documents, human experts define a number of category labels, and then manually assign these pre-defined labels to documents. We refer to these labeled documents as training dataset; (2) A supervised learning algorithm, e.g., support vector machines (SVMs), is trained on the training dataset, outputting a classifier for predicting future documents.

Manually labeling documents, i.e., step (1), is very expensive and time-consuming. Unfortunately, supervised learning algorithms often require massive labeled documents to avoid learning issues such as overfitting (Cawley and Talbot, 2010). A common way to reduce the labeling effort is developing semi-supervised learning algorithms, where one trains text classifiers on a mixture collection of a few labeled documents and a larger number of unlabeled documents (Nigam et al., 2000; Hu et al., 2017). However, semi-supervised learning still requires labeled documents, and manually labeling a small number of documents remains very expensive in many real world applications.

Recently, researchers have proposed a number of dataless text classification algorithms (Liu et al., 2004; Chang et al., 2008; Downey and Etzioni, 2008; Druck et al., 2008; Hingmire et al., 2013; Hingmire and Chakraborti, 2014; Chen et al., 2015; Li et al., 2016), which do not require labeled documents as training instances. Instead, they train text classifiers using unlabeled documents with seed words, i.e., the selected representative words for categories. Actually, manually choosing seed words is significantly cheaper than labeling documents (Raghavan et al., 2006; Druck et al., 2008), saving many human efforts. This kind of dataless algorithm has empirically achieved promising classification results, and it has become a practical alternative to supervised learning algorithms, especially when the labeled documents are extremely expensive to obtain.

* corresponding author

Proceedings of the 27th International Conference on Computational Linguistics, pages 1908–1917
Santa Fe, New Mexico, USA, August 20-26, 2018.

A straightforward way of dataless classification is to construct a pseudo training dataset using the supervision information provided by seed words, before applying a traditional supervised classification algorithm (Liu et al., 2004; Druck et al., 2008). However, this methodology suffers from two problems. First, the documents containing no seed words can not be marked with any category label, resulting in a waste of instances to some extent. Unfortunately, there may exist a large number of such "unlabeled" documents, especially when the seed words are scarce. Second, the seed word based pseudo-labels are often quite noisy. That is because many documents may not contain the seed words from the true categories, but only contain the ones from unassociated categories.

To address the problems mentioned above, we develop a novel extension of the Naive Bayes classifier for dateless text classification, named pseudo-label based dataless Naive Bayes (PL-DNB). First, following (Nigam et al., 2000) we employ the expectation maximization (EM) algorithm to train PL-DNB in a semi-supervised manner, so that all the documents with and without pseudo-labels can be used for creating a classifier. Second, we iteratively update the pseudo-labels with highly acceptable confidence. The confidence values of pseudo-labels are measured by a mixture of seed word occurrences and estimations of label posteriors. Additionally, we aim to further avoid noisy pseudo-labels by preserving local neighborhood structure of documents. That is, for each document we update its pseudo-label by also considering the information of its nearest neighboring documents. We empirically compare the proposed PL-DNB algorithm against the dataless classification algorithms and traditional supervised algorithms. Experimental results indicate that our PL-DNB outperforms the existing dataless algorithms with seed words, and especially performs well on the imbalanced dataset

The rest of this paper is organized as follows: In Section 2, we introduce recent related works. In Section 3, we describe the proposed PL-DNB algorithm for dataless text classification. Section 4 presents the empirical results, and the conclusion is shown in Section 5.

2 Related Work

In this section, we review recent related works on dataless text classification and semi-supervised naive Bayes.

Dataless text classification There are some previous dataless text classifiers based on pseudo-labels. For example, the seed word naive Bayes (SNB) (Liu et al., 2004) computes information gain values of words over document clusters computed by k-means, and manually selects some top-ranked words as seed words to represent each category. It assigns pseudo-labels or probabilistic pseudo-labels to documents using these selected seed words. A naive Bayes classifier is then trained by the EM algorithm in a semi-supervised manner. Another dataless classification algorithm (Ko and Seo, 2004) constructs context-clusters as the basic unit using sliding windows, instead of documents. These context-clusters are associated with pseudo-labels using representative words occurring in category labels and titles. A naive Bayes classifier is also trained over these bootstrapped context-clusters. Compared with these previous algorithms, our PL-DNB creates pseudo-labels by further considering the information of nearest neighboring documents.This can lead to more accurate pseudo-labels than only using seed words.

Specifically, other dataless text classifiers aims to accurately categorize text documents by understanding the labels, i.e., embedding documents and category labels into a same semantic space (Chang et al., 2008; Palatucci et al., 2009; Elhoseiny et al., 2013; Song and Roth, 2014). The work of (Chang et al., 2008) builds a semantic space by exploiting the concepts of Wikipedia, and uses explicit semantic analysis to measure the distance between documents and labels. In some sense, this algorithm can be considered as a text classifier based on distant supervision (Mintz et al., 2008). Recently, the authors of (Song and Roth, 2014) further investigate dataless hierarchical text classification tasks. These dataless algorithms empirically performed well, however, they require auxiliary knowledge bases that are not always available.

Additionally, researchers have investigated some dataless text classifiers based on topic models, such as latent Dirichlet allocation (LDA) (Blei et al., 2003). The dataless classification algorithm proposed in (Hingmire et al., 2013), named classifyLDA, involves three steps: 1) learn a set of original topics by inferring the unsupervised LDA model over the concerned dataset; 2) manually assign a category

label to each topic, and then combine the topics that are associated with a same label into a new single topic; 3) classify test documents using the document-level posterior of those combined topics learned by unsupervised LDA. An extension algorithm (Hingmire and Chakraborti, 2014) of classifyLDA allows the original topics to be assigned to more than one category label in step 2, making the algorithm more flexible. Recently, the authors of (Li et al., 2016) propose a seed-guided topic model (STM), which simultaneously considers category word probability and initial document category distribution based on seed words. In STM, each category label is associated with a category-topic, and it further incorporates a new type of topic, i.e., general-topic, to filter out the noise words during Gibbs sampling. Empirical study in (Li et al., 2016) shows that STM consistently outperforms many existing dataless text classification algorithms. We present comparison results of this state-of-the-art STM in the experiment section.

Semi-supervised naive Bayes To the best of our knowledge, an early semi-supervised naive Bayes proposed in (Nigam et al., 2000) uses the EM algorithm to train a classifier over labeled and unlabeled documents, named NB-EM. The pooling multinomials algorithm (Yao et al., 2009) incorporates training documents with auxiliary knowledge, building a composite naive Bayes classifier for semantic analysis tasks. Recently, (Zhao et al., 2016) extends NB-EM by leveraging the word-level statistical constraint. These semi-supervised naive Bayes algorithms can be directly applied to text classification applications with a few labeled documents. However, manually collecting a small number of labeled documents remains expensive in many text analysis applications. In contrast, our PL-DNB only requires a much cheaper set of seed words, making it more practical.

3 Algorithm

We briefly review the framework of semi-supervised naive Bayes, and then present the proposed pseudo-label based dataless naive Bayes (PL-DNB) algorithm for dataless text classification with seed words.

3.1 Semi-supervised Naive Bayes

Suppose that there exists a collection of text documents $\mathcal{D}$, consisting of a subset of labeled documents $\mathcal{D}_L = \{(d_i, y_i)\}_{i=1}^{i=|\mathcal{D}_L|}$ and a subset of unlabeled ones $\mathcal{D}_U = \{d_i\}_{i=1}^{i=|\mathcal{D}_U|}$. Let $\mathcal{C} = \{c_i\}_{i=1}^{i=|\mathcal{C}|}$ and $\mathcal{W} = \{w_i\}_{i=1}^{i=|\mathcal{W}|}$ denote the set of category labels and the vocabulary of words, respectively.

The goal of semi-supervised text classification is to train a classifier over $\mathcal{D}$, which can automatically assign a correct category label for any test document. Semi-supervised naive Bayes is built on the Bayesian rule. Thanks to the bag-of-words assumption, it can classify a document d by a fully factored posterior distribution of labels:

$$y = \max_{c_i \in \mathcal{C}} \Pr(y = c_i|d), \qquad \Pr(y = c_i|d) \propto \Pr(y = c_i) \prod_{j=1}^{|\mathcal{W}|} \Pr(w_j|y = c_i)^{N_{d,w_j}} \tag{1}$$

where N_{d,w_j} is the occurrence number of word w_j in document d.

To achieve the parameter $\theta = \{\Pr(y), \Pr(w_j|y)\}$, semi-supervised naive Bayes employs the EM algorithm to maximize the following objective $\mathcal{L}(\theta)$, a mixture of the joint likelihood of labeled documents and marginal likelihood of unlabeled ones:

$$\mathcal{L}(\theta) = \sum\nolimits_{d \in \mathcal{D}_L} \log p(d, y) + \lambda \sum\nolimits_{d \in \mathcal{D}_U} \log p(d) \tag{2}$$

where $\lambda \in [0, 1]$ is a tuning parameter, controlling the importance of unlabeled documents.

3.2 Pseudo-label based Dataless Naive Bayes

Algorithm outline Overall, our PL-DNB is built on semi-supervised naive Bayes.

In the context of dataless classification, there are no labeled documents available. To address this, we initialize pseudo-labels $\widehat{y}^{(0)}$ for documents using seed word occurrences, leading to a set of documents with pseudo-labels $\mathcal{D}_L^{(0)}$ and a set of unlabeled documents that contain no seed words $\mathcal{D}_U^{(0)}$. After this initialization, PL-DNB iteratively performs the following two steps until the maximum iterative number is reached.

- **Step 1**: At each iteration *t*, use the EM algorithm to estimate the naive Bayes parameter $\theta^{(t)}$ by maximizing the following semi-supervised objective:

$$\mathcal{L}^{(t)}(\theta) = \sum\nolimits_{d \in \mathcal{D}_L^{(t-1)}} \log p(d, \widehat{y}^{(t-1)}) + \lambda \sum\nolimits_{d \in \mathcal{D}_U^{(t-1)}} \log p(d) \tag{3}$$

- **Step 2**: Given the optimum of $\theta^{(t)}$, compute the label posterior distributions for each document. We update pseudo-labels $\widehat{y}^{(t)}$ using these label posteriors and seed word occurrences. Only the pseudo-labels with acceptable confidence are left, leading to new document sets of $\mathcal{D}_L^{(t)}$ and $\mathcal{D}_U^{(t)}$.

Given the final optimum of θ, we can classify test documents using the label posterior distributions computed by Eq.1. For clarity, PL-DNB is summarized in *Algorithm 1*. We then introduce details of pseudo-label initialization, naive Bayes parameter θ update and pseudo-label $\widehat{y}$ update .

Algorithm 1 PL-DNB outline

1: **Initialize** pseudo-labels $\widehat{y}^{(0)}$ by seed word occurrences, obtaining $\mathcal{D}_L^{(0)}$ and $\mathcal{D}_U^{(0)}$
2: **For** $t = 1, 2, \ldots,$ MaxIter
3: **Update** the naive Bayes parameter $\theta^{(t)}$ using EM. Details are outlined in *Algorithm 2*
4: **Set** $\mathcal{D}_L^{(t)} = \mathcal{D}_U^{(t)} = \emptyset$
5: **For** $i = 1, 2, \ldots, |\mathcal{D}|$
6: **Assign** d_i a pseudo-label $\widehat{y}_i^{(t)}$ with acceptable confidence and **add** $(d_i, \widehat{y}_i^{(t)})$ into $\mathcal{D}_L^{(t)}$, otherwise
7: **add** d_i into $\mathcal{D}_U^{(t)}$
8: **End For**
9: **End for**

Pseudo-label initialization To incorporate the supervision information provided by seed words, we initialize pseudo-labels for each document using seed word occurrences. Since the selected seed words are representatives for categories, we suppose that a document containing more seed words of a category is more likely to be associated with this category. Following this, for each document *d* we compute the normalized seed word occurrence vector π_d as follows:

$$\pi_d(c_i) = \frac{SF_d(c_i) + \gamma}{\sum_{j=1}^{|\mathcal{C}|} SF_d(c_j) + |\mathcal{C}|\gamma} \tag{4}$$

where $SF_d(c_i)$ denotes the number of times that the seed words of category c_i have occurred in document *d*, and γ is a smoothing parameter used to avoid dividing by zero.[1] Then, the pseudo-label is initialized by the following rule:

$$\widehat{y}_d^{(0)} = \begin{cases} \infty & \text{if } \pi_d(c_1) = \pi_d(c_2), \cdots, = \pi_d(c_{|\mathcal{C}|}) \\ \underset{c_i \in \mathcal{C}}{\operatorname{argmax}}\, \pi_d(c_i) & \text{otherwise} \end{cases} \tag{5}$$

We do not assign a pseudo-label to document *d* without any seed word occurrence, denoted by ∞. We can obtain $\mathcal{D}_L^{(0)}$ and $\mathcal{D}_U^{(0)}$ by applying this initialization rule to all documents in $\mathcal{D}$.

Naive Bayes parameter θ update Given $\mathcal{D}_L^{(t-1)}$ and $\mathcal{D}_U^{(t-1)}$, the optimization of Eq.3 becomes a standard semi-supervised naive Bayes. We can use the EM algorithm to find the (local) optimum of $\theta^{(t)}$. This is achieved by iterating the following E-step and M-step. Due to the space limitation, we directly present the update equations without derivations.

In the **E-step**, we estimate the label posterior for each unlabeled document *d* using the Bayesian rule:

$$\Pr(y = c_i|d) = \frac{\Pr(y = c_i) \prod_{j=1}^{|\mathcal{W}|} \Pr(w_j|y = c_i)^{N_{d,w_j}}}{\sum_{c_h \in \mathcal{C}} \Pr(y = c_h) \prod_{j=1}^{|\mathcal{W}|} \Pr(w_j|y = c_h)^{N_{d,w_j}}} \tag{6}$$

[1] We empirically set γ to 0.01 in this work.

In the **M-step**, we update the naive Bayes parameter $\theta = \{\Pr(y), \Pr(w_j|y)\}$ with Laplace smoothing:

$$\Pr(y = c_i) = \frac{1 + \widehat{N}_{c_i} + \lambda \sum_{d \in \mathcal{D}_U} \Pr(y = c_i|d)}{|\mathcal{C}| + |\mathcal{D}_L| + \lambda |\mathcal{D}_U|} \tag{7}$$

$$\Pr(w_j|y = c_i) = \frac{1 + N_{c_i,w_j} + \lambda \sum_{d \in \mathcal{D}_U} N_{d,w_j} \Pr(y = c_i|d)}{|\mathcal{W}| + N_{c_i} + \lambda \sum_{d \in \mathcal{D}_U} \sum_{j=1}^{|\mathcal{W}|} N_{d,w_j} \Pr(y = c_i|d)} \tag{8}$$

where $\widehat{N}_{c_i}$ is the number of labeled documents marked with label c_i; N_{c_i,w_j} and N_{c_i} are the number of word w_j occurring and total number of words occurring in documents marked with label c_i, respectively. For clarity, we outline this EM procedure in *Algorithm 2*.

Algorithm 2 EM procedure outline for θ update

1: **Repeat**
2: **E-step**: Estimate the label posteriors for unlabeled documents using Eq.6
3: **M-step**: Update θ using Eqs.7 and 8
4: **Until convergence**

Pseudo-label $\widehat{y}$ update To update pseudo-labels $\widehat{y}^{(t)}$, for each document d we compute a label weight vector $\widehat{\pi}_d$ as follows[2]:

$$\widehat{\pi}_d(c_i) = \frac{\Pr^{(t)}(y = c_i|d) + \pi_d(c_i)}{2} \tag{9}$$

where $\Pr^{(t)}(y|d)$ is the label posterior distribution of document d estimated using the current $\theta^{(t)}$, and π_d the normalized seed word occurrence vector that has been shown in Eq.4. The pseudo-label is updated by the following rule:

$$\widehat{y}_d^{(t)} = \begin{cases} \underset{c_i \in \mathcal{C}}{\operatorname{argmax}}\, \widehat{\pi}_d(c_i) & \text{if } \max\limits_{c_i \in \mathcal{C}} \widehat{\pi}_d(c_i) > \delta \\ \infty & \text{otherwise} \end{cases} \tag{10}$$

where δ denotes an acceptable confidence threshold. This means that the dominate label in $\widehat{\pi}_d$ becomes the new pseudo-label for document d, if its weight in $\widehat{\pi}_d$ is larger than δ.

To avoid noisy pseudo-labels, we further consider the information of nearest neighboring documents. Following the intuition that the neighboring documents are more likely from a same category, we re-write the equation of $\widehat{\pi}_d$ by:

$$\widehat{\pi}_d(c_i) = \frac{\Pr^{(t)}(y = c_i|d) + \pi_d(c_i) + \sum_{j \in \Omega_d} \left(\Pr^{(t)}(y = c_i|j) + \pi_j(c_i) \right)}{2 \times (1 + |\Omega_d|)} \tag{11}$$

where Ω_d denotes the set of nearest neighboring documents for document d. In this work, we find Ω using the cosine similarity of TF-IDF representations, and empirically set $|\Omega_d|$ to 5.

4 Experiment

In this section, we empirically compare the proposed PL-DNB against both the existing dataless classification algorithms and traditional supervised algorithms.

4.1 Experimental Setup

Dataset and seed word set We employed two datasets of *Reuters*[3] and *Newsgroup*[4], which are usually used in text classification evaluations. In terms of *Reuters*, we left the 10 largest categories in the original

[2]Note that $\sum_{i=1}^{|\mathcal{C}|} \widehat{\pi}_d(c_i) = 1$

[3]http://kdd.ics.uci.edu/database/reuters21578/reuters21578.html

[4]http://qwone.com/∼jason/20Newsgroups/

Table 1: Statistics of datasets and sets of seed words. *#Train* / *#Test*: number of training/test documents; *#AvgDoc*: average length of documents; *#Label*: number of categories; *#Word*: number of unique words; S^L/S^D: average number of seed words in S^L/S^D.

Dataset	*#Train*	*#Test*	*#AvgDoc*	*#Label*	*#Word*	S^L	S^D
Reuters	5228	2057	70.8	10	7419	1	6
Newsgroup	11314	7532	152.1	20	52761	1	4.75

dataset, obtaining a dataset of 7285 documents in total. The training and test sets consist of 5228 and 2057 documents, respectively. Besides, the *Reuters* dataset is very imbalanced, where the largest category has 3713 documents but the smallest one has only 90 documents. In terms of *Newsgroup*, we used the "bydate" version, which contains totally 18846 documents divided into 20 categories. The training and test sets consist of 11314 and 7532 documents, respectively. The *Newsgroup* dataset is extremely balanced, where each category has about 900 documents. We apply traditional pre-processing methods to both datasets, i.e., removal of the standard stopwords, and the words that are shorter than 2 characters or occur in less than 5 documents.

Following (Chen et al., 2015; Li et al., 2016), we used two sets of seed words, denoted by S^L and S^D respectively. The seed words of S^L are directly selected from the category labels. We only left a single seed word for each category. The seed words of S^D are manually selected with the domain knowledge from additional candidate sets obtained by unsupervised learning algorithms. In contrast to S^L, the set of S^D contains more seed words.

For clarity, we show the statistics of the datasets and sets of seed words in Table 1.

Baseline algorithm To evaluate the effectiveness of PL-DNB, we compare it against four baseline algorithms, described as follows:

- **Seed word based naive Bayes (SNB)** (Liu et al., 2004): SNB is an early dataless naive Bayes algorithm that directly trains a classifier over the training documents with pseudo-labels in a semi-supervised manner. For a fair comparison, we omit the representative word selection step of SNB, but use the seed words of S^L and S^D just as other dataless algorithms

- **Seed-guided topic model (STM)** (Li et al., 2016): STM is a topic model based dataless text classification algorithm.[5] All crucial parameters of STM are tuned following the suggestions provided by its authors.

- **Naive Bayes (NB)**: We implement an in-house code of the supervised version of NB. The TF-IDF representation is used for documents.

- **Support vector machines (SVMs)**: We employ the *sklearn* tool[6] of SVMs with default parameter settings. The TF-IDF representation is also used.

Following the settings of (Li et al., 2016), we run dataless algorithms (i.e., SNB, STM and PL-DNB) over all documents, and only evaluate the classification results of test documents. For supervised algorithms (i.e., NB and SVMs), we train them over the training set, and also evaluate the results of test documents.

For the proposed PL-DNB, the numbers of the outer iteration and inner EM iteration are set to 10 and 5, respectively. The acceptable confidence threshold δ is set to 0.3. Besides, we empirically set the tuning parameter λ to 0 for *Reuters* and 0.3 for *Newsgroup*, respectively.

[5]This code is available at https://github.com/ly233/Seed-Guided-Topic-Model

[6]This tool is available at http://scikit-learn.org/stable/

Table 2: Evaluation results of Micro-F1. The superscript "‡" denotes that the best result of PL-DNB is statistically significant at 0.01 level.

Dataset	S^L			S^D			NB	SVMs
	PL-DNB	SNB	STM	PL-DNB	SNB	STM		
Reuters	0.847	0.633‡	0.819‡	0.895	0.826‡	**0.900**	0.921	0.973
Newsgroup	0.710	0.562‡	0.702‡	0.756	0.709‡	**0.766**	0.764	0.823

Table 3: Evaluation results of Macro-F1. The superscript "‡" denotes that the best result of PL-DNB is statistically significant at 0.01 level.

Dataset	S^L			S^D			NB	SVMs
	PL-DNB	SNB	STM	PL-DNB	SNB	STM		
Reuters	0.758	0.539‡	0.697‡	**0.832**	0.769‡	0.815‡	0.907	0.973
Newsgroup	0.670	0.508‡	0.668‡	**0.728**	0.661‡	0.724	0.753	0.823

4.2 Comparison of Baseline Algorithms

In the experiment, we use Micro-F1 and Macro-F1 to evaluate the classification performance. For each algorithm, we independently run it 10 times, and show the average results in Tables 2 and 3.

Comparison of dataless algorithms In contrast to SNB, we can observe that the proposed PL-DNB algorithm performs better in all settings. Overall, the performance gain of Micro-F1 is about 0.05∼0.21, and that of Macro-F1 is about 0.06∼0.22. More importantly, PL-DNB significantly outperforms SNB with the seed word set of S^L. The result indicates that PL-DNB can achieve high classification performance with very few seed words. This makes PL-DNB more practical for the text analysis tasks, where even seed words are expensive to obtain. Additionally, since the main difference between SNB and PL-DNB is that PL-DNB exploits nearest neighboring documents to improve the pseudo-label update, the performance gain over SNB indicates that this nearest neighbor scheme is effective.

In contrast to STM, we can observe that PL-DNB gets higher scores in most settings. In terms of S^L, the Micro-F1 and Macro-F1 scores of PL-DNB are higher than those of STM across both datasets. For example, the improvements of Micro-F1 and Macro-F1 over STM are about 0.03 and 0.06 on *Reuters*, respectively. This further supports the above observation that PL-DNB can perform well with a few seed words. In terms of S^D, PL-DNB is competitive with STM, where PL-DNB gets higher Macro-F1 scores but slightly lower Micro-F1 scores.

Comparing between the two dataless baselines, STM gains better performance in all settings. This result is consistent with the empirical results reported in (Li et al., 2016). The possible reason is that STM incorporates the category word probability to exploit the discriminative power of words, which can effectively improve classification results.

Comparison of supervised algorithms First, we can observe that PL-DNB is a little worse than NB, where the performance gap is not very obvious. For example, the Micro-F1 and Macro-F1 scores of PL-DNB+S^D are only about 0.01 and 0.025 lower than those of NB on *Newsgroup*. That is, PL-DNB may approach the supervised version of NB given sufficient seed words. Second, PL-DNB performs worse than SVMs, however it uses very limited supervision during classifier training. In some sense, our PL-DNB can be considered as an important candidate algorithm for text classification, especially when collecting labeled documents for training is extremely expensive.

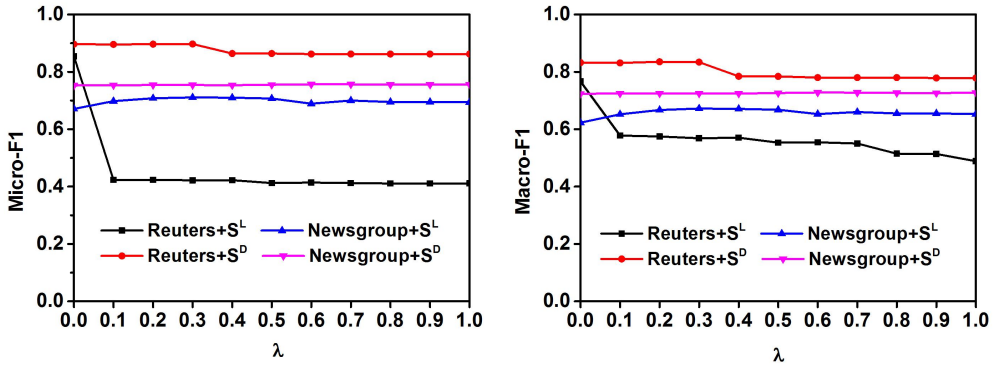

Figure 1: Evaluation on different values of λ

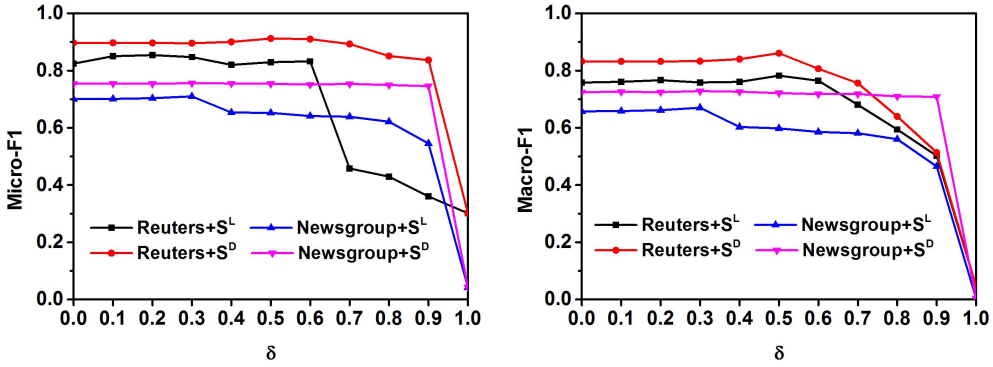

Figure 2: Evaluation on different values of δ

4.3 Evaluation of Parameters

In this section, we empirically evaluate two crucial parameters of PL-DNB, i.e., λ and δ.

Analysis of λ The parameter λ is used to control the importance of unlabeled documents during semi-supervised learning (ref. Eq.3). Note that $\lambda = 1$ means that unlabeled documents are equally important with labeled ones, and $\lambda = 0$ means that unlabeled documents are not utilized.

To evaluate the impact of λ, we examine the classification results of different values of $\lambda \in \{0, 0.1, 0.2 \cdots, 1\}$ when δ is fixed to 0.3. The experimental results of Micro-F1 and Macro-F1 are shown in Figure 1.

In terms of *Reuters*, we can observe that PL-DNB achieves higher scores when the value of λ is relatively smaller. Especially when the seed word set of S^L is utilized, the best scores are achieved at $\lambda = 0$, and significantly higher than scores of other λ values. The Micro-F1 score of $\lambda = 0$ is over 0.85, but those of other λ values are only about 0.4. Besides, the gap of Macro-F1 between $\lambda = 0$ and other values is about 0.2. Since S^L provides only one seed words for each category, a great majority of documents can not be initialized with pseudo-labels. At the beginning of optimization iterations, these unlabeled documents that are without any supervision prefer equally contributing to all categories, resulting in worse estimations of naive Bayes parameters. This problem grows worse for imbalanced datasets, because the smaller categories are much more sensitive to the noise. Additionally, smaller values of λ still perform better than larger ones when using the seed word set of S^D, but the gap is not obvious. The reason is that S^D can assign pseudo-labels to most of the documents. In summary, this evaluation result implies that for the imbalanced dataset PL-DNB prefers smaller λ values. We thus set $\lambda = 0$ for *Reuters* in our experiment.

In terms of *Newsgroup*, we can roughly observe that PL-DNB performs stable as λ varies. Overall, the

reason may be that *Newsgroup* is very balanced, so that PL-DNB is insensitive to unlabeled documents at the beginning of optimization iterations. For the setting of *Newsgroup*+S^D, both Micro-F1 and Macro-F1 scores are mostly the same with different values of λ. This is because only a small number of documents can not be initialized with pseudo-labels[7] by S^D. That is to say, λ only affects a few documents during the semi-supervised optimization step. Additionally, we see that the scores of PL-DNB are a bit higher when λ is in the interval $[0.1, 0.5]$ when using S^L. We empirically set $\lambda = 0.3$ for *Newsgroup* in our experiment.

Analysis of δ The parameter δ is a threshold used to determine whether a document can be marked with a pseudo-label in the pseudo-label update step (ref. Eq.10). On one hand, $\delta = 0$ means that all documents will be associated with pseudo-labels; On the other hand, a higher δ value implies that the updated pseudo-labels must be more "accurate".

We evaluate the classification results of different values of $\delta \in \{0, 0.1, 0.2 \cdots, 1\}$ by holding λ fixed to 0 for *Reuters* and 0.3 for *Newsgroup* respectively. The experimental results of Micro-F1 and Macro-F1 are shown in Figure 2.

Overall, we can observe that the classification performance is fast dropped as δ becomes larger. For example, the Micro-F1 score drops from 0.8 to 0.4 across *Reuters*+S^L when δ becomes larger; and especially all scores of $\delta = 1$ are very poor. The main reason of this trend is that the number of documents with pseudo-labels decreases as δ becomes larger. That is, less labeled documents can be used for training naive Bayes (ref. Eq.3). This result indicates that in the context of dataless classification, the number of labeled documents may be a bit more important then the quality of pseudo-labels.

In terms of *Reuters*, PL-DNB performs relatively stable as $\delta \in [0.1, 0.6]$, but worse when δ becomes larger than 0.6. In contrast, the performance of *Newsgroup* gets stable before δ achieves 0.9, especially when the set of S^D is utilized. This comparison indicates that the balanced dataset may be less sensitive to the value of δ than the imbalanced dataset. Additionally, the best scores of *Reuters* are achieved at $\delta = 0.5$ in most settings, but it is only a bit better than $\delta = 0.1, 0.2, 0.3, 0.4, 0.6$. The best scores of *Newsgroup* are mostly at $\delta = 0.3$. Empirically, we set δ to 0.3 for both datasets in our experiment.

5 Conclusion

In this paper, we develop a novel dataless PL-DNB classification algorithm, which can be directly trained on unlabeled documents with seed words, instead of labeled documents. In PL-DNB, we create pseudo-labels for documents and train a naive Bayes classifier over the pseudo training set in semi-supervised learning. The pseudo-labels are iteratively updated, and the information of nearest neighboring documents is considered to avoid noisy pseudo-labels. We examine the effective of PL-DNB on two popular datasets in classification evaluations, where one is imbalanced and the other is balanced. Empirical results indicate that PL-DNB performs better than the state-of-the-art dataless text classifiers, especially for the imbalanced dataset. Specially, PL-DNB achieves competitive performance with supervised naive Bayes in some settings.

Acknowledgements

We would like to acknowledge support for this project from the National Natural Science Foundation of China (NSFC) (grant numbers 61602204, 61572226 and 61472157).

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Gavin C. Cawley and Nicola L. C. Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107.

[7] In the current setting, only about 23% documents have no initialized pseudo-labels.

Ming-Wei Chang, Lev Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: dataless classification. In *AAAI Conference on Artificial Intelligence*, pages 830–835.

Xingyuan Chen, Yunqing Xia, Peng Jin1, and John Carroll. 2015. Dataless text classification with descriptive LDA. In *AAAI Conference on Artificial Intelligence*, pages 2224–2231.

Doug Downey and Oren Etzioni. 2008. Look ma, no hands: analyzing the monotonic feature abstraction for text classification. In *Neural Information Processing Systems*, pages 393–400.

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*, pages 595–602.

Mohamed Elhoseiny, Babak Saleh, and Ahmed M. Elgammal. 2013. Write a classifier: Zero-shot learning using purely textual descriptions. In *IEEE International Conference on Computer Vision*.

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*, pages 385–394.

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*, pages 877–880.

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*, pages 1865–1871.

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*.

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*, pages 85–94.

Bing Liu, Xiaoli Li, Wee Sun Lee, , and Philip S. Yu. 2004. Text classification by labeling words. In *AAAI Conference on Artificial Intelligence*, pages 425–430.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2008. Distant supervision for relation extraction without labeled data. In *Annual Meeting of the Association for Computational Linguistics*, pages 1003–1011.

Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, and Tom Mitchell. 2000. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3):103–134.

Mark Palatucci, Dean Pomerleau, Geoffrey Hinton, and Tom M. Mitchell. 2009. Zero-shot learning with semantic output codes. In *Neural Information Processing Systems*, pages 1410–1418.

Hema Raghavan, Omid Madani, and Rosie Jones. 2006. Active learning with feedback on features and instances. *Journal of Machine Learning Research*, 7:1655–1686.

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In *AAAI Conference on Artificial Intelligence*.

Limin Yao, David Mimno, and Andrew McCallum. 2009. Efficient methods for topic model inference on streaming document collections. In *International Conference on Knowledge Discovery and Data Mining*.

Li Zhao, Minlie Huang, Ziyu Yao, Rongwei Su, Yingying Jiang, and Xiaoyan Zhu. 2016. Semi-supervised multinomial naive bayes for text classification by leveragingword-level statistical constraint. In *AAAI Conference on Artificial Intelligence*.