Machine Learning (2021) 110:907–931
https://doi.org/10.1007/s10994-021-05967-y

Robust supervised topic models under label noise

Wei Wang[1,2,3] · Bing Guo[1] · Yan Shen[4] · Han Yang[2] · Yaosen Chen[1] · Xinhua Suo[1]

Received: 12 July 2020 / Revised: 26 February 2021 / Accepted: 3 March 2021 / Published online: 14 April 2021

Abstract
Recently, some statistical topic modeling approaches have been widely applied in the field of supervised document classification. However, there are few researches on these approaches under label noise, which widely exists in real-world applications. For example, many large-scale datasets are collected from websites or annotated by varying quality human-workers, and then have a few mislabeled items. In this paper, we propose two robust topic models for document classification problems: Smoothed Labeled LDA (SL-LDA) and Adaptive Labeled LDA (AL-LDA). SL-LDA is an extension of Labeled LDA (L-LDA), which is a classical supervised topic model. The proposed model overcomes the shortcoming of L-LDA, i.e., overfitting on noisy labels, through Dirichlet smoothing. AL-LDA is an iterative optimization framework based on SL-LDA. At each iterative procedure, we update the Dirichlet prior, which incorporates the observed labels, by a concise algorithm based on *maximizing entropy* and *minimizing cross-entropy* principles. This method avoids identifying the noisy label, which is a common difficulty existing in label noise cleaning algorithms. Quantitative experimental results on *noisy completely at random* (NCAR) and *Multiple Noisy Sources* (MNS) settings demonstrate our models have outstanding performance under noisy labels. Specially, the proposed AL-LDA has significant advantages relative to the state-of-the-art topic modeling approaches under massive label noise.

Keywords Supervised topic modeling · Document classification · Label noise

Editor: Derek Greene.

✉ Bing Guo
guobing@scu.edu.cn

Wei Wang
wong.wei@163.com

[1] College of Computer Science, Sichuan University, Chengdu, China

[2] Chengdu Sobey Digital Technology Co., Ltd., Chengdu, China

[3] Peng Cheng Laboratory, Shenzhen, China

[4] School of Computer Science, Chengdu University of Information Technology, Chengdu, China

1 Introduction

Much of texts is annotated with human interpretable labels; how to use these labels is an important technique with the digital text in Web growing explosively. Recently, some statistical topic modeling approaches, e.g., Latent Dirichlet allocation (LDA) (Blei et al. 2003), as interpretable, flexible, and easily extensible approaches have been widely applied in the field of document classification (Burkhardt and Kramer 2019). However, standard LDA is a completely unsupervised algorithm, and then how to incorporate prior labels into the topic modeling procedure is a popular research direction. To incorporate the prior information in the generative process, there are two kinds of approaches: one first generates the words, and then generates the response variables, i.e., the prior information, conditioned on the word space; the other generates the prior knowledge first, i.e., incorporates the prior side information, and then generates the words conditioned on them. Roughly speaking, in the second approach, the side information has more influence on the modeling procedure, and the model has better predictive ability (Soleimani and Miller 2019). Meanwhile, the first type of approaches is often designed for single label classification or regression (Blei and McAuliffe 2010; Li et al. 2018; Lacoste-Julien et al. 2008; Magnusson et al. 2016; Zhu et al. 2012), and the second can support multi-label classification.

To the best of our knowledge, Labeled LDA (L-LDA) introduced by Ramage et al. (2009) is the first supervised LDA model on multi-label document classification. It incorporates prior information by the second approach, i.e., simply defines a one-to-one correspondence between topics and observed labels, and then incorporates the observed label information by the document-topic distribution Dirichlet prior. L-LDA has been widely applied for efficiency and concision. However, it constrains the topic distributions in the observed labels that lead to over-focus on them. So noisy labels, which widely exist, would worsen the performance of L-LDA (Li et al. 2015b; Ramage et al. 2011).

Label noise is an unavoidable phenomenon in real-world applications. It may be generated by machines or non-expert annotators. These days label noise in training data is more prevalent as many datasets are annotated through crowd-sourcing (Kumar et al. 2020). Handling label noise is an important and challenging problem. There are two main approaches in the literatures of solving the problem: label noise cleaning algorithms and label noise robust algorithms (Zhang et al. 2019). However, there are few literatures that focus on topic models under label noise. In this paper, we propose robust topic modeling approaches for document classification under label noise. Firstly, we make L-LDA robust by easing overfitting on noisy labels. Label Smoothing Regularization (LSR) (Szegedy et al. 2016) is a well known technique using soft labels in place of one-hot labels to handle the issue. Following this method, we introduce Smoothed Labeled LDA (SL-LDA) building on Dirichlet smoothing. Secondly, we aim to further improve the model robustness by using data cleaning approaches. However, it is well known that data cleaning approaches suffer from a chicken-and-egg dilemma (Angelova et al. 2005), since good classifier depends on high quality datasets, but high quality datasets need a good classification filter. Sample reweighting approaches have been applied as a data cleaning strategy, which inclines to suppress noisy labels by small weights. However, the weighting function is often manually set that leads to poor generalizability (Shu et al. 2019), or trained by a small unbiased validation set, i.e., meta-data, which is actually uncommon to construct (Ren et al. 2018). Mikalsen et al. (2019) propose a graph-based label propagation method that can deal with noisy data by iterative computing soft labels, and allowing the labeled data to change labels during the propagation. Tanaka et al. (2018) propose joint optimization framework of learning deep

neural network (DNN) parameters and estimating true labels. Inspired by (Mikalsen et al. 2019) and (Tanaka et al. 2018), we propose Adaptive Labeled LDA (AL-LDA), which is an iterative optimization framework based on SL-LDA. To reduce the influence of noisy labels, we update the model prior probabilities, i.e., soft labels, at each iteration by a concise optimization method based on two principles, i.e., *maximizing entropy* and *minimizing cross-entropy*. We also introduce convergence conditions to avoid overfitting on noisy training labels. This method does not need to identify the noisy label, and then avoid the chicken-and-egg dilemma. Meanwhile, the proposed method does not need the meta-data to train the weighting function used in sample reweighting approaches. In addition, the proposed method is different from common label noise-tolerant algorithms, which need modeling label noise that often leads to model complexity (Frénay and Verleysen 2013).

Our contribution is summarized as follows. Firstly we propose a supervised topic model, i.e., SL-LDA, which is an extension of L-LDA, and eases its overfitting problem by Dirichlet smoothing. Then we extend the model to be more robust under label noise by an iterative optimization framework named AL-LDA, which has a novel optimization algorithm that avoids identifying noisy labels or constructing meta-data, and convergence conditions that avoid overtraining. We evaluate the models on *noisy completely at random* (NCAR) datasets, including single-label and multi-label collections. The experiments show the proposed models have excellent performance under massive label noise. We also evaluate the models on the document classification scenario on *Multiple Noisy Sources* (MNS) settings, where the results show our models have better performance than state-of-the-art works.

The rest of the paper is structured as follows. Section 2 reviews the related work; Sect. 3 describes the proposed methods, including SL-LDA and AL-LDA; Sect. 4 introduces the experiments and evaluation results. We discuss the results in Sect. 5. Finally, Sect. 6 gives concluding remarks and an outline of future work.

2 Related work

LDA (Blei et al. 2003) is a hierarchical Bayesian model that aims to map a text document into a latent low dimensional space based on a set of automatically learned topics. The model considers each document as random mixtures over topics, and each topic is a distribution over words. Given the number of topics K, a collection of documents D and the size of the vocabulary V, word generation is defined by the conditional distributions $P(w_n = w|z_n = t)$, denoted by the matrix $\Phi(K \times V)$. Similarly, topic generation is defined by the conditional distributions $P(z_n = t|d_n = d)$, denoted by the matrix $\Theta(D \times K)$. The joint probability of a corpus $\mathcal{W}$ and corresponding topics $\mathcal{Z}$ is $P(\mathcal{W}, \mathcal{Z}|\Phi, \Theta) = \prod_w \prod_t \prod_d \phi_{w|t}^{N_{w|t}} \theta_{t|d}^{N_{t|d}}$, where $N_{w|t}$ is the number of times that word w generated by topic t, and $N_{t|d}$ is the number of times that topic t in document d. The model places a Dirichlet prior $\beta \boldsymbol{n}$ over Φ, i.e., $P(\Phi|\beta \boldsymbol{n}) = \prod_t Dir(\boldsymbol{\phi}_t|\beta \boldsymbol{n})$, and another Dirichlet prior $\alpha \boldsymbol{m}$ over Θ, i.e., $P(\Theta|\alpha \boldsymbol{m}) = \prod_d Dir(\boldsymbol{\theta}_d|\alpha \boldsymbol{m})$.

For application in the field of document classification, several modifications of LDA to incorporate observed labels have been proposed by the second approach introduced in Sect. 1, i.e., incorporate the prior information first, and then generate the words conditioned on them (Burkhardt and Kramer 2018; Li et al. 2015a, b; Padmanabhan et al. 2017; Rubin et al. 2011; Ramage et al. 2009, 2011; Wang et al. 2020; Zhang et al. 2017). Labeled LDA (L-LDA) (Ramage et al. 2009) is an earlier supervised LDA model for multi-label

document classification. It simply constrains the topic distributions to pre-assigned labels that lead to over-focus on them. To overcome the problems of L-LDA, there are two kinds of improved approaches (Wang et al. 2020). One relaxes the topic sampling restriction of the document pre-assigned labels to avoid over-focus on them (Li et al. 2015b; Padmanabhan et al. 2017; Wang et al. 2020; Zhang et al. 2017). Label smoothing is also effective as a means of these methods, which prevent overconfidence on any one label. The other assumes the existence of another topic layer and aims to establish the relation between topics and prior labels (Burkhardt and Kramer 2018; Li et al. 2015a; Rubin et al. 2011; Ramage et al. 2011). Dependency-LDA (Rubin et al. 2011) incorporates another topic model to model the observed label correlations, which are deemed to be crucial for multi-label classifiers (Burkhardt and Kramer 2019). Consequently, Dependency-LDA is competitive with state-of-the-art discriminative methods, e.g., SVM, and is often selected as the baseline in related works (Burkhardt and Kramer 2018; Li et al. 2015a, b; Wang et al. 2020).

There have been many approaches proposed for robust learning of classifiers under label noise (Frénay and Verleysen 2013). Some data processing methods rely on cleaning noisy samples from training data (Brodley and Friedl 1999; Jeatrakul et al. 2010; Sun et al. 2007). However, they often suffer from a chicken-and-egg dilemma, since noisy sample classifier are hard to train from coarsely labeled dataset. Another strategy of data cleaning is sample reweighting approaches, which design a weighting function mapping from training loss to sample weight, since incorrect label often have large loss values (De La Torre and Black 2003; Jiang et al. 2014; Zhang and Sabuncu 2018). These methods often need to manually set a specific form of weighting function or hyper-parameters, which raise their application difficulty. Meta-data set (i.e., with clean labels and balanced data distribution) learning is an effective method to get weighting functions (Shu et al. 2019; Veit et al. 2017). However, the difficulty of getting meta-data limits its application. In addition, there are also some approaches that modify the existing algorithms to be robust to label noise (Boutell et al. 2004; Biggio et al. 2011; Khardon and Wachman 2007; Manwani and Sastry 2013). In particular, deep neural networks (DNNs) can easily overfit to noisy labels, so making them robust is a popular research direction these days (Ghosh et al. 2017; Li et al. 2019; Patrini et al. 2017; Tanaka et al. 2018). Ghosh et al. (2017) present analytical results on the noise-tolerance of loss functions in a multi-class classification scenario, and derive corresponding sufficient conditions. Patrini et al. (2017) introduce a loss correction approach that is robust to label noise. Meta-learning based noise-tolerant (MLNT) (Li et al. 2019) is a noise-tolerant training algorithm that can be theoretically applied to any model trained with the gradient-based rule. The joint optimization framework for training DNNs on noisy labeled datasets (Tanaka et al. 2018) obtains clean labels by updating them using the soft-label or hard-label method, which demonstrates state-of-the-art performance in image classification. NMLSDR proposed by Mikalsen et al. (2019) cleans noisy multi-labels and labels unlabeled data simultaneously. The method first constructs a neighbourhood graph, and then designs a label propagation algorithm for unlabeled and mislabeled data. However, the graph construction process is time-consuming.

In the experiments of these literatures, most studies use *noisy completely at random*(NCAR) label noise, i.e., randomly selecting instances and changing them to the other remaining labels (Golzari et al. 2009). Some literatures study the scenario where several heterogeneous annotators with varying qualities provide the labels, i.e., *Multiple Noisy Sources* (MNS). For simplicity, the experiments often assume a single coin model for annotators and also that the annotator qualities are independent of the class (Raykar et al. 2010).

Table 1 Notation descriptions

Notation	Description
K	Number of topics/labels
D	Number of documents
V	Number of words
$\mathcal{W}$	The corpus
$\mathcal{Z}$	The assigned topics of corpus $\mathcal{W}$
Θ	The matrix of document - topic/label distributions
Φ	The matrix of topic/label - word distributions
β	The concentration parameter of Dirichlet prior for topic/label - word distributions
$\boldsymbol{n}$	The base measure of Dirichlet prior for topic/label - word distributions
α	The concentration parameter of Dirichlet prior for document - topic/label distributions
$\mathbf{m}$	The base measure of document-specific Dirichlet prior for document - topic/label distributions
γ	The concentration parameter of Dirichlet prior for document-specific $\mathbf{m}$
$\boldsymbol{\mu}$	The uniform base measure of Dirichlet prior for document-specific $\mathbf{m}$
$\boldsymbol{\Lambda}$	The label presence/absence indicator

There are few supervised topic modeling approaches for robust classification with noisy labels. MRTM (Multiple Relational Topic Modeling) introduced by Liu et al. (2018) explores latent topics for noisy short texts in social networks. To our knowledge, ML-PA-LDA-MNS proposed by Padmanabhan et al. (2017) is the first supervised multi-label classification topic model that specially considers the presence of label noise, and aims to model it. It assumes the latent topic is generated by the document labels as well as absence labels, and for the multiple noisy sources, adopts a single coin model, which is suggested can learn the annotator qualities well. However, the reported experiments with artificial *Multiple Noisy Sources* (MNS) settings do not demonstrate the model has better performance than ML-PA-LDA, which is the basic of ML-PA-LDA-MNS without modeling the annotator qualities. In addition, the additional parameters of ML-PA-LDA-MNS for label noise increase the model complexity.

3 The proposed method

Firstly, we review the L-LDA model and introduce the Smoothed L-LDA (SL-LDA), then propose a novel optimization framework for supervised topic model with noisy labels, i.e., Adaptive Labeled LDA (AL-LDA). Lastly, building on the discussion of the proposed optimization algorithm, the convergence condition of AL-LDA is introduced. We summarize some important notations in Table 1.

3.1 SL-LDA

For the supervised extension of LDA, L-LDA defines a one-to-one correspondence between topics and labels. Given the number of labels K, a vector of binary label presence/absence indicator $\boldsymbol{\Lambda}_d = [l_1 l_2 ... l_K]^T, l_k \in \{0, 1\}$ is defined. To make this model fully generative, l_k can be generated by a Bernoulli distribution with a prior probability η_k. To incorporate label

information, L-LDA places each document assigned the same labels ($\mathbf{\Lambda}_d$) a same asymmetric Dirichlet prior $\mathbf{m}_d$. The distribution over $\boldsymbol{\theta}_d$ is given by

$$P(\boldsymbol{\theta}_d|\alpha\mathbf{m}_d) = Dirichlet(\boldsymbol{\theta}_d|\alpha\mathbf{m}_d),$$

with

$$m_{dt} = \frac{\Lambda_{dt}}{\sum_{t=1}^{K}\Lambda_{dt}}.$$

To overcome the shortcoming of L-LDA, i.e., over-focus on the pre-assigned labels, we add hidden topics besides the observed labels, and place Dirichlet smoothing on the document-specific $\mathbf{m}_d$, so

$$m_{dt} = \frac{\Lambda_{dt} + \gamma\mu_t}{\sum_{t=1}^{K+K_h}\Lambda_{dt} + \gamma}, \tag{1}$$

where K_h is the number of hidden topics, $\boldsymbol{\mu}$ is a uniform base measure, and γ is the concentration parameter. To incorporate the observed label frequencies, we just modify Equation (1) as

$$m_{dt} = \frac{\Lambda_{dt}N_t + \gamma\mu_t}{\sum_{t=1}^{K+K_h}\Lambda_{dt}N_t + \gamma}, \tag{2}$$

where N_t is the number of topic t in training data. This model is named Smoothed L-LDA (SL-LDA), which as a generative process is summarized as Algorithm 1.

Algorithm 1 Generative process of Smoothed L-LDA

$K = K + K_h$
for topic $t \in [1, K]$ **do**
  Generate $\boldsymbol{\phi} = \{\phi_w\}_{w=1}^{V} \sim Dirichlet(\cdot|\beta\boldsymbol{n})$
end for
for document $d \in [1, D]$ **do**
  for topic $t \in [1, K]$ **do**
    Generate $\Lambda_{dt} \sim Bernoulli(\eta_t)$
  end for
  Generate $\boldsymbol{m} = \{m_t\}_{t=1}^{K} \sim Dirichlet(\cdot|\gamma\boldsymbol{\mu}, \mathbf{\Lambda}_d, \{N_t\}_{t=1}^{K})$
  Generate $\boldsymbol{\theta} = \{\theta_t\}_{t=1}^{K} \sim Dirichlet(\cdot|\alpha\boldsymbol{m})$
  for word $n \in [1, N]$ **do**
    Sample topic $z_{dn} \sim Multinomial(\boldsymbol{\theta})$
    Sample word $w_{dn} \sim Multinomial(\boldsymbol{\phi}_{z_{dn}})$
  end for
end for

To learn the model, we use Collapsed Gibbs sampling (Griffiths and Steyvers 2004), which is a widely used approach to estimate LDA distributions. The Gibbs sampler is collapsed because the variables $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_t$ are analytically integrated out, and only the latent topic variables $z_n^{(d)}$ for the word $w_n^{(d)}$ are iteratively sampled based on the probability

$$P(z_n^{(d)}|\mathcal{W}, \mathcal{Z}_{\backslash d,n}, \alpha\boldsymbol{m}, \beta\boldsymbol{n}) \propto \frac{N^{\backslash d,n}_{w_n^{(d)}|z_n^{(d)}} + \beta n_w}{N^{\backslash d,n}_{\cdot|z_n^{(d)}} + \beta} \times \frac{N^{\backslash d,n}_{z_n^{(d)}|d} + \alpha m_t}{N_{\cdot|d} - 1 + \alpha}, \tag{3}$$

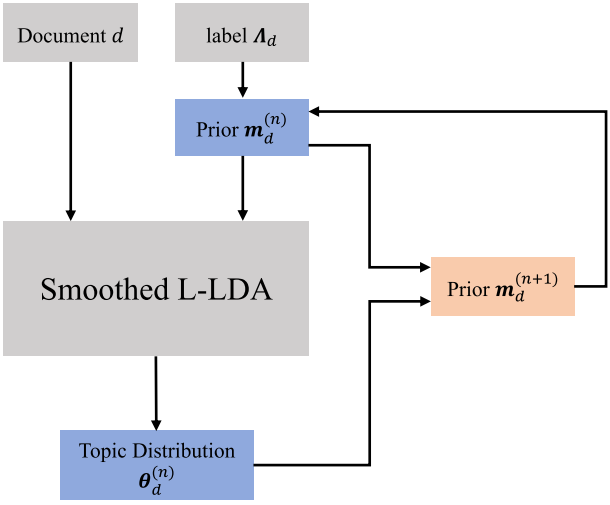

Fig. 1 The concept of the proposed optimization framework

where $N^{\backslash d,n}_{w_n^{(d)}|z_n^{(d)}}$ is the counts of $w_n^{(d)}$ in topic $z_n^{(d)}$ that dose not include the current assignment of the topic $z_n^{(d)}$ for $w_n^{(d)}$, $N^{\backslash d,n}_{\cdot|z_n^{(d)}}$ is the total number of times that any word has been generated by topic $z_n^{(d)}$ excluding the current assignment, $N^{\backslash d,n}_{z_n^{(d)}|d}$ is the number of times that topic $z_n^{(d)}$ in document d excluding the current assignment, and $N_{\cdot|d}$ is the total number of topics in document d. While training SL-LDA, we replace m_t with m_{dt} calculated by Eq. (2).

3.2 AL-LDA

To improve the performance of our model under label noise, an iterative optimization procedure is introduced. Figure 1 shows the concept of our proposal. Each document d and prior label indicator $\boldsymbol{\Lambda}_d$, which is incorporated by the Dirichlet prior $\boldsymbol{m}_d$, are observed. For each iterative procedure, we use Gibbs sampling to learn the proposed SL-LDA, which is a Bayesian inference model, where $\boldsymbol{m}_d$ is the prior probability of the document d - topic distribution, and $\boldsymbol{\theta}_d$, i.e., the dth row of matrix Θ, is the posterior probability of the same one. Thus, we suggest the divergence from $\boldsymbol{m}_d$ to $\boldsymbol{\theta}_d$ indicates the label qualities. In other words, the divergence is high for noisy labels and low for clean labels (Tanaka et al. 2018). To improve the quality of training data, we consider updating prior probability $\boldsymbol{m}_d^{(n+1)}$ by using the divergence from $\boldsymbol{m}_d^{(n)}$ to $\boldsymbol{\theta}_d^{(n)}$ of the nth iteration. After several iterations, the influence of noisy labels would be reduced.

To update prior probabilities, we propose a novel optimization algorithm to reduce the influence of noisy samples. Depending on Eq. (2), the asymmetric Dirichlet prior $\boldsymbol{m}_d$ for each document-topic distribution incorporates the prior labels of the instance d. In the context of noisy labels, we aim to eliminate their effects. Supposing we can affirm a noisy sample, we should replace $\boldsymbol{m}_d$ with symmetric Dirichlet prior without any information, i.e., clean the noisy labels. This would be like increasing the uncertainty of prior topic probability distributions, i.e., increase the entropy of Dirichlet prior $\boldsymbol{m}_d$, defined as

$$H(\boldsymbol{m}_d) = -\sum_t m_{dt} \log m_{dt}.$$

This idea follows the principle of *Maximum-Entropy*, proposed by Jaynes (1957).

On the contrary, if the training data have high quality labels and the model is well learned, the divergence from prior probability, which incorporates the observed labels in our model, to posterior probability of each document would be low. At this time, the optimal $\boldsymbol{m}_d^{(n+1)}$ would be close to $\boldsymbol{\theta}_d^{(n)}$ and $\boldsymbol{m}_d^{(n)}$. Consequently, we suggest the divergences from the optimal $\boldsymbol{m}_d^{(n+1)}$ to $\boldsymbol{\theta}_d^{(n)}$ as well as from $\boldsymbol{m}_d^{(n+1)}$ to $\boldsymbol{m}_d^{(n)}$ would be as small as possible. To express conveniently, we replace $\boldsymbol{m}_d^{(n+1)}$ with $\boldsymbol{m}_d$, $\boldsymbol{\theta}_d^{(n)}$ with $\boldsymbol{\theta}_d$, and $\boldsymbol{m}_d^{(n)}$ with $\boldsymbol{m}_d^{(old)}$. Measuring the divergence by *cross-entropy*, i.e.,

$$H(\boldsymbol{m}_d, \boldsymbol{\theta}_d) = -\sum_t m_{dt} \log \theta_{dt},$$
$$H(\boldsymbol{m}_d, \boldsymbol{m}_d^{old}) = -\sum_t m_{dt} \log m_{dt}^{old},$$

we can apply the principle of *minimizing cross-entropy* on the $\boldsymbol{m}_d$ and $\boldsymbol{\theta}_d$ as well as $\boldsymbol{m}_d$ and $\boldsymbol{m}_d^{old}$.

Intuitively, we get

$$\boldsymbol{m}_d^* = \arg\min_{\boldsymbol{m}_d}(-H(\boldsymbol{m}_d) + H(\boldsymbol{m}_d, \boldsymbol{\theta}_d) + H(\boldsymbol{m}_d, \boldsymbol{m}_d^{old})), \quad s.t. \quad \sum_t m_{dt} = 1. \tag{4}$$

We define the Lagrangian function as

$$\mathcal{L}(\boldsymbol{m}_d, \lambda) = \sum_t m_{dt} \log m_{dt} - \sum_t m_{dt} \log \theta_{dt} - \sum_t m_{dt} \log m_{dt}^{old} + \lambda(\sum_t m_{dt} - 1),$$

where λ is the Lagrangian multiplier. Clearly, setting the derivative of $\mathcal{L}$ with respect to m_{dt} and λ to zero, the result can be expressed conveniently by replacing m_{dt}^{old} with m_{dt} as follow,

$$m_{dt}^* = \frac{\theta_{dt} m_{dt}}{\sum_t \theta_{dt} m_{dt}}. \tag{5}$$

3.3 Discussion of the optimization algorithm

To investigate the effects of the optimization algorithm, Eq. (5), we introduce a simple example with two labels, i.e., L0 and L1. Figure 2a (left) is a noisy sample with the wrong label L0. Supposing the proposed model gives the right posterior probabilities in Fig. 2a (center), the updating prior probabilities are $m_{d0}^{(n+1)} = 0.5$ and $m_{d1}^{(n+1)} = 0.5$ (Fig. 2a, right) by Eq. (5). In other words, the noisy label is cleaned. Figure 2b demonstrates a right label sample is enhanced by a right model predictive label. In summary, if the model gives right posterior probabilities, the proposed algorithm could reduce the influence of noisy labels or enhance the right labels.

On the contrary, if the model mislabeled a document, the optimization algorithm would clean the right original label or even replace it with a wrong label. Figure 3 demonstrates an example. An instance with the right label L0 (Fig. 3a, left) is mislabeled by the model with the wrong label L1 (Fig. 3a, center). Depending on Eq. (5), the right label L0 is cleaned

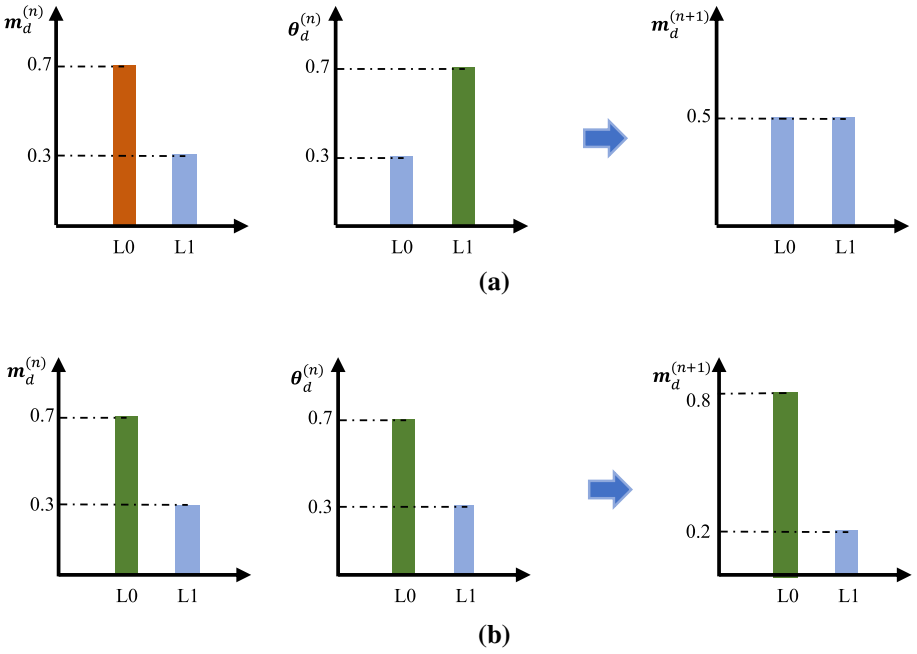

Fig. 2 Positive effects of AL-LDA optimization algorithm. Supposing a two class classification task, **a** A wrong prior label (orange) cleaned by a right predictive label (green) leads to no labels (blue). **b** A right prior label (green) is enhanced by a right predictive label (green) (Color figure online)

(Fig. 3a, right). Furthermore, if the effect of wrong posterior probabilities was greater than the right prior probabilities in Eq. (5), the right label L0 would be replaced with the wrong label L1 (Fig. 3b).

In general, it can be seen that in Equation (5), if $\theta_{dt} \geq \mathbb{E}_{\boldsymbol{m}}(\boldsymbol{\theta}_d)$, then $m^*_{dt} \geq m_{dt}$, and vice versa. So right posterior probabilities $\boldsymbol{\theta}_d$ are important for the proposed algorithm. It would decrease the noisy label probability ($\theta_{dt} < \mathbb{E}_{\boldsymbol{m}}(\boldsymbol{\theta}_d)$) or increase the right label probability ($\theta_{dt} > \mathbb{E}_{\boldsymbol{m}}(\boldsymbol{\theta}_d)$). At this time, the proposed algorithm is an effective and unified approach that optimizes the Dirichlet prior on both high quality labeled and mislabeled documents. However, wrong predictive labels could degrade the model. This finding suggests the model performance maybe stop improving after some point, or even getting worse. Consequently, we need to find the exact point to stop iteration.

3.4 The convergence condition of AL-LDA

To judge the convergence of AL-LDA, i.e., the exact stop point, we use the average Kullback-Leibler (KL) - divergence from $\boldsymbol{\theta}_d$ to original $\boldsymbol{m}_d^{(0)}$ as follow,

$$\overline{D}_{KL}(\mathcal{W}) = \frac{1}{D} \sum_d \sum_t \theta_{dt} \log \frac{\theta_{dt}}{m^{(0)}_{dt}}, \tag{6}$$

where D is the number of training documents. The reason of using average is avoiding the influence of the training data size. Generally, with respect to supervised learning, the

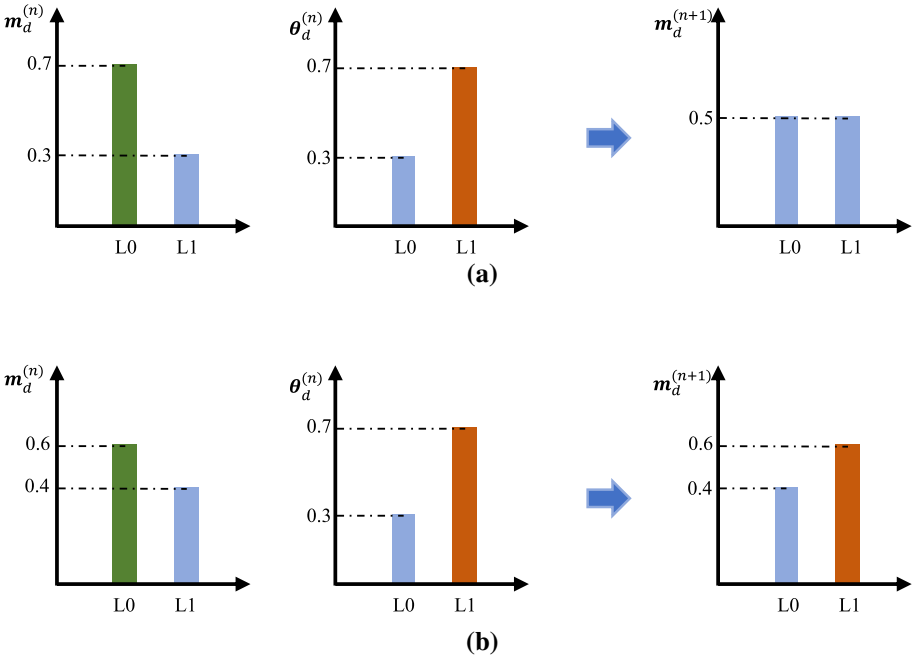

Fig. 3 Negative effects AL-LDA optimization algorithm. Supposing a two class classification task, **a** A right label (green) cleaned by a misjudged label (orange) leads to no labels (blue). **b** A right label (green) is replaced with a wrong label (orange) (Color figure online)

optimization problem is minimizing the divergence from ground-truth labels to the predictive labels, i.e., from $\boldsymbol{m}_d^{(0)}$ to $\boldsymbol{\theta}_d$ in the proposed model. It works well on clean labels; however, its performance decreases on coarsely labeled data, which lead to overfitting (Ying 2019) and model degradation introduced in Sect. 3.3 . To alleviate these issues, *early stopping* is employed. Inspired by the stopping criteria introduced by Prechelt (1998), we define two convergence conditions of the iterative optimization. One is the ratio between the current average KL-divergence $\overline{D}_{KL}^{(n)}$ and the first average KL-divergence $\overline{D}_{KL}^{(1)}$ as follow,

$$R_{\overline{D}_{KL}} = \frac{\overline{D}_{KL}^{(n)}}{\overline{D}_{KL}^{(1)}}. \tag{7}$$

If $R_{\overline{D}_{KL}}$ is less than a threshold $R^*_{\overline{D}_{KL}}$, which indicates there is a risk overfitting on noisy labels, the proposed model meets the convergence condition. The other is the difference between the current average KL-divergence $\overline{D}_{KL}^{(n)}$ and the next iteration average KL-divergence $\overline{D}_{KL}^{(n+1)}$ as follow,

$$\Delta_{\overline{D}_{KL}} = \overline{D}_{KL}^{(n)} - \overline{D}_{KL}^{(n+1)}. \tag{8}$$

If $\Delta_{\overline{D}_{KL}}$ is less than a threshold $\Delta^*_{\overline{D}_{KL}}$, which indicates we cannot optimize the model or even reduce the model performance, the model also meets the convergence condition. In

Section 4, we will introduce the suggested parameter values of the convergence condition, which works well on all experiments.

The algorithm of AL-LDA is summarized as Algorithm 2.

Algorithm 2 Inference algorithm of AL-LDA

```
for document d ∈ [1, D] do
    Calculate m_d using Equation (2)
end for
while not convergence for the model, judged by D̄_KL using Equation(7)(8) do
    while not convergence for each iterative procedure do
        for document d ∈ [1, D] do
            for word n ∈ [1, N] do
                Sample topic z_dn using Equation (3)
            end for
        end for
    end while
    Read out parameter set Θ
    for document d ∈ [1, D] do
        Calculate m*_d using Equation (5)
    end for
end while
```

4 Experiments

In this section, we evaluate SL-LDA and AL-LDA on several popular typical document classification tasks under label noise. Firstly, we introduce the collections and metrics, then the parameter settings are introduced. Thirdly, we list the results of our models and compared supervised topic models under NCAR and MNS settings. Fourthly, we introduce the study on the convergence condition of AL-LDA. Lastly, the further study of the optimization algorithm of AL-LDA for neural networks is introduced.

4.1 Collection and metric

We select five typical collections to evaluate the performance of proposed models. All these datasets are publicly available and have been widely used in existing document classification literatures, including some most classical topic modeling approaches (Li et al. 2015a, b; Padmanabhan et al. 2017; Rubin et al. 2011; Ramage et al. 2009; Wang et al. 2020).

Yahoo Arts and Health multi-label subsets are from Yahoo Collection (Ueda and Saito 2003). We used the same training and test data presented by Ji et al. (2008), where training data consists of 1,000 documents, ensuring that each label appeared at least once. The remaining documents were used as the test data. Fudan University Chinese Text Classification Corpus collected by Dr. Li Ronglu is a popular single-label multi-class dataset. We selected 8,000 items from 9 categories. After removing common stop words and the terms occurred less than 8 times, we randomly selected 1,000 articles for training, ensuring that the article in each category appeared at least once, and reserved the remaining articles for

Table 2 Summary of experimental datasets

	Training Data	Test Data	Labels	Mean Label number per Document	Mean Label Frequency
Yahoo Arts	1,000	6,441	19	1.7	530
Yahoo Health	1,000	8,109	14	1.6	500
Fudan	1,000	7,000	9	1	889
20NewsGroups	11,314	7,532	20	1	942
Reuters	5,237	1,310	10	1.1	596

testing. 20NewsGroups[1] is a collection of news articles across 20 different newsgroups, which are considered as 20 different classification labels. In our experiments, we used 18,846 samples, 60% of them were selected for training and the remaining items for testing. Reuters-21578 dataset (Asuncion and Newman 2007), which is a collection of documents with news articles, is selected to evaluate the performance of ML-PA-LDA-MNS under MNS settings. To compare with it, we follow the same preprocess steps as (Padmanabhan et al. 2017): after using the Porter Stemmer algorithm (Porter 1980), removed common stop words and the terms occurred less than 50 times, documents that contained more than 20 words were retained. We randomly selected 80% articles for training and the remaining items for testing. The datasets are summarized in Table 2.

We consider binary prediction metrics, i.e., Macro-F1 and Micro-F1 scores, to evaluate our models. Firstly we define the Recall(R), Precision(P) and F1-score($F1$) (Goutte and Gaussier 2005) for a document as follows:

$$R = \frac{|l_d \cap l_d^*|}{|l_d|},$$
$$P = \frac{|l_d \cap l_d^*|}{|l_d^*|},$$
$$F1 = \frac{2PR}{P+R},$$

where l_d and l_d^* denote the true and estimated label set respectively. The Macro-F1 metric is obtained by averaging the document F1 across all documents. Meanwhile, the Micro-F1 metric considers the full testing corpus as a document (Yang 1999). Larger values of Macro and Micro-F1 scores imply better performance. We also consider One Error, which measures how many times the top-ranked label is not in the true label set, and is denoted as a percentage. Smaller values imply better classification for this metric.

To compare with the reported results of ML-PA-LDA-MNS, we used the accuracy (AC) measure, which is the proportion of correct predictions (both true positives and true negatives) among the total number of cases examined. The document labels, including ground-truth labels and predictive labels, are vectors of binary label. Letting TP, TN, FP and FN denote true positive, true negative, false positive and false negative numbers, w.r.t.

[1] sklearn.datasets.fetch 20newsgroups.

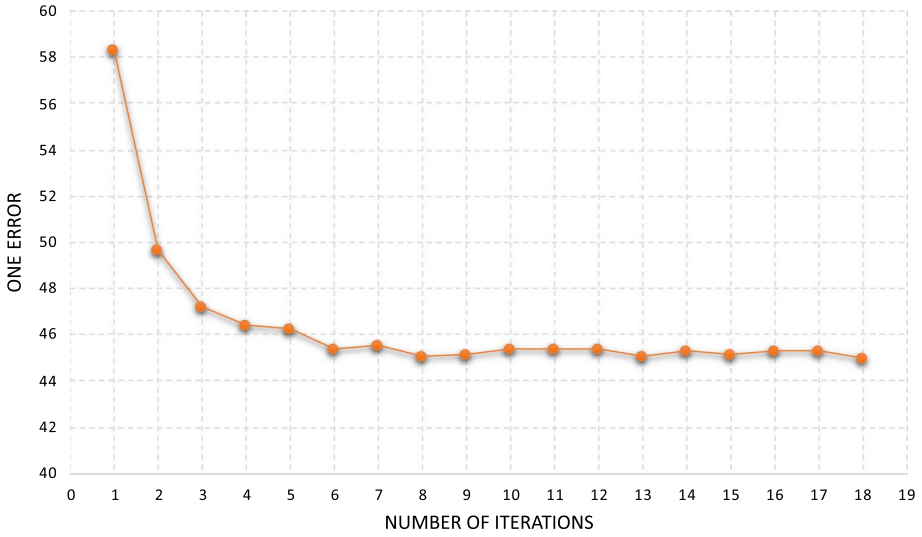

Fig. 4 One Error on Yahoo Arts with different iterations of SL-LDA Gibbs samplings

individual binary labels in document label vectors of the total test dataset, the overall accuracy is computed by

$$AC = \frac{TP + TN}{TP + TN + FP + FN}.$$

4.2 Hyperparameter and sampling parameter settings

Following the classical work (Griffiths and Steyvers 2004), the proposed model parameters originating from standard LDA are set as $\alpha = 50$[2], $\beta n_w = 0.01$, To smooth L-LDA, we set $\gamma = 50$, and add two hidden topics ($K_h = 2$). Our basic model, SL-LDA, has fast convergence. The performance with respect to the number of iterations is plotted in Fig. 4. After 8 iterations of Gibbs samplings, the proposed approach has stable performance. Zha and Li (2019) introduce similar results in their experiments of the supervised topic model. To train topic-word distributions Φ, we ran 3 independent MCMC chains. After an initial burn-in of 15 iterations, we took a single sample at the end of each chain, and averaged the samples to compute a single estimate for each iterative procedure. At test time, we ran 10 independent MCMC chains for document topic distributions Θ, and took 3 samples from each chain using an initial burn-in of 10 iterations and a 5 iteration lag between samples. All samples were averaged to estimate. To judge the AL-LDA convergence, we heuristically set $R^*_{\underline{D_{KL}}} = 0.7$ and $\Delta^*_{\underline{D_{KL}}} = 0.01$, which will be explained in detail in Sect. 4.6.

[2] Griffiths and Steyvers (2004) uses symmetric Dirichlet prior as $\alpha = 50/K$ for the standard LDA inference. To incorporate prior knowledge, we use asymmetric Dirichlet prior $\alpha \boldsymbol{m}$. In other words, α in our models is the concentration parameter of Dirichlet, and $\boldsymbol{m}$, which is not uniform, is the base measure vector of Dirichlet prior. So α in our models is similar to 50 in Griffiths and Steyvers (2004), and m_t, i.e., the tth element of $\boldsymbol{m}$, is similar to $1/K$.

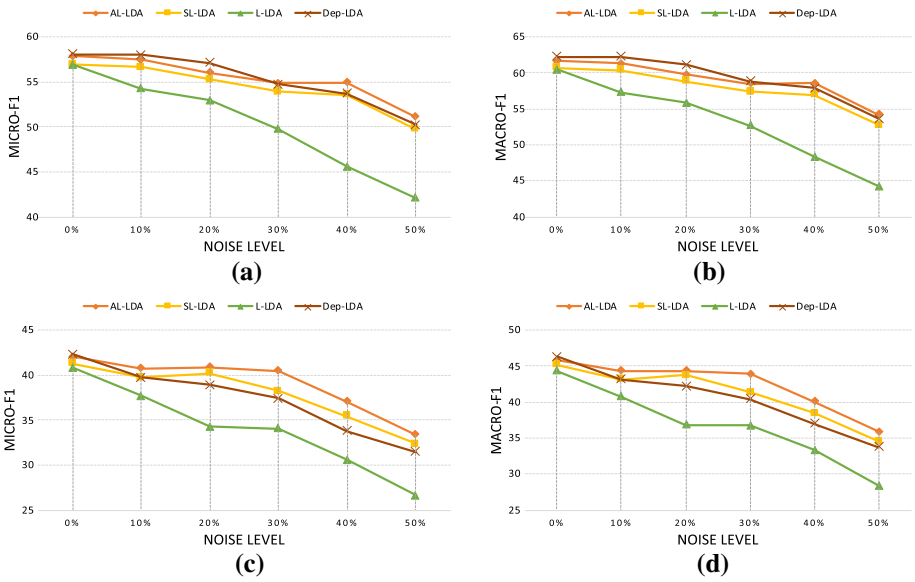

Fig. 5 Experimental binary prediction results on Yahoo Subsets with different noise levels, (a) Yahoo Health Micro-F1, (b) Yahoo Health Macro-F1, (c) Yahoo Arts Micro-F1, (d) Yahoo Arts Macro-F1

In multi-label classification experiments, the positive instance threshold-selection has effects on binary predictions. We use the proportional approach (Fürnkranz et al. 2008) on all compared algorithms. In this approach, the expected number of positive instances is $median(N_d^{TRAIN})$, where N_d^{TRAIN} is the number of labels for training document d.

4.3 Evaluation on NCAR settings

To evaluate the proposed models, we use Yahoo subsets, which represent multi-label classification tasks, as well as Fudan and 20newsgroups datasets, which represent single-label multi-class classification tasks. The training data was artificially corrupted by introducing random noise, i.e., NCAR. The noise level $p_e = x\%$ means that $x\%$ training articles are randomly selected and their prior labels are independently changed to the other remaining random labels. In other words, the probability of right labels of a document is $1 - p_e$, and the probability of each error label is about $\frac{p_e}{K-K_m}$, where K is the number of labels in the corpus, and K_m is the mean label number per document. Because $K > 2$ on multi-class datasets, and $K_m < \frac{K}{2}$ in the experimental datasets, we get if $p_e \leq 0.5, 1 - p_e > \frac{p_e}{K-K_m}$. In other words, the prior labels are useful while $p_e \leq 0.5$. To evaluate on massive label noise scenarios, the experiments were executed from 0% to 50% noise levels.

The state-of-the-art approach, Dependency-LDA (Rubin et al. 2011) that is known to outperform other topic modeling approaches (Burkhardt and Kramer 2018) is chosen as the main baseline. The open source code[3], which is the official release of the author, is implemented for inference without modification of hyper-parameters and sampling

[3] https://github.com/timothyrubin/DependencyLDA

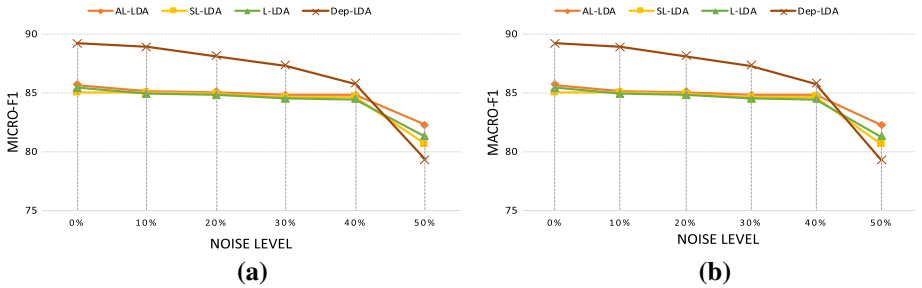

Fig. 6 Experimental binary prediction results on Fudan corpus with different noise levels, **a** Micro-F1, **b** Macro-F1

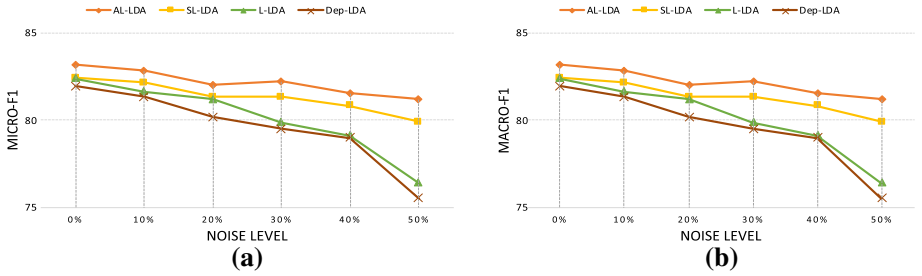

Fig. 7 Experimental binary prediction results on 20newsgroups corpus with different noise levels, **a** Micro-F1, **b** Macro-F1

parameters. In addition, the source code supports Flat-LDA, which is equal to L-LDA in practice (Rubin et al. 2011). We also list results of L-LDA (Flat-LDA) to compare.

Figure 5 shows the experimental binary prediction results, i.e., Micro-F1 and Macro-F1, for Yahoo subsets respectively. It can be seen that Micro-F1 (Fig. 5a,c) and Macro-F1 (Fig. 5b,d) have similar trends: SL-LDA performs better than L-LDA and AL-LDA performs better than SL-LDA under different noise levels. Furthermore, as the noise level increases, AL-LDA and SL-LDA perform significantly well relative to L-LDA. It is demonstrated that Micro-F1 difference between AL-LDA and L-LDA is about 1% under 0% noise level on Yahoo Health (Fig. 5a) , i.e., the original dataset, while the difference is more than 8% under 50% noise level. Similarly, Yahoo Arts Micro-F1 difference is about 1% under original dataset, while the difference is about 7% under 50% noise level (Fig. 5c). Dependency-LDA performs better under low noise levels, but under massive label noise, i.e., above 30% noise level on Yahoo Health (Fig. 5a,b) or above 0% noise level on Yahoo Arts (Fig. 5c,d), AL-LDA gets higher scores, and as the noise level increases, AL-LDA gets significantly advantage.

Figure 6 shows the Micro-F1 (a) and Macro-F1 (b) results for Fudan subset, which is a single-label multi-class dataset, so Micro-F1 and Macro-F1 have the same scores. It is clear that SL-LDA gets better scores than L-LDA on 4/6 noise levels, AL-LDA outperforms SL-LDA under all noise levels. Dependency-LDA gets the best scores below 40% noise level, and AL-LDA performs best on 50% noise level. 20newsgroups corpus is also a single-label multi-class. The results (Fig. 7) show our models perform better

Table 3 MNS settings, MNS1(top section) and MNS2(bottom section)

	Group1	Group2	Group3
Group probability	20%	40%	40%
Right label probability	51%-65%	66%-85%	86%-99.99%
Group probability	80%	10%	10%
Right label probability	51%-55%	56%-60%	61%-99.99%

Table 4 Experimental binary prediction results with MNS1 on Micro-F1 (top section) and Macro-F1 (bottom section), larger values imply better. All compared models get the same scores on Micro-F1 and Macro-F1 on Fudan and 20Newsgroups corpus, because they are single-label classification datasets

	Yahoo Health	Yahoo Arts	Fudan	20NewsGroups
AL-LDA	54.27	**39.40**	84.79	**82.20**
SL-LDA	53.63	38.40	84.49	81.15
L-LDA	46.05	34.59	88.16	81.34
Dep-LDA	**54.54**	38.58	**88.49**	80.50
AL-LDA	58.36	**43.20**	84.79	**82.20**
SL-LDA	57.76	42.20	84.49	81.15
L-LDA	48.72	37.53	88.16	81.34
Dep-LDA	**59.01**	42.48	**88.49**	80.50

Bold entries denote the best scores

than the compared algorithms, and the proposed models get more advantage while the noise level increases. Meanwhile, AL-LDA outperforms SL-LDA under all noise levels.

4.4 Evaluation on MNS settings

To evaluate the proposed models on Multiple Noisy Sources (MNS) settings, we also use Yahoo, Fudan, and 20Newsgroups datasets. Supposing several groups of annotators generate observed labels, each group has a probability range that the annotator gives right ones. Because we use the proportional approach in multi-label classification experiments, we also need to make sure that the training data and the test data have the similar mean label number per document. So we randomly selected the labels generated by annotators, ensuring the number of new labels is same as the original one. The simulation algorithm of multiple noisy sources is summarized as Algorithm 3. We simulated two MNS settings, i.e., MNS1 and MNS2, as Table 3. Each setting has five annotators. The label quality of MNS1 is significantly better than MNS2.

Table 5 Experimental binary prediction results with MNS2 on Micro-F1 (top section) and Macro-F1 (bottom section), larger values imply better. All compared models get the same scores on Micro-F1 and Macro-F1 on Fudan and 20Newsgroups corpus, because they are single-label classification datasets

	Yahoo Health	Yahoo Arts	Fudan	20NewsGroups
AL-LDA	51.69	**36.95**	**83.84**	**81.59**
SL-LDA	50.95	34.88	83.47	80.43
L-LDA	41.11	29.33	82.26	77.17
Dep-LDA	**51.94**	34.82	82.67	77.44
AL-LDA	55.65	**40.21**	**83.84**	**81.59**
SL-LDA	54.62	37.82	83.47	80.43
L-LDA	43.25	31.23	82.26	77.17
Dep-LDA	**55.75**	37.66	82.67	77.44

Bold entries denote the best scores

Table 6 Performances for different sizes of training data on Reuters-21578 with multiple noisy sources on Accuracy (top section) and Micro-F1 (bottom section), larger values imply better

% of Training data used	10%	30%	50%	70%	100%
AL-LDA	**96.7**	**97.3**	**97.3**	**97.3**	**97.4**
SL-LDA	96.4	**97.3**	97.2	**97.3**	97.2
ML-PA-LDA (Padmanabhan et al. 2017)	94.9	95.3	95.5	96.1	96.9
ML-PA-LDA-MNS (Padmanabhan et al. 2017)	92.7	93.0	93.6	93.7	94.2
AL-LDA	**84.6**	**87.4**	**87.3**	**87.7**	**87.7**
SL-LDA	83.4	**87.4**	87.2	87.4	87.0
ML-PA-LDA (Padmanabhan et al. 2017)	76.2	78.4	78.7	82.8	82.9
ML-PA-LDA-MNS (Padmanabhan et al. 2017)	61.6	61.9	62.9	65.0	66.9

Bold entries denote the best scores

Algorithm 3 Simulation algorithm of multiple noisy sources

```
for document d ∈ [1, D] do
  for each annotator do
    Decide which group the annotator belongs to
    Decide the probability p_r of right labels based on the right label probability range of the group
    Produce a random value r ∈ [0, 1]
    if r ≤ p_r then
      Give a right label (Select one from the original labels) to the candidate set of the document
    else
      Give a remaining random wrong label to the candidate set of the document
    end if
  end for
  for each original label do
    Replace with the random selection from the candidate set
  end for
end for
```

The binary predictions on MNS1 and MNS2 are listed in Tables 4 and 5 respectively. The results show AL-LDA performs best for 2/4 datasets on MNS1 (Table 4) and 3/4 datasets on MNS2 (Table 5). Dependency-LDA gets the best scores for 2/4 datasets on MNS1 and 1/4 datasets on MNS2. It suggests AL-LDA achieves competitive performance with Dependency-LDA on MNS1, which demonstrates relatively

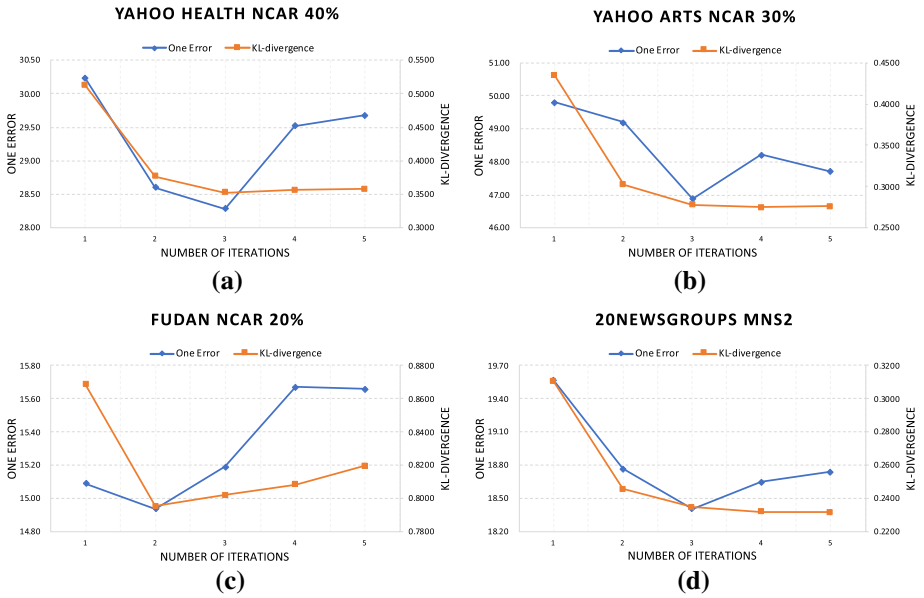

Fig. 8 AL-LDA performances and average KL-divergences for different iterations on four datasets. **a** Yahoo Health subset under 40% noise level, **b** Yahoo Arts subset under 30% noise level, **c** Fudan dataset under 20% noise level, **d** 20Newsgroups under MNS2 settings

high quality corpus, and outperforms Dependency-LDA on MNS2, which demonstrates massive noisy labels produced by low quality annotators. The results also show SL-LDA performs better for 2/4 datasets on MNS1 than L-LDA, and gets higher scores than L-LDA on MNS2 across four datasets. AL-LDA has significant advantages to SL-LDA on MNS1 and MNS2 across four datasets. It suggests the iterative optimization algorithm of AL-LDA helps to improve performance on MNS settings.

4.5 Comparisons with the reported results

To compare with ML-PA-LDA-MNS (Padmanabhan et al. 2017), which is a supervised topic model considering label noise, we use Reuters-21578 as MNS1 described in Sect. 4.4. Table 6 lists ML-PA-LDA and ML-PA-LDA-MNS reported results (Padmanabhan et al. 2017) as well as corresponding results of our models. Obviously, even considering the randomness of noisy label generation, our models perform better than the compared models. AL-LDA gets the best scores on both Accuracy and Micro-F1 measures across all settings of different training data sizes, and SL-LDA gets the second good results. Meanwhile, AL-LDA and SL-LDA get similar results on different sizes of training data. Actually, the AL-LDA AC difference between 10% and 100% training sizes is less than 1%. Furthermore, it is clear that our models have significant advantages on Micro-F1 relative to the compared models, i.e., AL-LDA and SL-LDA score about 20.1%-25.5% higher than ML-PA-LDA-MNS.

Table 7 AL-LDA iterative numbers of evaluations on NCAR settings across four datasets

	Original	10%	20%	30%	40%	50%
Yahoo Health	2	2	3	3	3	3
Yahoo Arts	2	2	2	3	2	3
Fudan	2	2	2	2	2	2
20NewsGroups	2	2	2	2	3	3

4.6 Study on the convergence condition of AL-LDA

Now we study the convergence condition of the proposed AL-LDA. We select four typical settings of experimental datasets to study. The other settings have similar results. Figure 8 shows the performances and model average KL-divergences defined by Eq. (6) with different iterations. Figure 8a,b,d get the best scores after three iterations, and more iterations worsen performances. Meanwhile, Fig. 8c, i.e., Fudan dataset under 20% noise level, gets the best scores after two iterations. In addition, Fig. 8a,c,d meet the convergence conditions by judging $\Delta_{\overline{D}_{KL}}$, which are -0.0035, -0.0066, and 0.0028 respectively. Meanwhile, Fig. 8b meets the convergence condition by judging $R_{\overline{D}_{KL}}$, which is 0.64. We set $\Delta^*_{\overline{D}_{KL}} = 0.01$ and $R^*_{\overline{D}_{KL}} = 0.7$ in the evaluations. The results show the proposed AL-LDA performs well on these hyper-parameter settings.

Table 7 lists the iterative numbers of evaluations on NCAR settings. It is demonstrated that the proposed AL-LDA meets the convergence conditions through 2 or 3 iterations in each of experiments, and often needs more iterations under more noisy labels.

4.7 Further study on the optimization algorithm of AL-LDA for neural networks

AL-LDA may be suggested to a generalized optimization framework for supervised document classification. To preliminarily study the proposed optimization algorithm for neural networks, we design a deep neural network (DNN) of 3 fully connected layers, 1 batch norm layer, and 1 softmax layer. The network uses soft plus in the hidden layers. For reducing the overtraining effect, we use dropout on the third fully connected layer. The input layer is a 8000 dimensions of one-hot word vectors, and the sizes of the hidden layers are 2000,1000, as well as 500. Adam optimizer is used with the second exponential decay rate of 0.99. To evaluate the optimization algorithm of AL-LDA, we use Yahoo, Fudan, and 20newsgroups datasets at NCAR settings of 30% noisy level.

The hard-label method of the joint optimization framework for learning with noisy labels (Tanaka et al. 2018) is chosen as the main baseline. Following it, we use the loss function as

$$\mathcal{L}(\boldsymbol{\omega}) = \mathcal{L}_c(\boldsymbol{\omega}|X, Y) + \alpha\mathcal{L}_p(\boldsymbol{\omega}|X) + \beta\mathcal{L}_e(\boldsymbol{\omega}|X),$$

where X denotes the training documents, Y denotes the ground-truth labels, and $\boldsymbol{\omega}$ denotes the network parameters. $\mathcal{L}_c$, $\mathcal{L}_p$, $\mathcal{L}_e$ denote the classification loss and two regularization terms respectively, as well as α, β denote hyper parameters. According to Tanaka et al. (2018), $\mathcal{L}_c$ is the KL-divergence from predictive label vectors $S(\boldsymbol{\omega}, X)$ to ground-truth label vectors Y. $\mathcal{L}_c$, $\mathcal{L}_p$, and $\mathcal{L}_e$ are defined as follows,

Table 8 Experimental One Error results of the basic DNN, the proposed optimization algorithm for the DNN, and the hard-label method (Tanaka et al. 2018) for the DNN on Yahoo, Fudan and 20Newsgroups datasets on NCAR settings with 30% noisy level, smaller values imply better

	DNN	Optimization algorithm of AL-LDA	Hard-label method of joint optimization framework
Yahoo health	41.06	39.65	**38.58**
Yahoo arts	61.67	**59.80**	60.45
Fudan	56.83	55.23	**54.60**
20 newsgroups	33.47	**28.63**	31.32

Bold entries denote the best scores

$$\mathcal{L}_c(\boldsymbol{\omega}|X,Y) = \frac{1}{n}\sum_{i=1}^{n} D_{KL}(\boldsymbol{y}_i||\boldsymbol{s}(\boldsymbol{\omega},\boldsymbol{x}_i)),$$
$$\mathcal{L}_p(\boldsymbol{\omega}|X) = D_{KL}(\boldsymbol{p}||\bar{\boldsymbol{s}}(\boldsymbol{\omega},X)),$$
$$\mathcal{L}_e(\boldsymbol{\omega}|X) = \frac{1}{n}\sum_{i=1}^{n} H(\boldsymbol{s}(\boldsymbol{\omega},\boldsymbol{x}_i),\boldsymbol{s}(\boldsymbol{\omega},\boldsymbol{x}_i)),$$

where $\boldsymbol{s}(\boldsymbol{\omega},\boldsymbol{x}_i)$ is the network output of $\boldsymbol{x}_i$, $\boldsymbol{p}$ is a distribution of classes among all training documents, and $\bar{\boldsymbol{s}}(\boldsymbol{\omega},X)$ is the mean probability output of the network.

To train the basic model, we set the learning rate as 0.002 for 3,000 iterations, batch size is 200, and $\alpha = \beta = 0.05$. To evaluate the optimization algorithm of AL-LDA and the baseline method, we take two steps. In the first step, we pre-train the model for learning rate 0.02, i.e., use high learning rate introduced by Tanaka et al. (2018), and update labels by the proposed optimization algorithm of AL-LDA and the hard-label method. In the second step, we initialize the network parameters and train the network as the basic model with the labels obtained in the first step.

The results (Table 8) show both the proposed optimization algorithm and the hard-label method perform well for the neural network. They all get better scores than the basic DNN. Our proposed algorithm performs best for 2/4 datasets, and the hard-label method outperforms ours for the others. Meanwhile, the two optimization algorithms have same time complexity $O(n)$ with $n = KD$, where K is the number of labels and D is the number of training documents. The comparisons demonstrate our algorithm is competitive with the baseline method in the experiment.

5 Discussion

The experimental results clearly demonstrate SL-LDA performs better than L-LDA across four datasets under NCAR settings, and even better than Dependency-LDA while massive label noise. The results also show SL-LDA outperforms L-LDA for 6/8 datasets, and even gets higher scores than Dependency-LDA for 4/8 datasets under MNS settings. It is well known that Dependency-LDA has significant advantages on multi-label classification among topic modeling approaches (Burkhardt and Kramer 2018). However, Dependency-LDA is also a complex model with many parameters which heavily depend on inference techniques (Li et al. 2015b). In other words, the proposed

Table 9 Experimental OneError_difference results on NCAR settings of Yahoo, Fudan, and 20Newsgroups datasets, larger values imply better

Noise level	Original	10%	20%	30%	40%	50%
Yahoo health	1.57	1.10	1.05	1.27	1.81	1.73
Yahoo arts	0.99	1.43	0.88	**2.92**	2.15	1.38
Fudan	0.55	0.08	0.15	0.22	0.16	1.61
20 Newsgroups	0.77	0.70	0.71	0.88	0.71	1.30

Bold entries denote the best score

Table 10 Experimental OneError_difference results on MNS settings of Yahoo, Fudan, and 20Newsgroups datasets, larger values imply better

	Yahoo health	Yahoo arts	Fudan	20 Newsgroups
MNS1	0.83	1.32	0.30	1.05
MNS2	0.98	**2.72**	0.37	1.16

Bold entries denote the best score

SL-LDA, which is a simple extension of L-LDA, outperforms complex Dependency-LDA under massive label noise. In addition, the proposed SL-LDA outperforms ML-PA-LDA-MNS, which models the multiple noisy sources and is obviously more complex than our model. In summary, SL-LDA is competitive with some complex classical topic modeling approaches under label noise. These results suggest the proposed model benefits from Dirichlet smoothing. Lukasik et al. (2020) present the label smoothing has a connection to loss correction and regularization techniques, and empirically demonstrate that label smoothing significantly improves performance under label noise. Our experimental results obviously accord with their conclusion.

Another interesting aspect of results is AL-LDA outperforms SL-LDA on all experiments. To demonstrate the effect of AL-LDA optimization framework, we define

$$OneError\_difference = OneError_{SL-LDA} - OneError_{AL-LDA}.$$

Because smaller values imply better classification for One Error, a positive value means AL-LDA performs better than SL-LDA, i.e., the proposed framework has performed well, and a negative value suggests the framework is detrimental. Table 9 lists *OneError_difference* of experiments on NCAR settings with different noise levels across four datasets, and Table 10 lists results of MNS settings. It is clear that AL-LDA outperforms SL-LDA in all cases. Specially, an improvement of around 3 percent in One Error under 30% noise level on Yahoo Arts subset. These results clearly show the optimization framework of AL-LDA further improves the model robustness.

Obviously, the experimental results show AL-LDA performs well under label noise. However, one limitation of AL-LDA is multiple iterations lead to longer training time. In the experiments, AL-LDA takes 2~4 times as much training time as SL-LDA to meet the convergence condition. SL-LDA is a relatively simple model like basic LDA and has high training efficiency, but the baseline approach, Dependency-LDA is a complex model with additional layer leading computational complexity. So the training efficiency of AL-LDA is competitive with the compared Dependency-LDA.

Our study on the impact of multiple iterations of AL-LDA suggests the optimization algorithm of AL-LDA has positive effects, i.e., reduce the influence of label noise, in the first few iterations; however, it has negative effects because of overfitting on noisy labels or cleaning right labels. To benefit from the optimization algorithm while suppressing its negative effects, we propose the convergence condition that works well on all experiments.

Furthermore, the optimization algorithm of AL-LDA may be suggested to a generalized optimization framework for supervised document classification. We evaluate the algorithm on a deep neural network. The results show our algorithm is competitive with the compared method (Tanaka et al. 2018), which is a state-of-the-art optimization framework for neural networks under label noise. However, the optimization framework (Tanaka et al. 2018) builds on the phenomenon that a DNN trained on noisy labeled datasets does not memorize noisy labels and maintains high performance for clean data under a high learning rate. So it has particular request for the learning rate on each iterations, and is only applicable to DNNs. On the contrary, the optimization algorithm of AL-LDA builds on general principles, i.e., *maximizing entropy* and *minimizing cross-entropy*, so it has better generalizability on supervised learning under label noise. The proposed algorithm not only works well on generative models such as statistical topic models, but also supports discriminative approaches such as DNNs. Still, more research will be needed to determine the impact of the proposed optimization algorithm on neural networks. We leave this research for future work.

6 Conclusion

Statistical topic models based on LDA have been widely developed in the field of document classification. They are interpretable generative models, which not only predict document classes, but also let us understand which words are important for the class, which parts of a text belong to the class. Some of them can support supervised learning and achieve competitive results with state-of-the-art approaches; however, massive label noise, which widely exists in the real world, has many negative consequences to the model performance. To address this issue, we propose robust topic models, i.e., Smoothed Labeled LDA (SL-LDA) and Adaptive Labeled LDA (AL-LDA). SL-LDA overcomes the problem of noisy label overfitting by Dirichlet smoothing, and AL-LDA is an optimization framework based on SL-LDA. AL-LDA reduces the noisy influence by iteratively optimize the model prior based on two principles, i.e., *maximizing entropy* and *minimizing cross-entropy*. It is worth noting that the proposed optimization framework avoids the chicken-and-egg dilemma, i.e., identifying noisy labels needs good classifier and classifier learning depends on cleansing data, which popularly exists in label noise cleaning approaches. In addition, to avoid overtraining, we study the convergence condition of AL-LDA and give the suggested parameters.

We evaluate the proposed models and compared methods on two noise settings, i.e., *noisy completely at random* (NCAR) and *Multiple Noisy Sources* (MNS), across different datasets. The experimental results show the proposed models have robust performances under label noise, and get better results under massive label noise than state-of-the-art topic models, which are more complex than ours. We also demonstrate the proposed optimization algorithm of AL-LDA has advantages on other supervised document classification methods, e.g., deep neural networks.

In the future, we intend to further research the extension of AL-LDA to other supervised document classifiers, and plan to apply the models to some other applications, e.g., news video segmentation and summarization.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant No. 61772352, 61971296, U19A2078; National Key Research and Development Project under Grant No. 2020YFB1711800 and 2020YFB1707900; the Science and Technology Planning Project of Sichuan Province under Grant No. 2019YFG0400, 2021YFG0152, 2020YFG0479, 2020YFG0322, 2020GFW035; and the R&D Project of Chengdu City under Grant No. 2019-YF05-01790-GX.

References

Angelova, A., Abu-Mostafam, Y., Perona, P., (2005) Pruning training sets for learning of object categories. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), IEEE*, vol 1, (pp. 494–501).

Asuncion A, Newman D (2007) Uci machine learning repository.

Biggio, B., Nelson, B., Laskov, P., (2011) Support vector machines under adversarial label noise. In: *Asian Conference on Machine Learning, PMLR*, (pp. 97–112).

Blei DM, McAuliffe JD (2010) Supervised topic models. arXiv preprint arXiv:10030783.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3,* 993–1022.

Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition, 37*(9), 1757–1771.

Brodley, C. E., & Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of Artificial Intelligence Research, 11,* 131–167.

Burkhardt, S., & Kramer, S. (2018). Online multi-label dependency topic models for text classification. *Machine Learning, 107*(5), 859–886.

Burkhardt, S., & Kramer, S. (2019). A survey of multi-label topic models. *ACM SIGKDD Explorations Newsletter, 21*(2), 61–79.

De La Torre, F., & Black, M. J. (2003). A framework for robust subspace learning. *International Journal of Computer Vision, 54*(1), 117–142.

Frénay, B., & Verleysen, M. (2013). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems, 25*(5), 845–869.

Fürnkranz, J., Hüllermeier, E., Mencía, E. L., & Brinker, K. (2008). Multilabel classification via calibrated label ranking. *Machine Learning, 73*(2), 133–153.

Ghosh, A., Kumar, H., Sastry, P., (2017) Robust loss functions under label noise for deep neural networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol 31(1)

Golzari, S., Doraisamy, S., Sulaiman, M. N., & Udzir, N. I. (2009). The effect of noise on rwtsairs classifier. *European Journal of Scientific Research, 31*(4), 632–641.

Goutte, C., Gaussier, E., (2005) A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In: *European Conference on Information Retrieval*, Springer, (pp. 345–359).

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(Supplement 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review, 106*(4), 620.

Jeatrakul, P., Wong, K. W., & Fung, C. C. (2010). Data cleaning for classification using misclassification analysis. *Journal of Advanced Computational Intelligence and Intelligent Informatics, 14*(3), 297–302.

Ji, S., Tang, L., Yu, S., Ye, J., (2008) Extracting shared subspace for multi-label classification. In: *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 381–389).

Jiang, L., Meng, D., Mitamura, T., Hauptmann, AG., (2014) Easy samples first: Self-paced reranking for zero-example multimedia search. In: *Proceedings of the 22nd ACM International Conference on Multimedia*, (pp. 547–556).

Khardon, R., & Wachman, G. (2007). Noise tolerant variants of the perceptron algorithm. *Journal of Machine Learning Research, 8*(2), 227–248.

Kumar, H., Manwani, N., Sastry, P., (2020) Robust learning of multi-label classifiers under label noise. In: *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, (pp. 90–97).

Lacoste-Julien, S., Sha, F., Jordan, MI., (2008) Disclda: Discriminative learning for dimensionality reduction and classification. In: *Advances in Neural Information Processing Systems*, (pp. 897–904).

Li, J., Wong, Y., Zhao, Q., Kankanhalli, MS., (2019) Learning to learn from noisy labeled data. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (pp. 5051–5059).

Li, X., Ouyang, J., & Zhou, X. (2015a). Supervised topic models for multi-label classification. *Neurocomputing, 149,* 811–819.

Li, X., Ouyang, J., Zhou, X., Lu, Y., & Liu, Y. (2015b). Supervised labeled latent dirichlet allocation for document categorization. *Applied Intelligence, 42*(3), 581–593.

Li, X., Ma, Z., Peng, P., Guo, X., Huang, F., Wang, X., & Guo, J. (2018). Supervised latent dirichlet allocation with a mixture of sparse softmax. *Neurocomputing, 312,* 324–335.

Liu, CY., Liu, Z., Li, T., Xia, B., (2018) Topic modeling for noisy short texts with multiple relations. In: *SEKE*, (pp. 610–609).

Lukasik, M., Bhojanapalli, S., Menon, A., Kumar, S., (2020) Does label smoothing mitigate label noise? In: *International Conference on Machine Learning, PMLR*, (pp. 6448–6458).

Magnusson, M., Jonsson, L., Villani, M., (2016) Dolda-a regularized supervised topic model for high-dimensional multi-class regression. arXiv preprint arXiv:160200260.

Manwani, N., & Sastry, P. (2013). Noise tolerance under risk minimization. *IEEE Transactions on Cybernetics, 43*(3), 1146–1151.

Mikalsen, K. Ø., Soguero-Ruiz, C., Bianchi, F. M., & Jenssen, R. (2019). Noisy multi-label semi-supervised dimensionality reduction. *Pattern Recognition, 90,* 257–270.

Padmanabhan, D., Bhat, S., Shevade, S., & Narahari, Y. (2017). Multi-label classification from multiple noisy sources using topic models. *Information, 8*(2), 52.

Patrini, G., Rozza, A., Krishna Menon, A., Nock, R., Qu, L., (2017) Making deep neural networks robust to label noise: A loss correction approach. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 1944–1952).

Porter, M. F. (1980). An algorithm for suffix stripping. *Program, 14*(3), 130–137.

Prechelt, L., (1998) Early stopping-but when? In: Neural Networks: Tricks of the trade, Springer, (pp. 55–69)

Ramage, D., Hall, D., Nallapati, R., Manning, CD., (2009) Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In: *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, (pp. 248–256).

Ramage, D., Manning, CD., Dumais, S., (2011) Partially labeled topic models for interpretable text mining. In: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (pp. 457–465).

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., & Moy, L. (2010). Learning from crowds. *Journal of Machine Learning Research, 11*(4), 1297–1322.

Ren, M., Zengm W., Yang, B., Urtasun, R., (2018) Learning to reweight examples for robust deep learning. In: *International Conference on Machine Learning, PMLR*, (pp. 4334–4343).

Rubin, T. N., Chambers, A., Smyth, P., & Steyvers, M. (2011). Statistical topic models for multi-label document classification. *Machine Learning, 88*(1–2), 157–208. https://doi.org/10.1007/s10994-011-5272-5.

Shu, J., Xie, Q., Yi, L., Zhao, Q., Zhou, S., Xu, Z., Meng, D., (2019) Meta-weight-net: Learning an explicit mapping for sample weighting. arXiv preprint arXiv:190207379

Soleimani, H., & Miller, D. J. (2019). Exploiting the value of class labels on high-dimensional feature spaces: Topic models for semi-supervised document classification. *Pattern Analysis and Applications, 22*(2), 299–309.

Sun, Jw., Zhao, Fy., Wang, Cj., Chen, Sf., (2007) Identifying and correcting mislabeled training instances. In: *Future Generation Communication and Networking (FGCN 2007)*, IEEE, vol 1, (pp. 244–250).

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., (2016) Rethinking the inception architecture for computer vision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 2818–2826).

Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K., (2018) Joint optimization framework for learning with noisy labels. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 5552–5560).

Ueda, N., Saito, K., (2003) Parametric mixture models for multi-labeled text. In: *Advances in Neural Information Processing Systems*, (pp. 737–744).

Veit, A., Alldrin, N., Chechik, G., Krasin, I., Gupta, A., Belongie, S., (2017) Learning from noisy large-scale datasets with minimal supervision. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (pp. 839–847).

Wang, W., Guo, B., Shen, Y., Yang, H., Chen, Y., & Suo, X. (2020). Twin labeled LDA: A supervised topic model for document classification. *Applied Intelligence, 50*(12), 4602–4615. https://doi.org/10.1007/s10489-020-01798-x.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. *Information Retrieval, 1*(1–2), 69–90.

Ying, X. (2019). An overview of overfitting and its solutions. *Journal of Physics: Conference Series, IOP Publishing, 1168*(2), 022022.

Zha, D., & Li, C. (2019). Multi-label dataless text classification with topic modeling. *Knowledge and Information Systems, 61*(1), 137–160.

Zhang, W., Wang, D., & Tan, X. (2019). Robust class-specific autoencoder for data cleaning and classification in the presence of label noise. *Neural Processing Letters, 50*(2), 1845–1860.

Zhang, Y., Ma, J., Wang, Z., & Chen, B. (2017). Lf-lda: A topic model for multi-label classification. In: *International Conference on Emerging Internetworking* (pp. 618–628). Data & Web Technologies: Springer.

Zhang, Z., Sabuncu, MR., (2018) Generalized cross entropy loss for training deep neural networks with noisy labels. arXiv preprint arXiv:180507836.

Zhu, J., Ahmed, A., & Xing, E. P. (2012). Medlda: Maximum margin supervised topic models. *The Journal of Machine Learning Research, 13*(1), 2237–2278.