Neural labeled LDA: a topic model for semi-supervised document classification

Wei Wang (wong.wei@163.com)
Sichuan University https://orcid.org/0000-0003-4718-5922

Bing Guo
Sichuan University

Yan Shen
Chengdu University of Information Technology

Han Yang
Chengdu Sobey Technology

Yaosen Chen
Sichuan University

Xinhua Suo
Sichuan University

Research Article

Keywords: Neural topic model, Semi-supervised learning, Document classification

Posted Date: April 26th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-171202/v1

Noname manuscript No.
(will be inserted by the editor)

Neural labeled LDA: a topic model for semi-supervised document classification

**Wei Wang · Bing Guo · Yan Shen · Han Yang ·
Yaosen Chen · Xinhua Suo**

Received: date / Accepted: date

Abstract Recently, some statistical topic modeling approaches based on LDA have been applied in the field of supervised document classification, where the model generation procedure incorporates prior knowledge to improve the classification performance. However, these customizations of topic modeling are limited by the cumbersome derivation of a specific inference algorithm for each modification. In this paper, we propose a new supervised topic modeling approach for document classification problems, Neural Labeled LDA (NL-LDA), which builds on the VAE framework, and designs a special generative network to incorporate prior information. The proposed model can support semi-supervised learning based on the *manifold* assumption and *low-density* assumption. Meanwhile, NL-LDA has a consistent and concise inference method while semi-supervised learning and predicting. Quantitative experimental results demonstrate our model has outstanding performance on supervised document classification relative to the compared approaches, including traditional statistical and neural topic models. Specially, the proposed model can support both single-label and multi-label document classification. The proposed NL-LDA performs significantly well on semi-supervised classification, especially under a small amount of labeled data. Further comparisons with related works also indicate our model is competitive with state-of-the-art topic modeling approaches on semi-supervised classification.

Keywords Neural topic model · Semi-supervised learning · Document classification

W. Wang · B. Guo · Y. Chen · X. Suo
College of Computer Science, Sichuan University, Chengdu, China

W. Wang · H. Yang
Media Intelligence Laboratory, Sobey Technology, Chengdu, China

W. Wang
Peng Cheng Laboratory, Shenzhen, China

Y. Shen
School of Computer Science, Chengdu University of Information Technology, Chengdu, China

W. Wang
E-mail: wong.wei@163.com
B. Guo
Corresponding author
Y. Shen
Co-corresponding author

1 Introduction

Statistical topic modeling approaches(Blei, 2012), e.g., Latent Dirichlet Allocation (LDA)(Blei et al, 2003), have been widely applied in the field of data mining, latent data discovery, and document classification(Jelodar et al, 2018). Standard LDA is a completely unsupervised algorithm, and then how to incorporate prior knowledge into the topic modeling procedure is a popular research direction(Burkhardt and Kramer, 2019b; Chen et al, 2019). A major challenge of these LDA customizations is the computational cost of computing the posterior distribution. For standard LDA, the popular inference methods include variational inference(Blei et al, 2003), collapsed Gibbs sampling(Griffiths and Steyvers, 2004), and collapsed variational Bayes(Teh et al, 2006). However, all these methods have a drawback that requires re-deriving the inference algorithms even if there is only a small change to the modeling procedure. Recently, Variational Auto-encoder (VAE)(Kingma and Welling, 2013; Rezende et al, 2014) has been considered as a new choice for topic models, because it is deemed to a black box inference method, and dose not need requiring model-specific derivations. To the best of our knowledge, the Neural Variational Document Model (NVDM) proposed by Miao et al (2016) is the first text topic model based on the VAE framework, but it dose not use the Dirichlet distribution, which promotes sparsity and leads to more interpretable topics as a prior in the modeling procedure of LDA. To handle this issue, there are many methods have been introduced and get competitive performance(Burkhardt and Kramer, 2019a). Furthermore, some topic modeling approaches based on the VAE framework, can incorporate prior knowledge for supervised learning(Card et al, 2018).

With the digital text grows explosively in Web, where unlabeled data is abundant, while only a limited subset of data samples has their corresponding labels, supporting semi-supervised classification is an increasing research field of topic modeling approaches these years(Pavlinek and Podgorelec, 2017; Soleimani and Miller, 2017). Standard LDA is an unsupervised algorithm, so combining standard unsupervised LDA and customized supervised LDA to support semi-supervised learning is a natural thought(Soleimani and Miller, 2016; Wang et al, 2012; Zhang and Wei, 2014). Meanwhile, there are also some semi-supervised approaches under the VAE framework(Kingma et al, 2014), especially for document classification(Xu et al, 2017). However, topic modeling approaches adopting the VAE framework for semi-supervised learning are still rare, or have a complicated model structure(Zhou et al, 2020). To address this challenge, we propose Neural Labeled LDA (NL-LDA). To handle the Dirichlet distribution, we employ a Laplace approximation following Srivastava and Sutton (2017). Inspired by SLDA(Mcauliffe and Blei, 2007), the proposed model incorporates the prior knowledge by an additional label generative network with a weight parameter, which leads to a flexible model that can employ various types of prior information. To support the semi-supervised learning, we apply the *manifold* assumption and *low-density* assumption. Learning on the unlabeled data is useful to discover the latent topics, i.e., the *manifold*, and then helps to improve the model classification performance. To further improve the performance of semi-supervised learning, we adopt the *low-density* assumption by extending the object function. Meanwhile, the model supports supervised, unsupervised learning, and prediction by a consistent and concise inference method.

Our contribution is summarized as follows. We propose a novel topic model, i.e., NL-LDA, which is an extension of SLDA for semi-supervised document classification. The proposed approach is a flexible model that allows a variety of extensions for incorporating prior knowledge, and has a consistent and concise inference method based on the VAE framework. The model has been evaluated on several kinds of typical document classification tasks, including single-label and multi-label classification. The experimental results demonstrate the proposed model has better performance than related works, including traditional and neural topic modeling approaches. Specially, the proposed model has significant advantages on semi-supervised document classification under a small amount of labeled data.

The rest of the paper is structured as follows. Section 2 reviews the related works; section 3 describes the proposed method; section 4 introduces the experiments and evaluation results. We discuss the results in Section 5. Finally, section 6 gives concluding remarks and an outline of future work.

2 Related work

Our work is related to two research lines, which are traditional statistical supervised topic models and neural topic modeling approaches.

LDA proposed by 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 topics. The model considers each document as random mixtures over topics, and each topic is a distribution over words. However, the automatically learned topics are hard to interpret and may not suit an end-user application, e.g. categorization. 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; the other generates the prior knowledge first, and then generates the words conditioned on them. The typical approach of the first type is Supervised Latent Dirichlet Allocation (SLDA)(Mcauliffe and Blei, 2007), where each document is paired with a response that infers topic prediction. Labeled LDA (L-LDA) introduced by Ramage et al (2009) is a typical second type approach. It 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, but it constrains the topic distributions in the observed labels that lead to over-focus on them. To alleviate this problem, Dependency-LDA(Rubin et al, 2011) incorporates another topic model to model the observed label correlations, which is deemed to be crucial for multi-label classifiers(Burkhardt and Kramer, 2019b). Another recent improved approach of L-LDA is Twin labeled LDA(Wang et al, 2020b), which employs two sets of parallel topic modeling processes, one incorporates the prior label information by hierarchical Dirichlet distributions, the other models the grouping tags that have prior knowledge about the label correlation.

Most LDA variants require approximate inference methods, which have the drawback that small changes to the modeling procedure result in a re-derivation of the inference algorithm. To overcome this challenge, some neural topic models that employ blackbox inference mechanism based on the VAE framework are proposed. The principle idea of VAEs is to build an inference neural network, which directly maps a document to an approximate posterior distribution of latent variables, and a generative neural network, which reproduces a document close to the observed document from latent variables. The two networks are jointly learned through the stochastic gradient descent method (SGD). Consequently, the VAE framework is considered to have the ability of discovering representative topics as topic modeling procedures. NVDM(Miao et al, 2016) successfully uses the idea of VAEs to train a topic model with a Gaussian prior for the latent variables. To overcome the problem that the Dirichlet distribution is not a location scale family, which hinders the reparameterization utilized in the VAE framework, Srivastava and Sutton (2017) employ a Laplace approximation for modeling a Dirichlet prior of the latent variables; Joo et al (2020) approximate the inverse cumulative distribution function of the Gamma distribution, which is a component of the Dirichlet distribution; Zhang et al (2018) utilize the Weibull distribution; and Burkhardt and Kramer (2019a) solve this problem based on rejection samples. Meanwhile, to utilize Dirichlet prior, Wang et al (2020a) abandon the VAE framework, and propose Bidirectional Adversarial Topic (BAT) model, which applies bidirectional adversarial training for neural topic modeling. All these methods of unsupervised neural topic models have achieved competitive results. Furthermore, SCHOLAR(Card et al, 2018) building on ProdLDA proposed by Srivastava and Sutton (2017) is a general neural framework for supervised topic models. It can use metadata as labels to help infer topics that are relevant in predicting those labels.

Semi-supervised learning is the branch of machine learning using labeled and unlabeled data to perform certain learning tasks, e.g. document classification. Researches on statistical topic modeling approaches focused on semi-supervised classification are well documented. Wang et al (2012) describe semi-supervised LDA based on standard LDA and L-LDA. HSLDA proposed by Zhang and Wei (2014) adopts joint distribution of LDA and SLDA to generate semi-supervised topic models. ST LDA proposed by Pavlinek and Podgorelec (2017) comprises a semi-supervised text classification algorithm based on self-training. MCCTM proposed by Soleimani and Miller (2017) is a semi-supervised model to jointly extract topics from a collection of text documents and classify new documents. They reports MCCTM outperforms other semi-supervised topic models with respect to classification performance. There are few neural topic modeling approaches for semi-supervised document classification. Zhou et al (2020) propose a semi-supervised topic model, S-VAE-GM, under the VAE framework. The approach assumes that a document is modeled as a mixture of classes, and a class is modeled as a mixture of latent topics under Gaussian mixture assumption.

3 The Proposed Method

Firstly, we review the SLDA model and introduce the Neural Labeled LDA (NL-LDA), then propose the inference method. Lastly, the semi-supervised learning is introduced. We summarize some important notations in Table 1.

Table 1: *Notation descriptions*

Notation	Description
K	Number of topics
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 distributions
Φ	The matrix of topic-word distributions
f_g	The generative network for documents
f_y	The generative network for labels
μ_0	The mean vector of the logistic normal for the variational approximation to the posterior
f_μ	The inference network for μ_0
Σ_0	The diagonal covariance of the logistic normal for the variational approximation to the posterior
f_Σ	The inference network for Σ_0

3.1 NL-LDA

Our model builds on SLDA, which is a supervised extension of LDA. To incorporate prior information, e.g. classification labels, SLDA adds to LDA a response variable associated with each document. Given the number of labels K, the number of documents D, and the number of words V, Θ is the matrix of document-topic distributions, Φ is the matrix of topic-word distributions. θ_{dt} is the topic proportion for topic t in document d with $\sum_{t=1}^{K} \theta_{dt} = 1$, ϕ_{tn} is the probability of word n under topic t with $\sum_{n=1}^{V} \phi_{tn} = 1$, and y_d is the response variable of document d, SLDA as a generative process is summarized as Algorithm 1.

Based on SLDA, we propose NL-LDA. To handle the Dirichlet within Variational Auto-encoder(VAE), we follow AVITM proposed by Srivastava and Sutton (2017), who employ a

Algorithm 1 Generative process of SLDA

```
for document $d \in [1, D]$ do
    Generate $\boldsymbol{\theta}_d = \{\theta_{dt}\}_{t=1}^{K} \sim Dirichlet(\cdot|\eta)$
    for word $n \in [1, N_d]$ do
        Sample topic $z_{dn} \sim Multinomial(\boldsymbol{\theta}_d)$
        Sample word $w_{dn} \sim Multinomial(\boldsymbol{\phi}_{z_{dn}})$
    end for
    Draw response variable $y_d | z_{1:N_d}, \gamma, \sigma^2 \sim \mathcal{N}(\gamma^T \bar{z}, \sigma^2)$, where $\bar{z} = \frac{1}{N_d} \sum_n z_n$
end for
```

logistic normal prior on θ_d instead of Dirichlet prior, and collapse z. We further replace the matrix product of Θ and Φ with a more flexible generative network, f_g, followed by a softmax transform represented by $\sigma(\cdot)$. Meanwhile, instead of using a normal linear model in SLDA, we suggest a new response variable function f_y, which is a more flexible multi-layer neural network followed by a softmax transform σ. NL-LDA as a generative story is summarized as Algorithm 2.

Algorithm 2 Generative process of NL-LDA

```
for document $d \in [1, D]$ do
    Generate $\boldsymbol{\theta}_d = \{\theta_{dt}\}_{t=1}^{K} = \mathcal{LN}(\boldsymbol{\mu}_1(\eta), diag(\Sigma_1(\eta)))$
    Generate $\delta_d = f_g(\boldsymbol{\theta}_d, \Phi)$
    for word $n \in [1, N_d]$ do
        Sample word $w_{dn} \sim Multinomial(\sigma(\delta_d))$
    end for
    Generate $\boldsymbol{y}_d = \sigma(f_y(\boldsymbol{\theta}_d))$
end for
```

In the proposed algorithm, $\boldsymbol{y}_d$ is the classification label probability distribution of document d with the simplex constraint. For single-label classification, we select the label that have the maximum probability, and for multi-label classification, we select the labels with a relatively high probability. To approximate a Dirichlet prior with hyper-parameter η, the mean and diagonal covariance terms of a multivariate normal prior, i.e., $\boldsymbol{\mu}_1(\eta)$ and $\Sigma_1(\eta)$ are given by the Laplace approximation of Hennig et al (2012) where

$$\mu_{1t}(\eta) = \log \eta_t - \frac{1}{K} \sum_t \log \eta_t,$$

$$\Sigma_{1tt} = \frac{1}{\eta_t}(1 - \frac{2}{K}) + \frac{1}{K^2} \sum_t \frac{1}{\eta_t}.$$

3.2 Inference method

For variational inference of the proposed model, the derivation of the *evidence lower bound* (ELBO) need be given. For document d,

$$\begin{aligned}
\log p(\boldsymbol{w}_d, \boldsymbol{y}_d) &= \log \int_{\boldsymbol{\theta}_d} p(\boldsymbol{w}_d, \boldsymbol{y}_d, \boldsymbol{\theta}_d) \frac{q(\boldsymbol{\theta}_d|\boldsymbol{w}_d)}{q(\boldsymbol{\theta}_d|\boldsymbol{w}_d)} d\boldsymbol{\theta}_d \\
&= \log(\mathbb{E}_{q(\boldsymbol{\theta}_d|\boldsymbol{w}_d)} \frac{p(\boldsymbol{w}_d, \boldsymbol{y}_d, \boldsymbol{\theta}_d)}{q(\boldsymbol{\theta}_d|\boldsymbol{w}_d)}) \\
&\geq \mathbb{E}_{q(\boldsymbol{\theta}_d|\boldsymbol{w}_d)} \log p(\boldsymbol{w}_d, \boldsymbol{y}_d, \boldsymbol{\theta}_d) - \mathbb{E}_{q(\boldsymbol{\theta}_d|\boldsymbol{w}_d)} \log q(\boldsymbol{\theta}_d|\boldsymbol{w}_d) \\
&= \mathbb{E}_{q(\boldsymbol{\theta}_d|\boldsymbol{w}_d)} \log p(\boldsymbol{w}_d|\boldsymbol{\theta}_d) + \mathbb{E}_{q(\boldsymbol{\theta}_d|\boldsymbol{w}_d)} \log p(\boldsymbol{y}_d|\boldsymbol{\theta}_d) - \mathbb{D}_{KL}(q(\boldsymbol{\theta}_d|\boldsymbol{w}_d)||p(\boldsymbol{\theta}_d)).
\end{aligned}$$

Following the VAE framework, $q(\boldsymbol{\theta}_d|\boldsymbol{w}_d)$ is "encoding" $\boldsymbol{w}_d$ to $\boldsymbol{\theta}_d$, and $p(\boldsymbol{w}_d|\boldsymbol{\theta}_d)$ is "decoding" it to reconstruct $\boldsymbol{w}_d$. Meanwhile, we incorporate the prior label information by $p(\boldsymbol{y}_d|\boldsymbol{\theta}_d)$, which is utilized to generate predictive labels.

We can write the modified variational objective function following AVITM, which is one of recent methods for efficient "black box" inference in topic models. Given two inference networks f_μ and f_Σ, for document d, we define the posterior distribution $q(\boldsymbol{\theta}_d)$ to be logistic normal with mean

$$\mu_0 = f_\mu(\boldsymbol{w}_d), \tag{1}$$

and diagonal covariance

$$\Sigma_0 = diag(f_\Sigma(\boldsymbol{w}_d)). \tag{2}$$

The ELBO is written as

$$\mathcal{L}_l(\boldsymbol{w}_d, \boldsymbol{y}_d) = \mathcal{L}_g(\boldsymbol{w}_d) + \mathcal{L}_y(\boldsymbol{w}_d, \boldsymbol{y}_d) + \mathcal{L}_i(\boldsymbol{w}_d), \tag{3}$$

where

$$\begin{aligned}
&\mathcal{L}_g(\boldsymbol{w}_d) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I})}(-\boldsymbol{w}_d^T \log(\sigma(f_g(\sigma(\boldsymbol{\mu}_0 + \Sigma_0^{1/2}\boldsymbol{\epsilon}))))),\\
&\mathcal{L}_y(\boldsymbol{w}_d, \boldsymbol{y}_d) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I})}(-\boldsymbol{y}_d^T \log(\sigma(f_y(\sigma(\boldsymbol{\mu}_0 + \Sigma_0^{1/2}\boldsymbol{\epsilon}))))),\\
&\mathcal{L}_i(\boldsymbol{w}_d) = \sum_{d=1}^{D}(\frac{1}{2}(tr(\Sigma_1^{-1}\Sigma_0) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\Sigma_1^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0) - K + \log\frac{|\Sigma_1|}{|\Sigma_0|})).
\end{aligned} \tag{4}$$

To enhance the effects of prior labels, we give $\mathcal{L}_y$ a hyper-parameters α, and rewrite Equation (3) as

$$\mathcal{L}_l(\boldsymbol{w}_d, \boldsymbol{y}_d) = \mathcal{L}_g(\boldsymbol{w}_d) + \alpha\mathcal{L}_y(\boldsymbol{w}_d, \boldsymbol{y}_d) + \mathcal{L}_i(\boldsymbol{w}_d). \tag{5}$$

To optimize Equation (5), we use stochastic gradient descent using Monte Carlo samples from $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \boldsymbol{I})$.

3.3 Semi-supervised learning

We assume that the training corpus $\mathcal{W}$ consisting of labeled and unlabeled documents, which are denoted as $\mathcal{W}_l$ and $\mathcal{W}_u$ respectively. While learning unlabeled documents, since $\mathcal{L}_y$ in Equation (3) is useless, the proposed model is similar to AVITM. Actually, the learned topics of a document by topic modeling approaches are a low dimensional representative space known as a *manifold*. The *manifold* assumption, i.e., all data points lie on multiple low dimensional manifolds and data points lying on the same manifold often have the same label, extends many algorithms to semi-supervised way(Engelen and Hoos, 2019; Sheikhpour et al, 2017). So learning on unlabeled data $\mathcal{W}_u$ is useful to construct the *manifold*, and then helps to improve the model performance.

To further improve the semi-supervised learning performance of the proposed model, we apply the *low-density* assumption(Grandvalet and Bengio, 2004) on the proposed model with an entropy regularization term,

$$\mathcal{L}_e(\boldsymbol{w}_d) = \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(0,\boldsymbol{I})}(-\sigma(f_y^T(\sigma(\boldsymbol{\mu}_0 + \Sigma_0^{1/2}\boldsymbol{\epsilon}))) \log(\sigma(f_y(\sigma(\boldsymbol{\mu}_0 + \Sigma_0^{1/2}\boldsymbol{\epsilon}))))). \tag{6}$$

Consequently, the object function while learning on unlabeled documents $\mathcal{W}_u$ is

$$\mathcal{L}_u(\boldsymbol{w}_d, \boldsymbol{y}_d) = \mathcal{L}_g(\boldsymbol{w}_d) + \beta\mathcal{L}_e(\boldsymbol{w}_d) + \mathcal{L}_i(\boldsymbol{w}_d), \tag{7}$$

where β is a hyper-parameter.

While training labeled data, the label generative network f_y is optimized by $\mathcal{L}_y$; however, it is optimized by $\mathcal{L}_e$ while unsupervised learning. This leads to the result that the network cannot converge. To address this issue, we optimize the inference networks, i.e., f_μ and f_Σ, as well as the generative network f_g except f_y while unsupervised learning. On the contrary, we optimize all networks using Equation (5) while supervised learning. The learning algorithm of NL-LDA is summarized as Algorithm 3.

Algorithm 3 Learning algorithm of NL-LDA

```
for each epoch do
    while document $d \in \mathcal{W}_l$ do
        Optimize $f_\mu$, $f_\Sigma$, $f_g$, and $f_y$ using Equation (5)
    end while
    while document $d \in \mathcal{W}_u$ do
        Optimize $f_\mu$, $f_\Sigma$, and $f_g$ using Equation (7)
    end while
end for
```

4 Experiments

In this section, we evaluate NL-LDA on several popular typical document classification tasks. Firstly, we introduce the collections and the metrics, then the implementation details are introduced. Lastly, we list the results of our models and compared approaches on supervised and semi-supervised classification.

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 and neural models under the VAE framework.

Yahoo Arts and Health multi-label subsets are from Yahoo Collection(Ueda and Saito, 2002). We randomly selected 6,441 and 8,109 documents for training respectively, ensuring that each label appeared at least once. 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. IMDB dataset[2] contains 50,000 movie reviews, which are either positive or negative sentiments, i.e., there are two classes in this dataset. We divided it equally for training and testing. AGNews is a massive collection of news articles. We used the same training and test data presented by Zhang et al (2015), who choose four largest classes from the original dataset. Each class contains 30,000 training documents and 1,900 testing items. After deleting the stop words, we removed low frequency words and about 8,000 words were retained in each dataset. The datasets are summarized in Table 2.

To compare classification accuracies, we compute the correct classification rate (CCR) as follows:

$$CCR = \frac{1}{n} \sum_{d=1}^{n} \delta(y_d, \hat{y}_d),$$

[1] sklearn.datasets.fetch_20newsgroups

[2] tensorflow.keras.datasets.imdb

Table 2: *Summary of experimental datasets*

	Training Data	Test Data	Labels	Mean Label number per Document
Yahoo Arts	6,441	1,000	19	1.7
Yahoo Health	8,109	1,000	14	1.6
20NewsGroups	11,314	7,532	20	1
IMDB	25,000	25,000	2	1
AGNews	120,000	7,600	4	1

where y_d and $\hat{y}_d$ denote the true and predicted class labels of document d respectively. While classifying single-label datasets, δ is an indicator variable such that $\delta(y_d, \hat{y}_d) = 1$ if $y_d = \hat{y}_d$ and zero otherwise. While classifying multi-label datasets, we define $\delta(y_d, \hat{y}_d) = 1$ if the top-ranked label of $\hat{y}_d$ is in the y_d, and zero otherwise. Larger values imply better performance.

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

$$R = \frac{|y_d \cap \hat{y}_d|}{|y_d|},$$

$$P = \frac{|y_d \cap \hat{y}_d|}{|\hat{y}_d|},$$

$$F1 = \frac{2PR}{P + R}.$$

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.

4.2 Implementation

To encode the posterior distribution over latent variables, we utilize two inference networks f_μ and f_Σ for the mean and log diagonal covariance[3] of the logistic normal respectively. The encoder f_μ and f_Σ are designed as multi-layer neural networks with two shared fully connected layers, as well as an exclusive batch norm layer. The networks utilize soft plus in the hidden layers. For reducing the overtraining effect, we utilize dropout on the second fully connected layer. The encoders utilize bag-of-words representation of documents as their inputs, and the size of output layers is constrained to be the same as the number of topics. To decode the hidden variable sampled from the logistic normal, we utilize two generative networks f_g and f_y for generating documents and labels respectively. The decoder f_g and f_y are designed as two connected layers as well as one batch norm layer. Like encoders, the decoders also utilize soft plus in the hidden layers, and utilize dropout on the input. The output of the decoder f_g is transformed to probabilities of each word by a softmax layer, and the decoder f_y output is transformed to the probabilities of labels after a softmax layer.

The sizes of the hidden layers of f_μ and f_Σ are 2,000, 1,000, as well as f_g and f_y are 1,000, 2,000 respectively. We heuristically set the number of topics to 200, batch size = 200, and the maximum epoch is 2,500, as well as $\alpha = 10$ and $\beta = 1$. To initialize the weights of the networks, we use the Xavier uniform initializer in Tensorflow(Abadi et al, 2016). Adam optimizer is used

[3] We replace the diagonal covariance with the log diagonal covariance in implementation for computation convenience.

with the second exponential decay rate of 0.99, and learning rate = 0.002. After the training of each epoch, we record the value of $\mathcal{L}_y$, which indicates the label prediction performance of the proposed model. We save the network models if the newly $\mathcal{L}_y$ is better than the existed one, i.e., the value of $\mathcal{L}_y$ is less than the recorded one. After training, the saved model is utilized for label prediction.

4.3 Evaluation on supervised document classification

To evaluate the proposed model on supervised classification, we use Yahoo subsets, which represent multi-label classification tasks, as well as 20newsgroups, IMDB, and AGNews datasets, which represent single-label classification tasks. Dependency-LDA(Rubin et al, 2011), TL-LDA(Wang et al, 2020b), and SCHOLAR(Card et al, 2018) are chosen as the baselines. The first two are state-of-the-art supervised topic modeling approaches, and the latter is a neural topic model that incorporates metadata including prior labels. We utilize the original author-provided implementations of Dependency-LDA[4], TL-LDA[5] without modification of hyper-parameters and sampling parameters, as well as SCHOLAR[6] with tone as a label, $vocabulary\_size = 5,000$, $embedding\_dim = 300$. The number of topics and training epochs have impacts on the performance of SCHOLAR, we experimentally set the topic number of 20newsgroups, IMDB, and AGNews to 50, 40, and 20 respectively, as well as the number of training epochs to 2,000, 500, and 200 respectively.

CCR and binary predictions are listed in Table 3 and 4 respectively. Our model performs well in CCR results (Table 3). It gets the best scores among all compared algorithms, including traditional statistical topic models, i.e., Dependency-LDA and TL-LDA, as well as the neural topic model, i.e., SCHOLAR. Furthermore, the proposed approach has significant advantage across all five datasets, while the compared models perform poorly on a particular dataset. For example, TL-LDA performs well on most of the datasets except IMDB, and SCHOLAR performs well on IMDB and AGNews, but poorly on 20NewsGroups. Table 4 demonstrates the proposed NL-LDA performs better than the compared models on Micro and Macro-F1. It is clear that the proposed approach performs well on multi-label document classification.

Table 3: *Experimental CCR results of Yahoo Arts and Health, 20Newsgroups, IMDB, and AGNews datasets, larger values imply better. There are no Yahoo subset results of SCHOLAR, which dose not support multi-label document classification.*

	Yahoo-Arts	Yahoo-Health	20NewsGroups	IMDB	AGNews
NL-LDA	**62.30**	**82.30**	**86.35**	**87.39**	**90.47**
Dep-LDA	56.30	81.80	81.08	80.06	82.19
TL-LDA	62.10	80.60	81.59	77.88	88.70
SCHOLAR	-	-	80.78	86.76	88.64

Bold entries denote the best scores.

4.4 Evaluation on semi-supervised document classification

To evaluate the proposed model performance on semi-supervised classification, we use 20Newsgroups, AGNews, and Yahoo Arts. The training data labels were randomly removed on a particular proportion. Figure 1 shows CCR results of the proposed NL-LDA with semi-supervised and supervised mode, i.e., learning only from labeled data, respectively. It is clear that the

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

[5] https://github.com/willwong76/tllda

[6] https://github.com/dallascard/scholar

Table 4: *Experimental binary prediction results on Micro-F1 (top section) and Macro-F1 (bottom section), larger values imply better.*

	NL-LDA	Dep-LDA	TL-LDA
Yahoo-Arts	**46.95**	43.16	46.80
Yahoo-Health	**63.63**	62.90	62.31
Yahoo-Arts	**51.29**	45.04	50.74
Yahoo-Health	**67.80**	67.02	66.71

Bold entries denote the best scores.

semi-supervised mode performs significantly well under a small amount of labeled data relative to supervised mode of NL-LDA. For example, it is demonstrated that CCR difference between the two modes is about 1% above 20% labels reserved on 20Newsgroups datasets, while the difference is more than 7% under 5% labels reserved, i.e., 565 labeled training samples. AGNews and Yahoo Arts have similar trends. It is shown that the semi-supervised mode scores about 17% higher than the supervised mode under 0.2% labels reserved, i.e., 240 labeled documents, on AGNews dataset; and about 13% higher under 5% labels reserved, i.e., 322 documents, on Yahoo Arts.

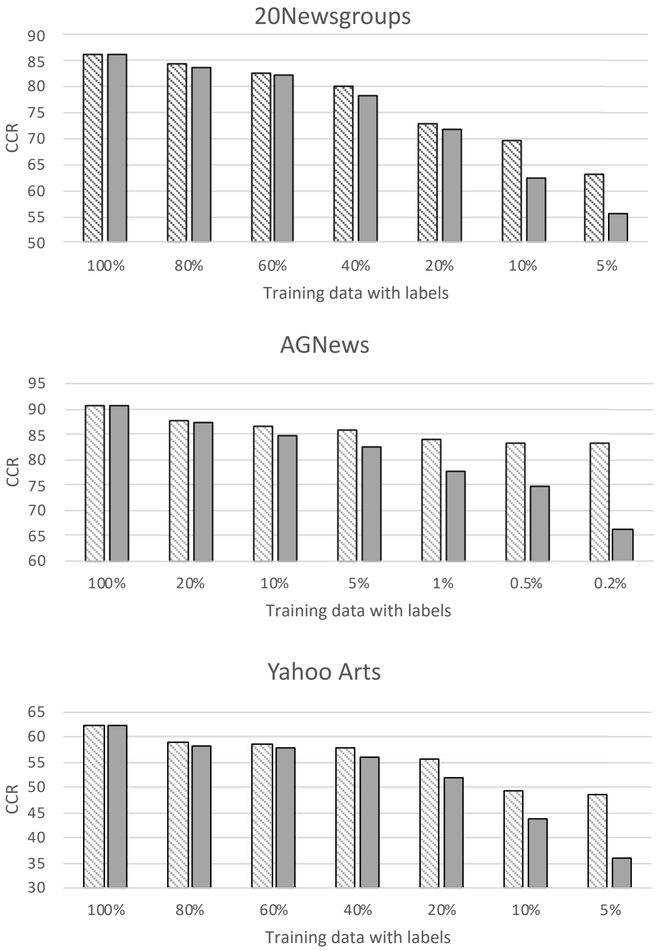

Fig. 1: *Experimental CCR results of semi-supervised (twill) and supervised (grey) modes of NL-LDA with different labeled data proportion.*

We further select existing results of semi-supervised topic modeling approaches to compare, including a neural topic model and two traditional statistical topic modeling approaches.

S-VAE-GM(Zhou et al, 2020), which is a semi-supervised topic model under the VAE framework with Gaussian mixture assumption, demonstrates competitive performances on 20newsgroups, IMDB, and AGNews datasets. To compare with reported results of S-VAE-GM, we computed label-pivoted F1-scores, i.e. F1-score computed for specific labels, under 20% labels reserved on three datasets, and utilized box plots following Zhou et al (2020). In Figure 2, box plots indicate distributions of label-pivoted F1-scores of our model as well as corresponding results of S-VAE-GM. Obviously, even considering the randomness of the training and test data selection, our model performs better than S-VAE-GM.

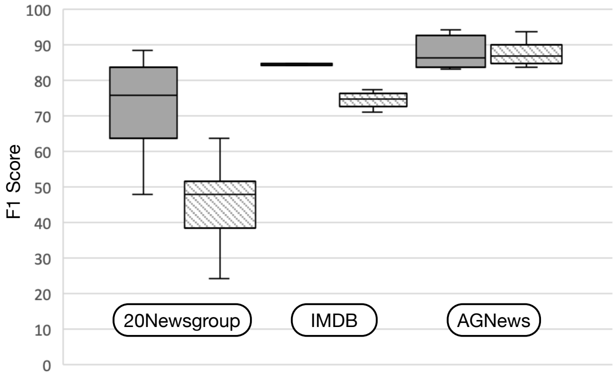

Fig. 2: *Comparison of NL-LDA label-pivoted F1 scores (grey), with results of S-VAE-GM (twill) from Zhou et al (2020) on 20Newsgroups, IMDB, and AGNews, larger values imply better.*

MCCTM(Soleimani and Miller, 2017), is a class-based mixture of topic models for classifying documents using both labeled and unlabeled examples. The reported results show it achieves better CCR than some state-of-the-art semi-supervised topic modeling approaches. ssLDA is an extension of Supervised LDA to a semi-supervised framework for document classification(Mcauliffe and Blei, 2007; Wang et al, 2009). To compare with reported results of MCCTM and ssLDA, we use 20Newsgroups following Soleimani and Miller (2017), i.e., 13,105 and 5,063 documents in the training and test sets respectively. Table 5 lists the results, which show the proposed NL-LDA has significant advantages to the compared approaches on 7/8 reserved label proportions. It scores about 10% higher than the second good results above 30% labels reserved. However, MCCTM performs best under a very small amount of labeled data, i.e., 5% labeled samples. It gets 2% higher score than ours.

Table 5: *Comparison of CCR results on 20Newsgroups with results of MCCTM and ssLDA from Soleimani and Miller (2017), larger values imply better.*

	90%	70%	50%	30%	20%	10%	5%
NL-LDA	**83**	**82**	**82**	**80**	**76**	**74**	64
MCCTM	74	72	71	70	68	67	**66**
ssLDA	67	67	67	66	65	64	61

Bold entries denote the best scores.

5 Discussion

The experimental results clearly demonstrate that the proposed NL-LDA has significant advantages on supervised document classification relative to the compared topic modeling approaches, including traditional statistical and neural topic models. Statistical topic modeling approaches based on LDA have been widely applied in the field of document classification, including multi-label and single-label multi-class classification; however, they require deriving specific inference methods for customizations of LDA. Meanwhile, our work not only supports single-label and multi-label document classification, but also dose not require specific derivations for customized model modification by the VAE framework.

One striking aspect of the experimental results is NL-LDA outperforms SCHOLAR, which is also a neural topic modeling approach based on the VAE framework. One difference between the two models is the method incorporating prior labels. SCHOLAR incorporates prior information to the input of the model. However, the proposed model incorporates prior labels by the objective function $\mathcal{L}_y$ of the model, which suggests direct effects on the output of neural networks, and helps to improve the performance while back propagation training. This result is different from traditional statistical topic modeling approaches. Roughly speaking, incorporating the prior side information, and then generating the words has better predictive ability in statistical supervised topic models(Soleimani and Miller, 2017).

The other difference is the proposed model can adjust the weight of prior knowledge by the hyper-parameter α, but SCHOAR has no corresponding mechanism. Furthermore, to test data without prior labels, SCHOLAR needs specially prepare an encoder without labels or replace the label vector for the input of inference network with a vector of all zeros. This is a cumbersome procedure. On the contrary, our model can easily used for prediction without any modification.

Another interesting aspect of results is the proposed NL-LDA performs significantly well on semi-supervised document classification. Firstly, the proposed approach can make good use of unlabeled data to improve the model performance, especially under a small amount of labeled data (Figure 1). Secondly, NL-LDA scores better than the compared existing semi-supervised topic modeling approaches, including neural and classical statistical topic models (Figure 2, Table 5). The results suggest the two assumptions, i.e., the *manifold* and *low-density* assumption, have helped to improve the performance of semi-supervised document classification. Our model applies the *low-density* assumption by the entropy regularization term, which is weighted by the hyper-parameter β. To demonstrate the effects of the entropy regularization term on semi-supervised document classification, we use Yahoo subsets, 20Newsgroups, IMDB, and AGNews under 10% labels reserved. Table 6 lists the results. It is clear that the entropy regularization term has positive effects on semi-supervised classification tasks.

Table 6: *Experimental CCR results of NL-LDA ($\beta = 1$) and NL-LDA without the entropy regularization term ($\beta = 0$) on Yahoo Arts and Health, 20Newsgroups, IMDB, and AGNews datasets under 10% labeled samples, larger values imply better.*

	Yahoo-Arts	Yahoo-Health	20NewsGroup	IMDB	AGNews
$\beta = 1$	**49.30**	**71.60**	**69.51**	**83.30**	**86.64**
$\beta = 0$	47.80	69.10	68.31	82.92	86.05

Bold entries denote the best scores.

However, one limitation of neural topic models is many network parameters need be trained. The traditional statistical semi-supervised modeling approach, i.e., MCCTM, performs 2% higher score than our model under a very small amount of labeled data on 20Newsgroups dataset. Another statistical model, ssLDA, only gets 3% lower score than ours, while the proposed model scores 15% higher than ssLDA above 50% labeled data. These results suggest statistical approaches models better than neural topic models under a very small amount of data because neural models have more parameters, and need a certain amount of labeled data to train.

The proposed model may be suggested to a generalized supervised topic model framework, which can incorporate other metadata as a new parallel generative process like f_g and f_y. In this paper, we use bag-of-words representation as the model inputs. Furthermore, we can utilize other word embeddings obtained from pre-trained model, e.g. BERT(Devlin et al, 2018), to further improve the model classification performance.

6 Conclusion

Statistic topic models based on LDA have been widely developed in the field of document classification. Some of them can support semi-supervised classification and achieve competitive results with state-of-the-art approaches. However, these customizations of LDA have a drawback that small changes to the modeling procedure result in a re-derivation of the inference algorithm leading to the lack of applications. To address this issue, we propose a novel semi-supervised topic model, i.e., Neural Labeled LDA (NL-LDA), based on the VAE framework, which is a black box inference method. An additional label generative network with a weight parameter is employed to incorporate prior knowledge, and results in flexibility, simplicity, and outstanding performance. NL-LDA supports semi-supervised learning based on two assumptions, i.e., the *manifold* assumption and *low-density* assumption. It is worth noting that we utilize different object functions and optimize different sub networks while learning from labeled and unlabeled documents respectively, and the proposed model has a consistent and concise inference method while semi-supervised learning and predicting.

We evaluate the proposed model and compared methods, including traditional statistical and neural topic models, on supervised and semi-supervised document classification. The results show the proposed model has significant advantages across all experimental datasets, including single-label and multi-label datasets. The proposed NL-LDA performs significantly well on semi-supervised classification, especially under a small amount of labeled data.

In conclusion, the proposed NL-LDA has three significant advantages. Firstly, our model has concise inference method by the VAE framework. Secondly, our model has good flexibility because of incorporating metadata by a generative network. Thirdly, our model performs well on supervised and semi-supervised document classification.

In the future, we intend to further improve the performance of the proposed model by adopting pre-trained language models, and plan to apply the models to some other applications, e.g., news video segmentation and summarization.

Declarations

Funding

This work was funded in part by the National Natural Science Foundation of China under Grant No. 61772352; the Science and Technology Planning Project of Sichuan Province under Grant No. 2019YFG0400, 2018GZDZX0031, 2018GZDZX0004, 2017GZDZX0003, 2018JY0182, 19ZDYF1286.

Conflicts of interest/Competing interests

The authors declared that they have no conflicts of interest/competing interests to this work.

Availability of data and material

The datasets used during the current study are publicly available.

Code availability

The source code used in the current study is available from the first author or corresponding author on reasonable request.

Authors' contributions

Wei Wang: Conceptualization, Investigation, Methodology, Software, Datasets preparation, Formal analysis, Writing - original draft. **Bing Guo**: Methodology, Formal analysis, Investigation, Writing - review and editing, Supervision, Funding acquisition. **Yan Shen**: Writing - review and editing, Supervision, Funding acquisition. **Han Yang**: Validation, Software, Datasets preparation, Writing - review and editing. **Yaosen Chen**: Validation, Datasets preparation, Formal analysis, Visualization. **Xinhua Suo**: Validation, Formal analysis.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant No. 61772352; the Science and Technology Planning Project of Sichuan Province under Grant No. 2019YFG0400, 2018GZDZX0031, 2018GZDZX0004, 2017GZDZX0003, 2018JY0182, 19ZDYF1286.

References

Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al (2016) Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), pp 265–283

Blei DM (2012) Probabilistic topic models. Communications of the ACM 55(4):77–84

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of machine Learning research 3(Jan):993–1022

Burkhardt S, Kramer S (2019a) Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. Journal of Machine Learning Research 20(131):1–27

Burkhardt S, Kramer S (2019b) A survey of multi-label topic models. ACM SIGKDD Explorations Newsletter 21(2):61–79, DOI 10.1145/3373464.3373474

Card D, Tan C, Smith NA (2018) Neural models for documents with metadata. arXiv preprint arXiv:170509296

Chen J, Zhang K, Zhou Y, Chen Z, Liu Y, Tang Z, Yin L (2019) A novel topic model for documents by incorporating semantic relations between words. Soft Computing 24(15):11407–11423, DOI 10.1007/s00500-019-04604-0

Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:181004805

Engelen JEV, Hoos HH (2019) A survey on semi-supervised learning. Machine Learning 109(2):373–440, DOI 10.1007/s10994-019-05855-6

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

Grandvalet Y, Bengio Y (2004) Semi-supervised learning by entropy minimization. Advances in neural information processing systems 17:529–536

Griffiths TL, Steyvers M (2004) Finding scientific topics. Proceedings of the National Academy of Sciences 101(Supplement 1):5228–5235, DOI 10.1073/pnas.0307752101

Hennig P, Stern D, Herbrich R, Graepel T (2012) Kernel topic models. In: Artificial Intelligence and Statistics, pp 511–519

Jelodar H, Wang Y, Yuan C, Feng X, Jiang X, Li Y, Zhao L (2018) Latent dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications 78(11):15169–15211, DOI 10.1007/s11042-018-6894-4

Joo W, Lee W, Park S, Moon IC (2020) Dirichlet variational autoencoder. Pattern Recognition 107:107514, DOI 10.1016/j.patcog.2020.107514

Kingma DP, Welling M (2013) Auto-encoding variational bayes. arXiv preprint arXiv:13126114

Kingma DP, Rezende DJ, Mohamed S, Welling M (2014) Semi-supervised learning with deep generative models. arXiv preprint arXiv:14065298

Mcauliffe J, Blei D (2007) Supervised topic models. Advances in neural information processing systems 20:121–128

Miao Y, Yu L, Blunsom P (2016) Neural variational inference for text processing. In: International conference on machine learning, pp 1727–1736

Pavlinek M, Podgorelec V (2017) Text classification method based on self-training and lda topic models. Expert Systems with Applications 80:83–93

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

Rezende DJ, Mohamed S, Wierstra D (2014) Stochastic backpropagation and approximate inference in deep generative models. arXiv preprint arXiv:14014082

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

Sheikhpour R, Sarram MA, Gharaghani S, Chahooki MAZ (2017) A survey on semi-supervised feature selection methods. Pattern Recognition 64:141–158, DOI 10.1016/j.patcog.2016.11.003

Soleimani H, Miller DJ (2016) Semi-supervised multi-label topic models for document classification and sentence labeling. In: Proceedings of the 25th ACM international on conference on information and knowledge management, pp 105–114

Soleimani H, Miller DJ (2017) 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, DOI 10.1007/s10044-017-0629-4

Srivastava A, Sutton C (2017) Autoencoding variational inference for topic models. arXiv preprint arXiv:170301488

Teh Y, Newman D, Welling M (2006) A collapsed variational bayesian inference algorithm for latent dirichlet allocation. Advances in neural information processing systems 19:1353–1360

Ueda N, Saito K (2002) Parametric mixture models for multi-labeled text. Advances in neural information processing systems 15:737–744

Wang C, Blei D, Li FF (2009) Simultaneous image classification and annotation. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp 1903–1910

Wang D, Thint M, Al-Rubaie A (2012) Semi-supervised latent dirichlet allocation and its application for document classification. In: 2012 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, IEEE, vol 3, pp 306–310

Wang R, Hu X, Zhou D, He Y, Xiong Y, Ye C, Xu H (2020a) Neural topic modeling with bidirectional adversarial training. arXiv preprint arXiv:200412331

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

Xu W, Sun H, Deng C, Tan Y (2017) Variational autoencoder for semi-supervised text classification. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol 31(1)

Yang Y (1999) An evaluation of statistical approaches to text categorization. Information retrieval 1(1-2):69–90

Zhang H, Chen B, Guo D, Zhou M (2018) Whai: Weibull hybrid autoencoding inference for deep topic modeling. arXiv preprint arXiv:180301328

Zhang X, Zhao J, LeCun Y (2015) Character-level convolutional networks for text classification. Advances in neural information processing systems 28:649–657

Zhang Y, Wei W (2014) A jointly distributed semi-supervised topic model. Neurocomputing 134:38–45

Zhou C, Ban H, Zhang J, Li Q, Zhang Y (2020) Gaussian mixture variational autoencoder for semi-supervised topic modeling. IEEE Access 8:106843–106854, DOI 10.1109/access.2020.3001184

Figures

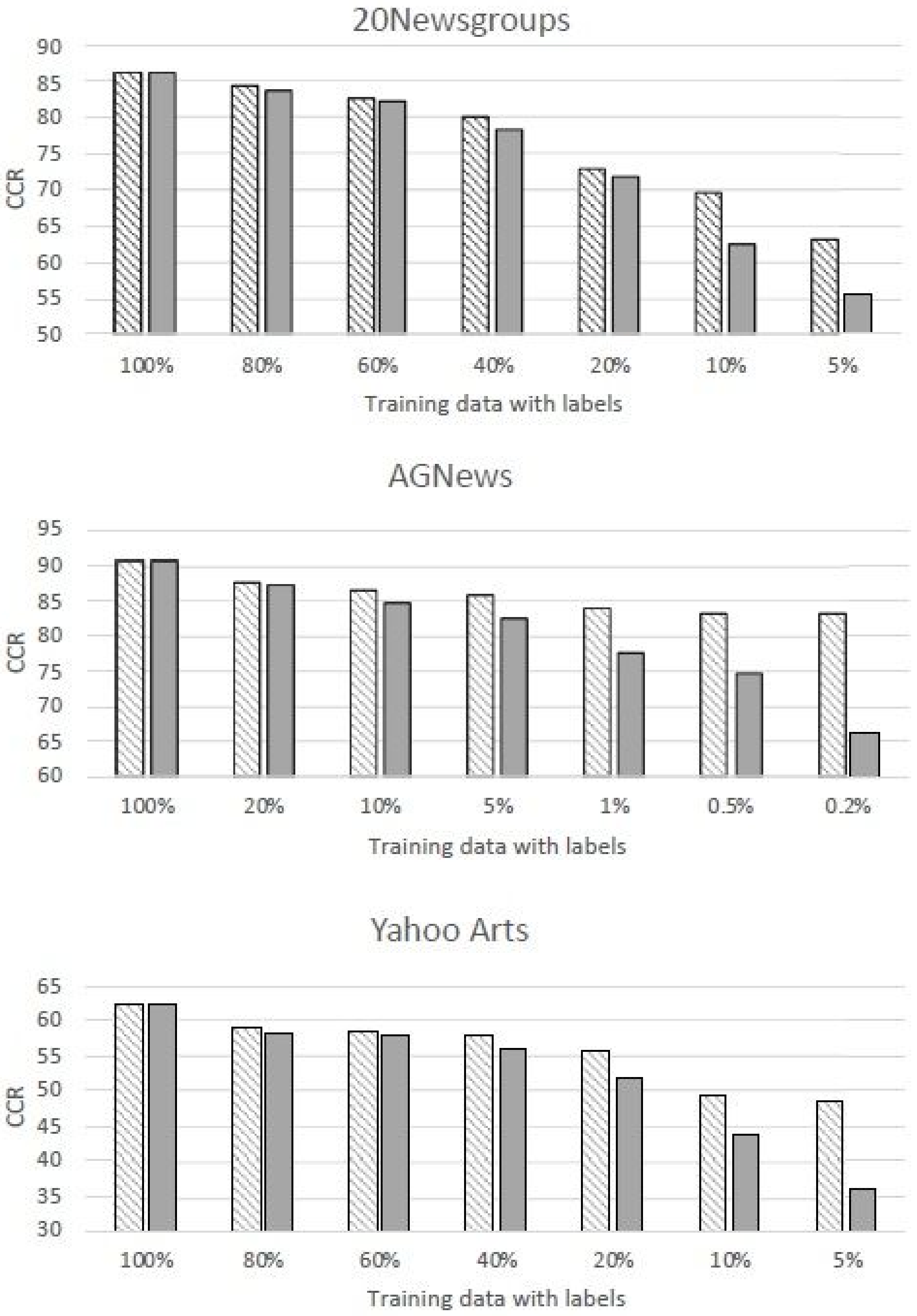

Figure 1

Experimental CCR results of semi-supervised (twill) and supervised (grey) modes of NL-LDA with different labeled data proportion.

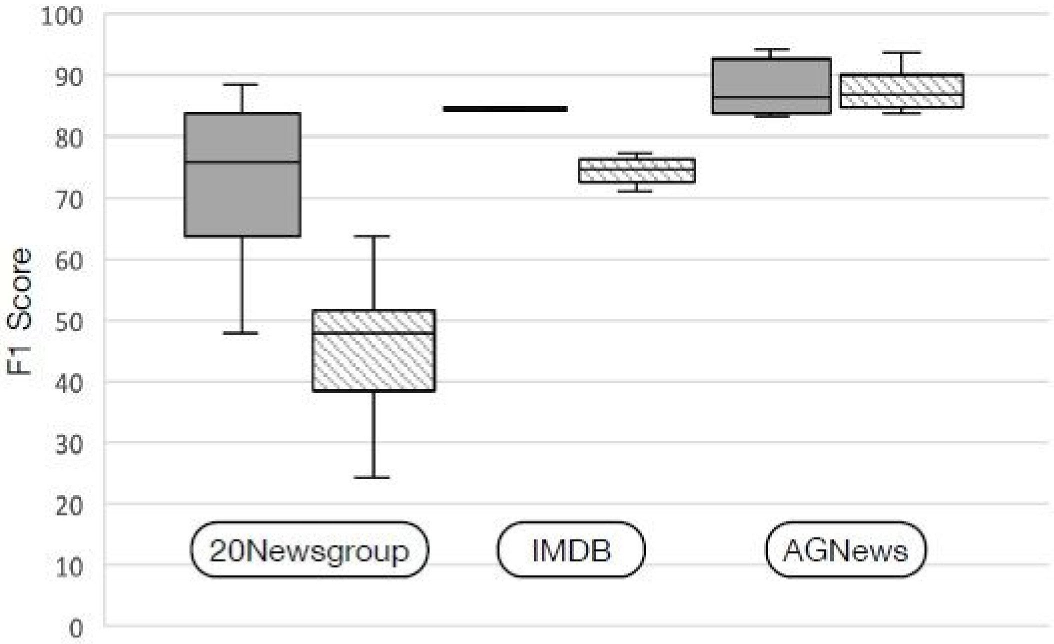

Figure 2

Comparison of NL-LDA label-pivoted F1 scores (grey), with results of S-VAE-GM (twill) from Zhou et al (2020) on 20Newsgroups, IMDB, and AGNews, larger values imply better.