KDSTM: NEURAL SEMI-SUPERVISED TOPIC MODELING WITH KNOWLEDGE DISTILLATION

Weijie Xu[1], Xiaoyu Jiang[1], Jay Desai[1], Bin Han[2], Fuqin Yan[1] & Francis Iannacci[1]
[1] Amazon Inc. [2] University of Washington
{weijiexu,billyjia,jdesa,fqinyan,iannacci}@amazon.com
{binhan96816}@gmail.com

ABSTRACT

In text classification tasks, fine tuning pretrained language models like BERT and GPT-3 yields competitive accuracy; however, both methods require pretraining on large text datasets. In contrast, general topic modeling methods possess the advantage of analyzing documents to extract meaningful patterns of words without the need of pretraining. To leverage topic modeling's unsupervised insights extraction on text classification tasks, we develop the Knowledge Distillation Semi-supervised Topic Modeling (KDSTM). KDSTM requires no pretrained embeddings, few labeled documents and is efficient to train, making it ideal under resource constrained settings. Across a variety of datasets, our method outperforms existing supervised topic modeling methods in classification accuracy, robustness and efficiency and achieves similar performance compare to state of the art weakly supervised text classification methods.

1 INTRODUCTION

The current state-of-the-art language modeling methods often require transfer learning Brown et al. (2020), large amount of labels Yang et al. (2019) and pretrained embeddings Cao et al. (2020). Consequently, they are difficult to apply in low resource settings, where many endangered languagesAustin & Sallabank (2011) lack both pre trained language models and sufficient labeled documents. For semi-supervised method Meng et al. (2018) tailored to limited label scenario, it is time consuming to both tune and train.

Topic modeling is an unsupervised method for discovering latent structure within the training document sets and achieves great empirical performance in many fieldsBlei et al. (2009), including finance Aziz et al., healthcare Bhattacharya et al. (2017), education Zhao et al. (2020b), marketing Reisenbichler (2019) and social science Roberts et al. (2013). Jelodar et al. (2018) provides a survey on the applications of topic modeling.

Latent Dirichlet Allocation (LDA) Blei et al. (2003) is the most fundamental topic modeling approach based on Bayesian inference on Markov chain Monte Carlo (MCMC) and variational inference; however, it is hard to be expressive or capture large vocabularies. Neural topic model (NTM) Miao et al. (2018) leverages auto-encoding Kingma et al. (2014) framework to approximate intractable distributions over latent variables. Recently, embedded topic model (ETM) Dieng et al. (2020) uses word embedding during the reconstruction process to make topic more coherent and reduce the influence of stop words. The goal of unsupervised topic modeling methods Blei et al. (2003); Teh et al. (2006); Miao et al. (2018); Gemp et al. (2019) is to maximize the probability of the observed data, resulting in the tendency to identify obvious and superficial aspects of a corpus. To incorporate users' domain knowledge of documents into the model, supervised modeling Blei & McAuliffe (2010); Zhu et al. (2012); Wang & Yang (2020a) has been studied. However, supervised methods do not perform well when the labeled set is small.

In this work, we propose knowledge distillation semi-supervised topic modeling (KDSTM), which only requires a few labeled documents for each topic as input. KDSTM utilizes knowledge distillation and optimal transport to guide topic extraction with seed documents. It achieves state of the art results when benchmarking with supervised topic modeling and weakly supervised text classification methods.

Advantages of KDSTM are summarized as follows:

- KDSTM is a novel architecture which incorporates knowledge distillation and optimal transport into the neural topic modeling framework.
- KDSTM consistently achieves better topics classification performance on different datasets when compared to supervised topic modeling methods or weakly supervised text classification methods.
- KDSTM only requires a limited number of labeled documents as input, making it more practical in low resource settings.
- KDSTM does not rely on any transfer learning or pre trained language models. The embedding is trained on the dataset, making it suitable for less common/endangered languages.
- KDSTM is efficient to train and fine-tune compared to existing methods. This makes it suitable to be trained and run inference on resource constrained devices.

2 PRELIMINARY

Supervised Topic Modeling Since topic modeling reduces the dimensionality of the text, the learned low dimensional topic distributions can be used in the downstream tasks. Blei & McAuliffe (2010) adds a response variable associated with each document and assumes that the variable can be fitted by Gaussian distribution to make it tractable. Labeled LDA Ramage et al. (2009) assumes that each document can be associated with one topic and uses this information to create the model. Recently, BP-SLDA Chen et al. (2015) uses back propagation to make LDA supervised. Dieng Dieng et al. (2017) incorporates RNN with LDA to make latent variables more suitable for downstream tasks. TAM Wang & Yang (2020b) combines GSM Liu et al. (2019) and RNN Sherstinsky (2020) to do the supervised topic modeling. To be specific, it uses GSM to fit a document generative process and estimates document specific topic distribution. It uses GRU Chung et al. (2014) to encoded word tokens. After that, it jointly optimizes two components by using an attention mechanisms. Recently, topic modeling is also combined with Siamese network Huang et al. (2018) to achieve better prediction performance. However, these methods' performances drops when training set is small.

Optimal Transport To avoid matching labels with multiple topics , we consider the optimal transport distance Chen et al. (2019); Torres et al. (2021), which has been widely used for comparing the distribution of probabilities. Specifically, let $U(r, c)$ be the set of positive $m \times n$ matrices for which the rows sum to r and the columns sum to c: $U(r, c) = \{P \in R_{>0}^{m \times n} | P1_t = r, P^T 1_s = c\}$ For each position t, s in the matrix, it comes with a cost $M_{t,s}$. Our goal is to solve $d_M(r, c) = min_{P \in U(r,c)} \sum_{t,s} P_{t,s} M_{t,s}$. To make distribution homogeneous Cuturi (2013), we let

$$d_M^{\lambda}(r, c) = min_{P \in U(r,c)} \sum_{t,s} P_{t,s} M_{t,s} - \frac{1}{\lambda} h(P) \tag{1}$$

, where $h(P) = -\sum_{t,s} P_{t,s} \log P_{t,s}$. Optimal Transport induces good robustness and semantic invariance in NLP related tasks Chen et al. (2019) or topic modeling Zhao et al. (2020a); Xu et al. (2018).

Knowledge Distillation Knowledge distillation is the process of transferring knowledge from a large model to a smaller one. Given a large model trained for a specific classification task, the final layer of the network is a softmax in the form:

$$y(x|\tau) = \frac{e^{\frac{s(x)}{\tau}}}{\sum_j e^{\frac{s(x)}{\tau}}} \tag{2}$$

where t is the temperature parameter. The softmax operator converts the logit values s to pseudo-probabilities. Knowledge distillation consists of training a smaller network, called the distilled model, on a separate dataset called the transfer set. Cross entropy is used as the loss function between the output of the distilled model and output produced by the large model on the transfer set, using a high value of softmax temperature τ for both models. Hinton et al. (2015)

$$E(\mathbf{x}|\tau) = -\sum_i \hat{y}_i(\mathbf{x}|\tau) \log y_i(\mathbf{x}|\tau) \tag{3}$$

where $\hat{y}_i$ is generated by the large model and y_i is generated by the distilled model. Instead of using the prediction itself, few methods leverage similarity Tung & Mori (2019); Chen et al. (2018); Passalis et al. (2020) or features Romero et al. (2015); Passban et al. (2020); Chen et al. (2021) as guidance.

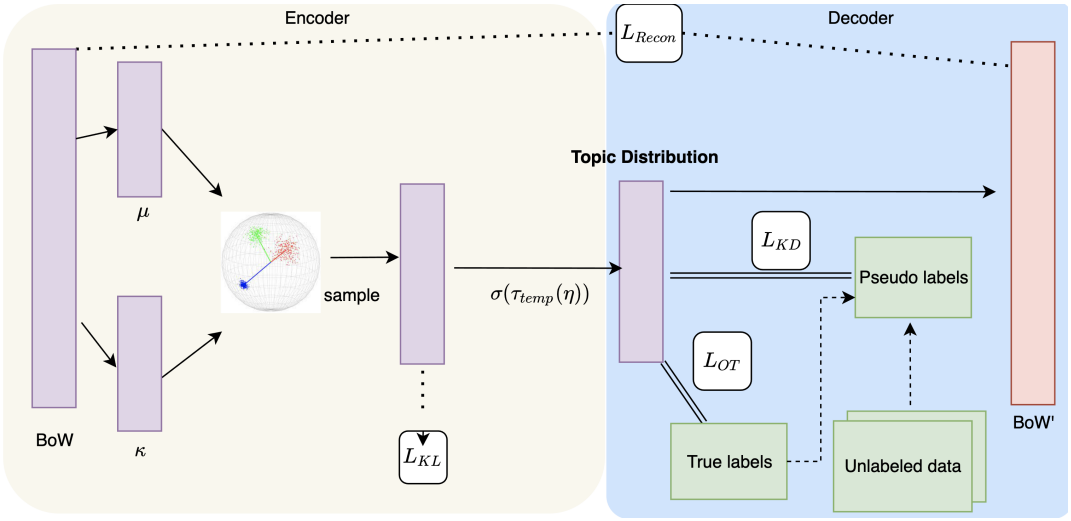

Figure 1: The Architecture of KDSTM with four main loss function including reconstruction loss, KL divergence, optimal transport loss and knowledge distillation loss

3 METHOD

The encoder network ϕ encodes the bag of words representation of any document x_d and output parameters for latent distribution which can be used to sample the topic distribution t_d. Following Dieng et al. (2020), the decoder is represented by a vocabulary embedding matrix e_W and a topic embedding matrix e_T. We use spherical word embedding Meng et al. (2019) to create e_W. We train it on the corpus and keep it fixed during the training. W is the corpus and T contains all topics. We also use VMF distribution (Appendix C) instead of normal distribution as the latent distribution for better clusterability Xu & Durrett (2018); Davidson et al. (2018); Reisinger et al. (2010); Batmanghelich et al. (2016); Ennajari et al. (2021). In this notation, our modified ETM's algorithm can be described as follows: for every document d, 1) Generate t_d using sampled direction parameter μ and scale parameter κ from ϕ. 2) Reconstruct bag of words by $t_d \times softmax(e_T e_W^T)$. The goal of ETM is to maximize the marginal likelihood of the documents: $\sum_{d=1}^{D} \log p(x_d|e_T, e_W)$. To make it tractable, the loss function combines reconstruction loss with KL divergence. The description of notation can be found in Table 1 in appendix.

In KDSTM, we adopt optimal transport to assign topics to labels. Each entry in M is defined as $M_{g,t} = 1 - mean_{x \in g} \phi_t(x)$ where g is one of the labeled documents' group. Let $M_{g,t}$ represent the weights of words in labeled documents in group g on topic t. We use sinkhorn distance as loss function and give high entropy λ to make sure that each labeled comment falls into separate topics. Thus,

$$L_{OT} = min_{P \in U(|T|,|G|)} \sum_{t,g} P_{t,g} M_{t,g} - \frac{1}{\lambda} h(P) \tag{4}$$

where $|G|$ is the number of labeled groups and g represents one group of labeled comments.

The next step is to ensure that a test input text that is similar to our labeled document has high probability of being classified as the same topic. To achieve that, we borrow the idea from similarity and feature based knowledge distillation Mun et al. (2018); Tung & Mori (2019). To be specific, we first train the unsupervised topic modeling till convergence. Then, we store the direction parameters

μ from ϕ and use it to calculate cosine similarity s between unlabeled and labeled documents. We use the maximum similarity in each labeled group as guidance for knowledge distillation. The benefits of this approach include: 1) latent distribution from unsupervised topic model can be used for label classification Lacoste-Julien et al. (2008); Chen et al. (2015). These distributions can serve as teachers. 2) we do not need a separate and larger model, making it more resource efficient.
For each text x_d, we find the most similar document i in each labeled group g and their similarity $s_i(x)$. Then we define

$$\hat{y^g}(x;\tau) = \frac{e^{\frac{\max_{i\in g} s_i(x)}{\tau}}}{\sum_{g\in G} e^{\frac{\max_{i\in g} s_i(x)}{\tau}}} \tag{5}$$

where G is all labeled text groups. The knowledge distillation loss is measured by :

$$L_{KD} = -\tau^2 \sum_{d\in D}\sum_{g\in G} I(s_g(x_d) \geq thresh)\hat{y^g}(x_d;\tau) log(\phi(x_d)) \tag{6}$$

We use the indicator function $I(s_i(x_d) \geq thresh)$. Since we only have few labels available, some of documents may not be related to any of the labeled documents. Thus, we only care about documents that are relevant to existing labels to provide a better teaching experience. We split training into 3 stages: 1) we train the standard topic modeling with KL Divergence and reconstruction loss $L_{KD} + L_{Recon}$ till convergence. We calculate similarity matrix s after this stage. 2) We add optimal transport loss $L_{KD} + L_{Recon} + \alpha L_{OT}$ and train for few epochs. 3) We add knowledge distillation loss $L_{KD} + L_{Recon} + \alpha L_{OT} + \beta L_{KD}$ and train for few epochs. In practice, step 2 and step 3 are less time consuming and thus, this helps user optimize their labels in online settings. The architecture is illustrated in Figure 1. Hoyle et al. (2020) also leverages knowledge distillation, but uses bag of words representation, is of unsupervised nature, and using a BERT-based auto-encoder as the teacher.

4 EXPERIMENTS

Settings In this section, we report the experimental results for our methods and three additional state of the arts methods (WestClass, LLDA, TAM). To form the vocabulary, we keep all words that appear more than a certain number of times and vary the threshold from 20 to 100 depending on the size of the dataset. We remove documents that are less than 2 words. We also remove stop words, digits, time and symbols from vocabulary and use a fully-connected neural network with two hidden layers of [256, 64] unit and ReLU as the activation function followed by a dropout layer(rate = 0.5). The hyperparameter setting used for all baseline models and vNTM are similar to Burkhardt & Kramer (2019). We use AdamKingma & Ba (2017) as the optimizer with learning rate 0.002 and use batch size 256. We use Smith & Topin (2018) as scheduler and use learning rate 0.01 for maximally iterations equal to 50. For each run, we sample 5 documents per class and use them as inputs, calculate performance metrics on the rest of unlabeled documents, run each algorithm 10 times and report the result in the bin plot. We report accuracy, aucroc and averaged micro f1 score. For hyperparameters, we use $\lambda = 50$, $\alpha = \beta = 10$, $thresh = 0$ and $\tau = 1$ for our method and perform moderate tuning on parameters presented in the original papers of other methods. Our code is written in PyTorch and all the models are trained on AWS using ml.p2.8xlarge (NVIDIA K80).

Datasets (1) **AG's News** We use AG's News dataset from Zhang et al. (2016). It has 4 classes and 30000 documents per class. Class categories include World, Sports, Business and Sci/Tech for evaluation; (2) **DBLP** Tang et al. (2008; 2010) dataset consists of bibliography data in computer science. DBLP selects a list of conferences from 4 research areas, database, data mining, artificial intelligence, and computer vision. With a total 60,744 papers averaging 5.4 words in each title, DBLP tests the performance on small text corpus. See Appendix D. (3) **20News** Lang (1995) is a collection of newsgroup posts. We only select 4 categories here. Compared to the other two datasets, 4 categories newsgroup is small so that we can check the performance of our methods on small datasets.

Methods (1) **WestClass** Meng et al. (2018): This method is a weakly-supervised neural text classification method. The weak supervision source can come from any of the three sources: label surface names, class related keywords, and labeling documents. Output will be document labels consistent with cluster of inputs. We will use this method to benchmark the performance on document classification accuracy. (2) **L-Label** Ramage et al. (2009): This method accepts labeled documents and is

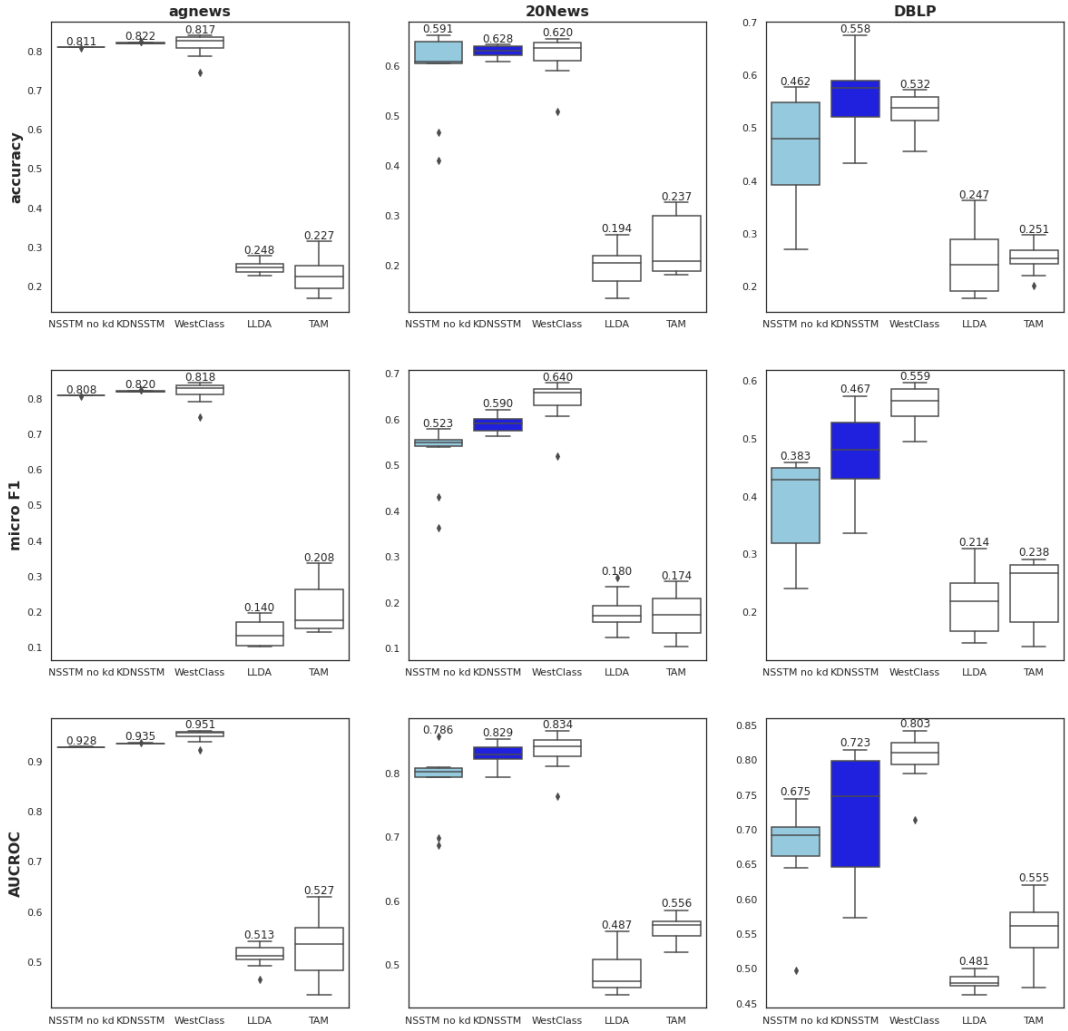

Figure 2: From left to right, is NSSTM without knowledge distillation (Sky Blue), KDNSSTM (Blue), WestClass, LabeledLDA, TAM. We average result as text on top of each bin plot

used to compare KDSTM. To better benchmark the performance, we fine-tune alpha, eta and number of iterations to get the best accuracy performance. (3) **TAM** Wang & Yang (2020a): This method achieves the better performance than existing supervised topic modeling methods. To benchmark the performance for a strongly performing model, we fine-tune the dimension of GRU gate, learning rate and number of epochs. See Appendix E.

Results As can be seen from Figure 2, KDSTM consistently performs significantly better than existing supervised topic modeling methods on all 3 classification metrics (accuracy, F1, and AUC). On standard dataset like AG's News, our performance is higher than WestClass on accuracy and micro F1. Despite similar overall performance, our method is on average 4 times faster than WestClass (Table 2 in appendix). If we finetune τ, our method can further improve which show its potential to beat WestClass (Table 3 in appendix). Our method has high variance in DBLP where each document has on average 5.4 words. We also compare the performance without knowledge distillation. Knowledge distillation increases the performance of the model on all 3 metrics. Knowledge distillation provides less benefit for small dataset such as 20News.

5 CONCLUSION

In this work, we develop Knowledge Distillation Semi-supervised Topic Modeling (KDSTM) using knowledge distillation and optimal transport. Our method achieves improved performance across several classification metrics compared to existing supervised topic modeling methods and more efficient than existing weakly supervised text classification methods. Our method does not require transfer learning or pretrained embeddings and is faster to train and fine-tune, making it ideal in low resource scenarios. For future work, we will extend this work to include sequential information to further improve its performance and stability.

REFERENCES

Peter K Austin and Julia Sallabank. *The Cambridge handbook of endangered languages*. Cambridge University Press, 2011.

Saqib Aziz, Michael Dowling, Helmi Hammami, and Anke Piepenbrink. Machine learning in finance: A topic modeling approach. *European Financial Management*, n/a(n/a). doi: https://doi.org/10.1111/eufm.12326. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/eufm.12326.

Kayhan Batmanghelich, Ardavan Saeedi, Karthik Narasimhan, and Sam Gershman. Nonparametric spherical topic modeling with word embeddings, 2016.

Moumita Bhattacharya, Claudine Jurkovitz, and Hagit Shatkay. Identifying patterns of associated-conditions through topic models of electronic medical records. *CoRR*, abs/1711.10960, 2017. URL http://arxiv.org/abs/1711.10960.

David M. Blei and Jon D. McAuliffe. Supervised topic models, 2010.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

David M. Blei, Thomas L. Griffiths, and Michael I. Jordan. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies, 2009.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sophie Burkhardt and Stefan Kramer. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *Journal of Machine Learning Research*, 20(131):1–27, 2019. URL http://jmlr.org/papers/v20/18-569.html.

Steven Cao, Nikita Kitaev, and Dan Klein. Multilingual alignment of contextual word representations. *arXiv preprint arXiv:2002.03518*, 2020.

Defang Chen, Jian-Ping Mei, Yuan Zhang, Can Wang, Yan Feng, and Chun Chen. Cross-layer distillation with semantic calibration, 2021.

Jianshu Chen, Ji He, Yelong Shen, Lin Xiao, Xiaodong He, Jianfeng Gao, Xinying Song, and Li Deng. End-to-end learning of lda by mirror-descent back propagation over a deep architecture, 2015.

Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. Improving sequence-to-sequence learning via optimal transport, 2019.

Yuntao Chen, Naiyan Wang, and Zhaoxiang Zhang. Darkrank: Accelerating deep metric learning via cross sample similarities transfer. In *AAAI*, 2018.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances, 2013.

Tim R Davidson, Luca Falorsi, Nicola De Cao, Thomas Kipf, and Jakub M Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

Adji B. Dieng, Chong Wang, Jianfeng Gao, and John Paisley. Topicrnn: A recurrent neural network with long-range semantic dependency, 2017.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

Hafsa Ennajari, Nizar Bouguila, and Jamal Bentahar. Combining knowledge graph and word embeddings for spherical topic modeling. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021. doi: 10.1109/TNNLS.2021.3112045.

Ian Gemp, Ramesh Nallapati, Ran Ding, Feng Nan, and Bing Xiang. Weakly semi-supervised neural topic models. 2019.

Kelvin Guu, Tatsunori B. Hashimoto, Yonatan Oren, and Percy Liang. Generating sentences by editing prototypes, 2018.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

Alexander Hoyle, Pranav Goel, and Philip Resnik. Improving neural topic models using knowledge distillation, 2020.

Minghui Huang, Yanghui Rao, Yuwei Liu, Haoran Xie, and Fu Lee Wang. Siamese network-based supervised topic modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4652–4662, 2018.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey, 2018.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. Semi-supervised learning with deep generative models. In *Advances in neural information processing systems*, pp. 3581–3589, 2014.

Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In *Proceedings of the 21st International Conference on Neural Information Processing Systems*, NIPS'08, pp. 897–904, Red Hook, NY, USA, 2008. Curran Associates Inc. ISBN 9781605609492.

Ken Lang. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, pp. 331–339. Elsevier, 1995.

Luyang Liu, Heyan Huang, Yang Gao, Yongfeng Zhang, and Xiaochi Wei. Neural variational correlated topic modeling. In *The World Wide Web Conference*, pp. 1142–1152, 2019.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised hierarchical text classification, 2018.

Yu Meng, Jiaxin Huang, Guangyuan Wang, Chao Zhang, Honglei Zhuang, Lance Kaplan, and Jiawei Han. Spherical text embedding, 2019.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference, 2018.

Jonghwan Mun, Kimin Lee, Jinwoo Shin, and Bohyung Han. Learning to specialize with knowledge distillation for visual question answering. *Advances in neural information processing systems*, 31, 2018.

Nikolaos Passalis, Maria Tzelepi, and Anastasios Tefas. Probabilistic knowledge transfer for lightweight deep representation learning. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5):2030–2039, 2020.

Peyman Passban, Yimeng Wu, Mehdi Rezagholizadeh, and Qun Liu. Alp-kd: Attention-based layer projection for knowledge distillation, 2020.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 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, 2009.

Reutterer Reisenbichler, M. Topic modeling in marketing: recent advances and research opportunities. *J Bus Econ*, 89, 2019. URL https://doi.org/10.1007/s11573-018-0915-7.

Joseph Reisinger, Austin Waters, Bryan Silverthorn, and Raymond J Mooney. Spherical topic models. In *ICML*, 2010.

M. Roberts, B. Stewart, D. Tingley, and E. Airoldi. The structural topic model and applied social science. *Neural Information Processing Society*, 2013.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. Fitnets: Hints for thin deep nets, 2015.

Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, Mar 2020. ISSN 0167-2789. doi: 10.1016/j.physd.2019.132306. URL http://dx.doi.org/10.1016/j.physd.2019.132306.

Leslie N. Smith and Nicholay Topin. Super-convergence: Very fast training of neural networks using large learning rates, 2018.

Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: Extraction and mining of academic social networks. In *KDD'08*, pp. 990–998, 2008.

Jie Tang, Limin Yao, Duo Zhang, and Jing Zhang. A combination approach to web user profiling. *ACM TKDD*, 5(1):1–44, 2010.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006.

Luis Caicedo Torres, Luiz Manella Pereira, and M. Hadi Amini. A survey on optimal transport for machine learning: Theory and applications, 2021.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation, 2019.

Xinyi Wang and Yi Yang. Neural topic model with attention for supervised learning. In Silvia Chiappa and Roberto Calandra (eds.), *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 1147–1156. PMLR, 26–28 Aug 2020a. URL https://proceedings.mlr.press/v108/wang20c.html.

Xinyi Wang and Yi Yang. Neural topic model with attention for supervised learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1147–1156. PMLR, 2020b.

Hongteng Xu, Wenlin Wang, Wei Liu, and Lawrence Carin. Distilled wasserstein learning for word embedding and topic modeling. *arXiv preprint arXiv:1809.04705*, 2018.

Jiacheng Xu and Greg Durrett. Spherical latent spaces for stable variational autoencoders, 2018.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification, 2016.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. Neural topic model via optimal transport, 2020a.

Jinjin Zhao, Kim Larson, Weijie Xu, Neelesh Gattani, and Candace Thille. Targeted feedback generation for constructed-response questions. 2020b.

Jun Zhu, Amr Ahmed, and Eric P. Xing. Medlda: Maximum margin supervised topic models. *Journal of Machine Learning Research*, 13(74):2237–2278, 2012. URL http://jmlr.org/papers/v13/zhu12a.html.

Table 1: Description of the notations used in this work.

Notion	Description
W	Corpus
M	cost matrix for topic modeling
s	similarity matrix between labeled documents and unlabeled documents
Z	topic proportions
t_d	topic distribution
X	bow of words representation for all documents
x_d	bag of words representation for a document
ϕ	encoder
$L_{Recon}(X)$	reconstruction loss
e_W	word embedding matrix
e_T	topic embedding matrix
μ	vmf direction parameter
κ	vmf concentration parameter
τ	temperature parameter for knowledge distillation
P	wight matrix for optimal transport
g	a labeled document group
G	all labeled document group
T	group of all topics
D	all documents
L_{OT}	optimal transport loss
λ	entropy penalty weights
L_{KL}	KL divergence
i	a labeled document from a group g
$s_i(x)$	the similarity of document x on labeled document i
L_{KD}	knowledge distillation loss
α, β	coefficients for L_{KD} and L_{OT}
$thresh$	threshold for knowledge distillation

Appendix

A Notations Tables

Table refsample-table summarizes the notation used in the paper.

B Time Tables

We document the training time on our machine for all these methods and averaged across 10 runs. We also capture finetune time of KDSTM. To be specific, finetune time is stage 2 and 3 of KDSTM. Before this stage, model does not require any labels to train.

	20News	AG's News	DBLP
TAM	98.98	842.64	378.05
KDSTM	20.75	252.87	117.36
KDSTM finetune	9.04	106.01	43.72
WestClass	104.30	888.61	87.17

Table 2: We capture the training time(seconds) of methods WestClass, KDSTM overall/finetune stage and TAM.

C von Mises-Fisher Distribution

In low dimensions, the Gaussian density presents a concentrated probability mass around the origin. This is problematic when the data is partitioned into multiple clusters. An ideal prior should be non informative and uniform over the parameter space. Thus, the von Mises-Fisher(vMF) is used

in VAE. vMF is a distribution on the (M-1)-dimensional sphere in R^M, parameterized by $\mu \in R^M$ where $||\mu|| = 1$ and a concentration parameter $\kappa \in R_{\geq 0}$. The probability density function of the vMF distribution for $z \in R^D$ is defined as:

$$q(Z|\mu, \kappa) = C_M(\kappa) exp(\kappa \mu^T Z)$$

$$C_M(\kappa) = \frac{\kappa^{\frac{M}{2}-1}}{(2\pi)^{\frac{M}{2}} I_{\frac{M}{2}-1}(\kappa)} + log2$$

where I_v denotes the modified Bessel function of the first kind at order v. The KL divergence with vMF(., 0) Davidson et al. (2018) is

$$KL(vMF(\mu, \kappa)|vMF(., 0)) = \kappa \frac{I_{\frac{M}{2}}(\kappa)}{I_{\frac{M}{2}-1}(\kappa)}$$

$$+(\frac{M}{2} - 1)log\kappa - \frac{M}{2}log(2\pi) - logI_{\frac{M}{2}-1}(\kappa)$$

$$+\frac{M}{2}log\pi + log2 + log\Gamma(\frac{M}{2})$$

vMF based VAE has better clusterability of data points especially in low dimensions Guu et al. (2018).

D Dataset Details

Corpus Name	Class name (Number of documents in the class)	Average Document Length
20News	Atheism (689), Religion (521), Graphics (836), Space (856)	55.6
AG's News	Politics (30000), Sports (30000), Business (30000), Technology (30000)	45.0
DBLP	Database (11981), Data Mining (4763), Artificial Intelligence (20890), Computer Vision (16961)	5.4

Table 3: Details of selected datasets

E Code for Benchmarks

WestClass Meng et al. (2018): This method is a weakly-supervised neural text classification method. The weak supervision source can come from any of the three sources: label surface names, class related keywords, and labeling documents. Output will be document labels consistent with cluster of inputs. We will use this method to benchmark the performance on document classification accuracy. The code we use is: `https://github.com/yumeng5/WeSTClass`

L-Label Ramage et al. (2009): This method accepts labeled documents and is used to compare KDSTM. To better benchmark the performance, we fine-tune alpha, eta and number of iterations to get the best accuracy performance. The code we use is: `https://github.com/JoeZJH/Labeled-LDA-Python`

TAM Wang & Yang (2020a): This method achieves the best performance compare to other supervised topic modeling methods. To better benchmark the performance, we fine-tune dimension of GRU gate, learning rate and num of epochs to get the best performance. Since this method is not stable, we show the median of 10 runs to get the baselines. The code we use is: `https://github.com/WANGXinyiLinda/Neural-Topic-Model-with-Attention-for-Supervised-Learning`

F Scale VS Performance

Figure 3 shows the variation of metrics wrt scale. As scale increase, micro F1 and accuracy increase while AUC decreases.

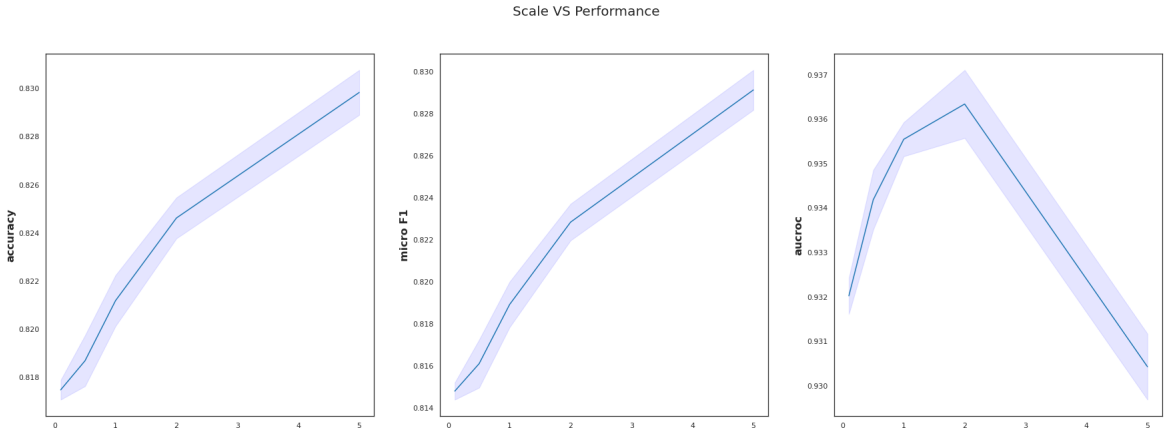

Figure 3: Performance vs Scale