Diversity-Aware Coherence Loss for Improving Neural Topic Models

Raymond Li[†], **Felipe González-Pizarro**[†], **Linzi Xing**[†], **Gabriel Murray**[‡], **Giuseppe Carenini**[†]
[†] University of British Columbia, Vancouver, BC, Canada
[‡] University of Fraser Valley, Abbotsford, BC, Canada
{raymondl, felipegp, lzxing, carenini}@cs.ubc.ca
gabriel.murray@ufv.ca

Abstract

The standard approach for neural topic modeling uses a variational autoencoder (VAE) framework that jointly minimizes the KL divergence between the estimated posterior and prior, in addition to the reconstruction loss. Since neural topic models are trained by recreating individual input documents, they do not explicitly capture the coherence between topic words on the corpus level. In this work, we propose a novel diversity-aware coherence loss that encourages the model to learn corpus-level coherence scores while maintaining a high diversity between topics. Experimental results on multiple datasets show that our method significantly improves the performance of neural topic models without requiring any pretraining or additional parameters.

1 Introduction

The main goal of topic modeling is to discover latent topics that best explain the observed documents in the corpus. The topics, conceptualized as a multidimensional distribution over the vocabulary, are useful for many downstream applications, including summarization (Wang et al., 2020; Xiao et al., 2022), text generation (Wang et al., 2019; Nevezhin et al., 2020), dialogue modeling (Xu et al., 2021; Zhu et al., 2021), as well as analyzing the data used for pretraining large language models (Chowdhery et al., 2022). When presented to humans, they are often represented as lists of the most probable words to assist the users in exploring and understanding the underlying themes in a large collection of documents. While the extrinsic quality of topics can be quantified by the performance of their downstream tasks, the intrinsic interpretability of topics appears to be strongly correlated with two important factors, namely *coherence* and *diversity* (Dieng et al., 2020).

The topic *coherence* measures to what extent the words within a topic are related to each other in a meaningful way. Although human studies provide a direct method for evaluation, they can be costly, especially when a large number of models are waiting to be assessed. Therefore, various automatic metrics have been developed to measure topic coherence (Newman et al., 2010; Mimno et al., 2011; Xing et al., 2019; Terragni et al., 2021). For instance, the well-established Normalized Pointwise Mutual Information (NPMI) metric (Lau et al., 2014), based on word co-occurrence within a fixed window, has been found to have a strong correlation with human judgment (Röder et al., 2015). On the other hand, topic *diversity* measures to what extent the topics are able to capture different aspects of the corpus based on the uniqueness of the topic words (Nan et al., 2019). Importantly, studies have shown that optimizing for coherence can come at the expense of diversity (Burkhardt and Kramer, 2019). Even without accounting for topic diversity, directly optimizing for topic coherence by itself is a non-trivial task, due to the computational overhead and non-differentiability of the score matrix (Ding et al., 2018).

While traditional topic modeling algorithms are in the form of statistical models such as the Latent Dirichlet Allocation (LDA) (Blei et al., 2003), advancements in variational inference methods (Kingma and Welling, 2014; Rezende et al., 2014) have led to the rapid development of neural topic model (NTM) architectures (Miao et al., 2016, 2017; Srivastava and Sutton, 2017). More recently, follow-up works have focused on the integration of additional knowledge to improve the coherence of NTMs. Their attempts include the incorporation of external embeddings (Ding et al., 2018; Card et al., 2018; Dieng et al., 2020; Bianchi et al., 2021a,b), knowledge distillation (Hoyle et al., 2020), and model pretraining (Zhang et al., 2022). However, as the model is designed to operate on a document-level input, one significant limitation of NTMs is their inability to explicitly capture the corpus-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 1710–1722

level coherence score, which assesses the extent to which words within specific topics tend to occur together in a comparable context within a given corpus. For example, semantically irrelevant words such as "*politics*" and "*sports*" might be contextually relevant in a given corpus (e.g., government funding for the national sports body). Recently, one closely related work addresses this gap by reinterpreting topic modeling as a coherence optimization task with diversity as a constraint (Lim and Lauw, 2022).

While traditional topic models tend to directly use corpus-level coherence signals, such as factorizing the document-term matrix (Steyvers and Griffiths, 2007), and topic segment labeling with random walks on co-occurrence graphs (Mihalcea and Radev, 2011; Joty et al., 2013), to the best of our knowledge, no existing work have explicitly integrated corpus-level coherence scores into the training of NTMs without sacrificing topic diversity. To address this gap, we propose a novel **coherence-aware diversity loss**, which is effective to improve both the coherence and diversity of NTMs by adding as an auxiliary loss during training. Experimental results show that this method can significantly improve baseline models without any pretraining or additional parameters[1].

2 Background

Latent Dirichlet Allocation (LDA) (Blei et al., 2003) is a simple yet effective probabilistic generative model trained on a collection of documents. It is based on the assumption that each document w in the corpus is described by a random mixture of latent topics z sampled from a distribution parameterized by θ, where the topics β are represented as a multidimensional distribution over the vocabulary V. The formal algorithm describing the generative process is presented in Appendix A. Under this assumption, the marginal likelihood of the document $p(w|\alpha, \beta)$ is described as:

$$\int_\theta \left(\prod_i^{|V|} \sum_{z_i}^{K} p(w_i|z_i, \beta) p(z_i|\theta) \right) p(\theta|\alpha) d\theta \quad (1)$$

However, since the posterior distribution $p(z_i|\theta)$ is intractable for exact inference, a wide variety of approximate inference algorithms have been used for LDA (e.g., Hoffman et al. (2010)).

A common strategy to approximate such posterior is employing the variational auto-encoder (VAE) (Kingma and Welling, 2014). In particular, NTMs use an encoder network to compress the document representation into a continuous latent distribution and pass it to a generative decoder to reconstruct the bag-of-words (BoW) representation of the documents. The model is trained to minimize the evidence lower bound (ELBO) of the marginal log-likelihood described by the LDA generative process:

$$\begin{aligned} L_{\text{ELBO}} = &- D_{\text{KL}}[q(\theta, z|w)||p(\theta, z|\alpha)] \\ &+ \mathbb{E}_{q(\theta,z|w)}[\log p(w|z, \theta, \alpha, \beta)] \end{aligned} \quad (2)$$

In Equation 2, the first term attempts to match the variational posterior over latent variables to the prior, and the second term ensures that the variational posterior favors values of the latent variables that are good at explaining the data (i.e., reconstruction loss). While standard Gaussian prior has typically been used in VAEs, **ProdLDA** (Srivastava and Sutton, 2017) showed that using a Laplace approximation of the Dirichlet prior achieved superior performance. To further improve topic coherence, **CombinedTM** (Bianchi et al., 2021a) concatenated the BoW input with contextualized SBERT embeddings (Reimers and Gurevych, 2019), while **ZeroshotTM** (Bianchi et al., 2021b) used only contextualized embeddings as input. These are the three baselines included in our experiments.

3 Proposed Methodology

Despite the recent advancements, one significant limitation of the NTM is that since the model is trained on document-level input, it does not have direct access to corpus-level coherence information (i.e., word co-occurrence). Specifically, the topic-word distribution β is optimized on the document-level reconstruction loss, which may not be an accurate estimate of the true corpus distribution due to the inherent stochasticity of gradient-descent algorithms. We address this problem by explicitly integrating a corpus-level coherence metric into the training process of NTMs using an auxiliary loss.

3.1 Optimizing Corpus Coherence

To improve the topic-word distribution β, we maximize the corpus-level coherence through the well-established NPMI metric[2] (Bouma, 2009; Lau et al.,

[1]The implementation of our work is available at: https://github.com/raymondzmc/Topic-Model-Diversity-Aware-Coherence-Loss

[2]Detailed definition of NPMI is presented in Appendix B.

2014). After computing the pairwise NPMI matrix $N \in \mathbb{R}^{|V|\times|V|}$ on the corpus, we use the negative β-weighted NPMI scores of the top-n words within each topic as the weight for the coherence penalty of β, where n is a hyperparameter that equals to the number of topic words to use. Specifically, we apply a mask M_c to keep the top-n words of each topic and apply the row-wise softmax operation σ to ensure the value of the penalty is always positive. We define the coherence weight W_C in Equation 3.

$$W_C = 1 - \text{normalize}(\sigma(\beta \odot M_c)N) \quad (3)$$

Intuitively, each value in $\sigma(\beta \odot M_k)N$ represents the β-weighted average NPMI score with other words in the topic. Then we use row-wise normalization to scale the values, so $W_C \in [0, 1]$.

3.2 Improving Topic Diversity

One problem with the coherence weight W_C is that it does not consider the diversity across topics. To account for this, we propose an additional method to simultaneously improve topic diversity by encouraging words unused by other topics to have higher probabilities. To achieve this, we bin the words within each topic into two groups, where the words in the first group consist of those that already have a high probability in other topics (i.e., appear within top-n words), while the second group does not. The intuition is that we want to penalize the words in the first group more than the words in the second group. In practice, we use a mask $M_d \in \mathbb{R}^{K\times V}$ for selecting β logits in the first group, where hyperparameter $\lambda_d \in [0.5, 1]$ is a balancing constant between the two groups and n is the number of topic words to use. We then compute the diversity-aware coherence weight W_D as the λ_d-weighted sum of W_C:

$$W_D = \lambda_d M_d \odot W_C + (1-\lambda_d)(\neg M_d) \odot W_C \quad (4)$$

From Equation 4, we see that when $\lambda_d = 0.5$, there are no constraints on diversity since the two groups are penalized equally ($2W_D = W_C$).

3.3 Auxiliary Loss

From the two definitions of coherence weight (W_C, W_D), we propose an auxiliary loss that can be directly combined with the ELBO loss (Equation 2) when training the NTM. Since β are unnormalized logits containing negative values, we apply the softmax operation $\sigma(\beta)$ to avoid unbound optimization.

$$L_{\text{AUX}} = \frac{1}{2}[\sigma(\beta)]^2 \odot W_D \quad (5)$$

In Equation 5, the topic probabilities are penalized by their negative weighted coherence score with the top-n words. The square operation ensures that words with very high probability are penalized to avoid the global minima, we justify this decision based on its partial derivatives in the next subsection.

The final objective function is the multitask loss consisting of the ELBO and our defined auxiliary loss:

$$L = L_{\text{ELBO}} + \lambda_a L_{\text{AUX}} \quad (6)$$

During training, we employ a linear warm-up schedule to increase λ_a gradually, so the model can learn to reconstruct the BoW representation based on the topic distribution α before optimizing for coherence and diversity.

3.4 Derivatives

We justify our auxiliary loss defined in Equation 5 using the derivatives w.r.t. the β parameters. For simplicity, we define $p_{k,i} = \sigma(\beta_k)_i$ as the softmax probability for word i in topic k. Since we detach the gradients when computing W, it can be treated as a constant w in the derivatives.

$$\begin{aligned}\frac{\partial L_{\text{AUX}}}{\partial \beta_{k,i}} = \; & w \cdot p_{k,i} \cdot p_{k,i}(1 - p_{k,i}) + \\ & w \cdot \sum_{j \neq i} p_{k,j}(-p_{k,j} p_{k,i})\end{aligned} \quad (7)$$

In Equation 7, the partial derivatives w.r.t. $\beta_{k,i}$ can be broken down into two terms. In the first term, the softmax derivative $p_{k,i}(1 - p_{k,i})$ is zero when $p_{k,i}$ is either 0 or 1 (really small or really large). The additional $p_{k,i}$ (from the square operation) penalizes over-confident logits and leads to better topics. Similarly for the second term, since $\sum_i p_{k,i} = 1$, $\sum_{j\neq i} \left(p_{k,j} p_{k,i}\right)$ is zero (global minima) when one logit dominates the others. Therefore, the additional $p_{k,j}$ has the same penalizing effect on the over-confident logits.

4 Experiments

In this section, we describe the experimental settings and present the quantitative results to assess the benefits of our proposed loss.

Dataset	20NewsGroup				Wiki20K				GoogleNews			
Metrics	NPMI	WE	I-RBO	TU	NPMI	WE	I-RBO	TU	NPMI	WE	I-RBO	TU
LDA	.0426	.1624	**.9880**	**.8077**	-.0470	.1329	.9934	**.8664**	-.2030	.0989	.9973	**.9065**
ProdLDA	.0730	.1626	.9923	.7739	.1712	.1883	**.9948**	.7674	.0919	.1240	.9974	.8460
CombinedTM	.0855	.1643	.9922	.7705	.1764	.1893	.9941	.7509	.1062	.1316	.9943	.7498
ZeroshotTM	.1008	.1749	.9910	.7214	.1783	.1896	.9916	.6999	.1218	.1321	.9967	.8200
ProdLDA + W_C	.1233	.1775	.9916	.7526	.2386	.2094	.9905	.6933	.1236	.1262	.9973	.8400
CombinedTM + W_C	.1301	.1781	.9910	.7477	.2392	.2113	.9890	.6748	.1378	.1339	.9938	.7421
ZeroshotTM + W_C	.1456	.1882	.9895	.6975	.2455	.2147	.9862	.6350	.1562	.1349	.9964	.8131
ProdLDA + W_D	.1235	.1786	.9940	.7901	.2367	.2101	.9929	.7556	.1275	.1274	**.9975**	.8504
CombinedTM + W_D	.1309	.1790	.9935	.7833	.2404	.2137	.9918	.7366	.1429	.1354	.9942	.7541
ZeroshotTM + W_D	**.1482**	**.1899**	.9919	.7343	**.2460**	**.2156**	.9890	.6904	**.1569**	**.1350**	.9967	.8228

Table 1: Average results over 5 number of topics ($K = 25, 50, 75, 100, 150$), where the results for each K are averaged over 10 random seeds. The results are reported for $\lambda_d = 0.7$, a mid-range value in the $[0.5, 1]$ interval.

4.1 Datasets and Evaluation Metrics

To test the generality of our approach, we train and evaluate our models on three publicly available datasets: 20NewsGroups, Wiki20K (Bianchi et al., 2021b), and GoogleNews (Qiang et al., 2022). We provide the statistics of the three datasets in Table 2[3].

Dataset	Domain	Docs	Vocabulary
20Newsgroups	Email	18,173	2,000
Wiki20K	Article	20,000	2,000
Google News	News	11,108	8,110

Table 2: Statistics of the three datasets used in our experiments.

We use automatic evaluation metrics to measure the topic coherence and diversity of the models. For coherence, we use the NPMI and Word Embedding (WE) (Fang et al., 2016) metrics, which measure the pairwise NPMI score and word embedding similarity, respectively, between the top-10 words of each topic. For diversity, we use Topic Uniqueness (TU) (Dieng et al., 2020), which measures the proportion of unique topic words, and Inversed Rank-Biased Overlap (I-RBO) (Terragni et al., 2021; Bianchi et al., 2021a), measuring the rank-aware difference between all combinations of topic pairs.

4.2 Baselines

We plug our proposed auxiliary loss to three baseline NTMs' training process to demonstrate the benefits of our approach across different settings. Specifically, the three models are (1) ProdLDA (Srivastava and Sutton, 2017), (2) CombinedTM (Bianchi et al., 2021a), and (3) ZeroshotTM (Bianchi et al., 2021b). For comparison, we also include the results of the standard LDA algorithm (Blei et al., 2003).

4.3 Hyperparemeter Settings

We follow the training settings reported by Bianchi et al. (2021a), with 100 epochs and a batch size of 100. The models are optimized using the ADAM optimizer (Kingma and Ba, 2015) with the momentum set to 0.99 and a fixed learning rate of 0.002. We do not modify the architecture of the models, where the inference network is composed of a single hidden layer and 100 dimensions of softplus activation units (Zheng et al., 2015). The priors over the topic and document distributions are learnable parameters. A 20% Dropout (Srivastava et al., 2014) is applied to the document representations. During our evaluation, we follow the same setup and used the top-10 words of each topic for the coherence and diversity metrics.

For the hyperparameters introduced in the diversity-aware coherence loss, both M_c and M_d are computed using the top-**20** words of each topic. The scaling factor λ_a is linearly increased for the first 50 epochs and kept constant for the last 50 epochs, we set λ_a to be **100** in order to balance the loss magnitude of L_{ELBO} and L_{AUX}. The λ_d in the diversity loss is set by taking a mid-range value of **0.7** in the $[0.5, 1]$ range. We do not perform any searches over our defined hyperparameters; we believe that additional experiments will yield better results (i.e., by using a validation set).

4.4 Results

Table 1 shows improvements across all settings. However, with the basic coherence loss (W_C), the

[3]Detailed description of the three datasets is provided in Appendix C.

significant coherence increase comes at the expense of topic diversity, where a slight decrease can be observed in the I-RBO and TU scores. In contrast, with the diversity-aware coherence loss (W_D), we observe that the model improves in coherence while having a significantly higher diversity over the basic loss (W_C). The further coherence improvements can be attributed to the regularization effects, where words with a high probability of belonging to another topic are less likely to be related to words in the current topic. Lastly, it is worth noting that due to the gradual increase in λ_a, our proposed loss has a negligible effect on the original document-topic distribution θ, and only modifies the word distribution within the established topics. We provide some sample model outputs in Appendix D.

4.5 Coherence and Diversity Trade-off

To study the effects of λ_d on the trade-off between coherence and diversity, we perform experiments with different values of λ_d with the ZeroshotTM baseline, which has the best overall performance. Note that when $\lambda_d = 0.5$, the objective is equivalent to the basic coherence loss. From results on the 20NewsGroups Dataset (Table 3), we see that coherence peaks at $\lambda_d = 0.7$ before the diversity penalty begins to dominate the loss. Further, while a higher value of λ_d leads to a lower coherency score, both coherency and diversity are still improved over the baselines for all values of λ_d, demonstrating the effectiveness of our method without the need for extensive hyperparameter tuning. We observe an identical trend in other datasets.

	NPMI	WE	I-RBO	TU
ZeroshotTM	.1008	.1749	.9910	.7214
$\lambda_d = 0.5$	.1456	.1882	.9895	.6975
$\lambda_d = 0.6$	.1428	.1875	.9908	.7198
$\lambda_d = 0.7$	**.1482**	**.1899**	.9919	.7343
$\lambda_d = 0.8$	.1443	.1890	.9925	.7499
$\lambda_d = 0.9$	.1369	.1867	.9933	.7724
$\lambda_d = 1.0$	.1193	.1816	**.9951**	**.8086**

Table 3: Results on the 20NewsGroups dataset for different values of λ_d with ZeroshotTM.

4.6 Comparison with Composite Activation

The recent work by Lim and Lauw (2022) proposed a model-free technique to refine topics based on the parameters of the trained model. Specifically, they solve an optimization problem (with the NPMI score as the objective) using a pool of candidates while setting the diversity score as a constraint.

Since their goal is similar to ours, we run further evaluations to compare the respective approaches. In particular, we experiment with ZeroshotTM on the 20NewsGroups dataset for $K = 25, 50$. For comparison, we use their Multi-Dimensional Knapsack Problem (MDKP) formulation, since it achieved the best overall performance. Regrettably, considering larger topic numbers was not possible due to the NP-hard runtime complexity of MDKP. From the results in Table 4, we see that while our methods have similar coherence scores, MDKP archives higher topic diversity due to its selectivity of less-redundant topics. However, when combining MDKP with our proposed loss (+ W_D + MDKP), we achieve the highest overall performance across all metrics. This is expected since the pool of potential topic candidates is generated based on the trained model, and better-performing models lead to superior candidates.

$K = 25$	NPMI	WE	I-RBO	TU
ZeroshotTM	.1059	.1791	.9927	.9152
+ MDKP	.1481	.1895	**.9991**	.9804
+ W_D	.1433	.1921	.9981	.9688
+ W_D + MDKP	**.1657**	**.2043**	.9989	**.9808**
$K = 50$	NPMI	WE	I-RBO	TU
ZeroshotTM	.1109	.1746	.9937	.8498
+ MDKP	.1578	.1903	.9983	.9452
+ W_D	.1581	.1921	.9963	.8840
+ W_D + MDKP	**.1783**	**.1932**	**.9985**	**.9500**

Table 4: Results for comparing our approach with Composite Activation on the 20NewsGroups dataset.

5 Conclusion and Future Work

In this work, we present a novel diversity-aware coherence loss to simultaneously improve the coherence and diversity of neural topic models. In contrast to previous methods, our approach directly integrates corpus-level coherence scores into the training of Neural Topic Models. The extensive experiments show that our proposal significantly improves the performance across all settings without requiring any pretraining or additional parameters.

For future work, we plan to perform extensive user studies to examine the extent to which improvements in quantitative metrics affect human preference. Further, we would like to extend our approach to other quantitative metrics (e.g., semantic similarity), and perform extrinsic evaluation to study the effects of our approach when the topics are used for downstream tasks (e.g., summarization, dialogue modeling, text generation).

Limitations

We address several limitations with regard to our work. First, the publicly available datasets used in our experiments are limited to English. Documents in different languages (i.e., Chinese) might require different segmentation techniques and may contain unique characteristics in terms of vocabulary size, data sparsity, and ambiguity. Secondly, we only evaluate the quality of the topic models in terms of coherence and diversity. Future work should explore how our method impacts other characteristics, such as document coverage (i.e., how well documents match their assigned topics) and topic model comprehensiveness (i.e., how thoroughly the model covers the topics appearing in the corpus).

Ethics Statement

The datasets used in this work are publicly available and selected from recent literature. There could exist biased views in their content, and should be viewed with discretion.

Our proposed method can be applied to extract topics from a large collection of documents. Researchers wishing to apply our method should ensure that the input corpora are adequately collected and do not violate any copyright infringements.

References

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

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

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

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

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040, Melbourne, Australia. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

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

Ran Ding, Ramesh Nallapati, and Bing Xiang. 2018. Coherence-aware neural topic modeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 830–836, Brussels, Belgium. Association for Computational Linguistics.

Anjie Fang, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. Using word embedding to evaluate the coherence of topics from twitter data. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, page 1057–1060, New York, NY, USA. Association for Computing Machinery.

Matthew Hoffman, Francis Bach, and David Blei. 2010. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Alexander Miserlis Hoyle, Pranav Goel, and Philip Resnik. 2020. Improving Neural Topic Models using Knowledge Distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1752–1771, Online. Association for Computational Linguistics.

Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. 2013. Topic segmentation and labeling in asynchronous conversations. *Journal of Artificial Intelligence Research*, 47:521–573.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539, Gothenburg, Sweden. Association for Computational Linguistics.

Jia Peng Lim and Hady Lauw. 2022. Towards reinterpreting neural topic models via composite activations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3688–3703, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 2410–2419. PMLR.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1727–1736, New York, New York, USA. PMLR.

Rada Mihalcea and Dragomir Radev. 2011. *Graph-based Natural Language Processing and Information Retrieval*. Cambridge University Press.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6345–6381, Florence, Italy. Association for Computational Linguistics.

Egor Nevezhin, Nikolay Butakov, Maria Khodorchenko, Maxim Petrov, and Denis Nasonov. 2020. Topic-driven ensemble for online advertising generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2273–2283, Barcelona, Spain (Online). International Committee on Computational Linguistics.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, Los Angeles, California. Association for Computational Linguistics.

Jipeng Qiang, Zhenyu Qian, Yun Li, Yunhao Yuan, and Xindong Wu. 2022. Short text topic modeling techniques, applications, and performance: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):1427–1445.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1278–1286, Bejing, China. PMLR.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. In *Handbook of Latent Semantic Analysis*, pages 439–460. Psychology Press.

Silvia Terragni, Elisabetta Fersini, and Enza Messina. 2021. Word embedding-based topic similarity measures. In *Natural Language Processing and Information Systems: 26th International Conference on Applications of Natural Language to Information Systems, NLDB 2021, Saarbrücken, Germany, June 23–25, 2021, Proceedings*, pages 33–45. Springer.

Wenlin Wang, Zhe Gan, Hongteng Xu, Ruiyi Zhang, Guoyin Wang, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2019. Topic-guided variational auto-encoder for text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 166–177, Minneapolis, Minnesota. Association for Computational Linguistics.

Zhengjue Wang, Zhibin Duan, Hao Zhang, Chaojie Wang, Long Tian, Bo Chen, and Mingyuan Zhou. 2020. Friendly topic assistant for transformer based abstractive summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 485–497, Online. Association for Computational Linguistics.

Wen Xiao, Lesly Miculicich, Yang Liu, Pengcheng He, and Giuseppe Carenini. 2022. Attend to the right context: A plug-and-play module for content-controllable summarization. *arXiv preprint arXiv:2212.10819*.

Linzi Xing, Michael J. Paul, and Giuseppe Carenini. 2019. Evaluating topic quality with posterior variability. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3471–3477, Hong Kong, China. Association for Computational Linguistics.

Yi Xu, Hai Zhao, and Zhuosheng Zhang. 2021. Topic-aware multi-turn dialogue modeling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14176–14184.

Linhai Zhang, Xuemeng Hu, Boyu Wang, Deyu Zhou, Qian-Wen Zhang, and Yunbo Cao. 2022. Pre-training and fine-tuning neural topic model: A simple yet effective approach to incorporating external knowledge. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5989, Dublin, Ireland. Association for Computational Linguistics.

Hao Zheng, Zhanlei Yang, Wenju Liu, Jizhong Liang, and Yanpeng Li. 2015. Improving deep neural networks using softplus units. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–4. IEEE.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582, Online. Association for Computational Linguistics.

A LDA Generative Process

The formal generative process of a corpus under the LDA assumption can be described by the following algorithm.

Algorithm 1 Generative process of LDA

for each document w **do**
  Sample topic distribution $\theta \sim \text{Dirichlet}(\alpha)$
  for each word w_i **do**
    Sample topic $z_i \sim \text{Multinomial}(\theta)$
    Sample word $w_i \sim \text{Multinomial}(\beta_{z_i})$
  end for
end for

B Normalized Pointwise Mutual Information

Normalized Pointwise Mutual Information (NPMI) (Lau et al., 2014) measures how much more likely the most representative terms of a topic co-occur than if they were independent. The method for computing the NPMI score between word w_i and w_j is described in Equation 8, where $P(w_i, w_j)$ is computed using a window size of 10. This metric ranges from -1 to 1.

$$\text{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)} \quad (8)$$

In practice, the pairwise NPMI matrix is computed by first counting the word co-occurrence of all words in the corpus and then calculating the pairwise score following Equation 8. In summary, the NPMI matrix can be computed in $\mathcal{O}(|W| + |V|^2)$ for a corpus of $|W|$ words and vocab size $|V|$. Since the matrix is computed only once for each corpus prior to training, it does not increase the runtime complexity of training time.

C Datasets

This section provides details regarding the datasets we used. The 20NewsGroup[4] dataset is a collection of email documents partitioned evenly across 20 categories (e.g., electronics, space), we use the same filtered subset provided by Bianchi et al. (2021a). The Wiki20K dataset[5] contains randomly sampled subsets from the English Wikipedia abstracts from DBpedia[6]. GoogleNews[7] (Qiang et al., 2022) is downloaded from the Google news site by crawling the titles and snippets. We do not perform any additional pre-processing and directly use the data provided by the sources to create contextualized and BoW representation.

D Sample Output

Table 5 provides a qualitative comparison of the topics generated by our proposed method using ZeroshotTM on the 20NewsGroups dataset.

E Implementation Details

We base our implementation using the code provided by the authors of ZeroshotTM and CombinedTM (Bianchi et al., 2021a,b). Their repository[8] also provides the evaluation metrics used in our experiments. Our Python code base includes external open-source libraries including NumPy[9], SciPy[10], PyTorch[11], SentenceTransformers[12], Pandas[13], Gensim[14] and scikit-learn[15].

F Computing Details

All our experiments are run on Linux machines with single 1080Ti GPU (CUDA version 11.4). Each epoch with 100 batch size on the most computationally intensive setting (GoogleNews with $K = 150$) takes on average 3 seconds to run for the baselines models and 8 and 15 seconds, for W_C and W_D, respectively. Under this setting, a maximum VRAM usage of 800MB was recorded.

[4] http://qwone.com/~jason/20Newsgroups
[5] https://github.com/vinid/data
[6] https://wiki.dbpedia.org/downloads-2016-10
[7] https://github.com/qiang2100/STTM/tree/master/dataset
[8] https://github.com/MilaNLProc/contextualized-topic-models
[9] https://numpy.org/
[10] https://scipy.org/
[11] https://pytorch.org/
[12] https://www.sbert.net/
[13] https://pandas.pydata.org/
[14] https://radimrehurek.com/gensim/
[15] https://scikit-learn.org/stable/

Table 5: Sample model output $K = 25$ by running ZeroshotTM (Z) with our proposed method ($+W_C$ and $+W_D$) on the 20NewsGroups dataset. We visualize the top-10 keywords of each topic with unique keywords in **bold**.

Model	Top-10 Topic Keywords
Z	newsletter, aids, hiv, medical, cancer, disease, page, health, volume, patients
Z + W_C	newsletter, aids, hiv, medical, cancer, disease, page, health, volume, patients
Z + W_D	newsletter, hiv, aids, medical, cancer, disease, health, page, volume, patients
Z	mary, sin, god, heaven, lord, christ, jesus, grace, spirit, **matthew**
Z + W_C	mary, sin, heaven, god, christ, lord, jesus, spirit, grace, **matthew**
Z + W_D	mary, heaven, sin, christ, god, spirit, lord, jesus, **holy**, grace
Z	engine, car, bike, cars, oil, ride, road, dealer, **miles**, riding
Z + W_C	engine, bike, car, cars, oil, ride, dealer, road, riding, **driving**
Z + W_D	engine, bike, car, cars, oil, ride, dealer, riding, road, **driving**
Z	game, baseball, ball, season, fans, team, year, playing, players, **winning**
Z + W_C	game, baseball, fans, ball, season, team, playing, **teams**, players, year
Z + W_D	baseball, game, fans, season, **teams**, ball, team, playing, players, year
Z	fbi, koresh, batf, trial, compound, gas, investigation, **media**, branch, agents
Z + W_C	fbi, batf, koresh, compound, gas, agents, trial, branch, investigation, **waco**
Z + W_D	fbi, koresh, batf, compound, gas, agents, trial, branch, **waco**, investigation
Z	entry, rules, entries, email, build, info, file, char, program, section
Z + W_C	entry, rules, entries, email, info, build, file, char, section, program
Z + W_D	entry, rules, entries, email, build, info, file, char, program, section
Z	army, turkey, muslim, jews, greek, jewish, genocide, **professor**, ottoman, greece
Z + W_C	army, muslim, turkey, ottoman, jews, greek, genocide, jewish, greece, **muslims**
Z + W_D	muslim, turkey, ottoman, genocide, army, jews, greek, jewish, greece, **muslims**
Z	board, driver, video, cards, card, monitor, windows, drivers, screen, **resolution**
Z + W_C	board, video, driver, cards, monitor, card, windows, drivers, screen, **printer**
Z + W_D	video, board, driver, cards, monitor, card, drivers, **printer**, screen, windows
Z	frequently, previously, suggested, **announced**, **foundation**, **spent**, **contain**, **grant**, **consistent**, authors
Z + W_C	**basically**, previously, frequently, **generally**, suggested, **primary**, authors, **appropriate**, **kinds**, **greater**
Z + W_D	**essentially**, **basically**, **kinds**, **consistent**, frequently, authors, previously, **primary**, **equivalent**, suggested
Z	sale, condition, offer, asking, offers, shipping, items, price, **email**, sell
Z + W_C	sale, condition, offer, shipping, asking, items, offers, sell, **email**, price
Z + W_D	sale, condition, shipping, offer, asking, items, offers, sell, price, **excellent**
Z	application, window, xterm, motif, font, manager, widget, **root**, event, server
Z + W_C	xterm, application, window, motif, font, widget, manager, **x11r5**, server, event
Z + W_D	xterm, motif, font, application, window, widget, manager, **x11r5**, event, server
Z	gun, amendment, constitution, firearms, right, militia, guns, weapon, bear, weapons
Z + W_C	amendment, constitution, firearms, gun, militia, right, guns, weapon, bear, weapons
Z + W_D	amendment, firearms, constitution, gun, militia, guns, right, weapon, bear, weapons
Z	**suggested**, **frequently**, previously, **authors**, **foundation**, **consistent**, **spent**, **join**, **et**, **announced**
Z + W_C	**suggested**, previously, **frequently**, **greater**, **requirements**, **consistent**, **opportunity**, **authors**, **particularly**, **appropriate**
Z + W_D	**spent**, **greater**, **association**, **appropriate**, **opportunity**, **requirements**, **posts**, previously, **success**, **training**
Z	objective, atheist, atheism, morality, exists, belief, does, exist, atheists, existence
Z + W_C	objective, atheist, atheism, morality, exists, belief, atheists, does, exist, existence
Z + W_D	atheist, objective, atheism, belief, morality, exists, atheists, existence, exist, does
Z	think, president, people, Stephanopoulos, dont, jobs, just, know, mr, myers
Z + W_C	think, president, Stephanopoulos, people, dont, jobs, just, know mr, myers
Z + W_D	think, president, Stephanopoulos, people, dont, jobs, just, know, mr, myers
Z	board, drive, ide, scsi, bus, isa, mhz, motherboard, **internal**, **pin**
Z + W_C	board, drive, ide, scsi, motherboard, bus, isa, mhz, **hd**, **controller**
Z + W_D	board, drive, ide, motherboard, scsi, mhz, bus, **hd**, isa, **controller**
Z	jpeg, images, image, formats, gif, format, software, conversion, quality, color
Z + W_C	jpeg, images, formats, image, gif, format, conversion, software, quality, color
Z + W_D	jpeg, images, formats, gif, image, format, conversion, software, quality, color
Z	msg, food, doctor, vitamin, doctors, medicine, diet, **insurance**, treatment, studies
Z + W_C	msg, food, doctor, medicine, doctors, vitamin, diet, studies, treatment, **insurance**
Z + W_D	msg, food, doctor, medicine, doctors, vitamin, diet, studies, **patients**, treatment
Z	agencies, encryption, keys, secure, algorithm, **chip**, enforcement, nsa, **clipper**, **secret**
Z + W_C	agencies, encryption, secure, keys, algorithm, nsa, enforcement, **encrypted**, **escrow**, **chip**
Z + W_D	secure, encryption, keys, agencies, algorithm, **escrow**, **encrypted**, enforcement, nsa, **clipper**
Z	windows, dos, nt, network, card, disk, pc, software, modem, operating
Z + W_C	windows, dos, nt, card, network, disk, pc, modem, software, operating
Z + W_D	windows, dos, nt, card, network, disk, pc, modem, software, operating
Z	address, site, thanks, looking, newsgroup, appreciate, advance, mailing, **obtain**, **domain**
Z + W_C	address, thanks, newsgroup, site, appreciate, advance, looking, mailing, **thank**, **reply**
Z + W_D	address, appreciate, site, thanks, advance, newsgroup, looking, mailing, **thank**, **obtain**

Model	Top-10 Topic Keywords
Z	launch, nasa, shuttle, mission, satellite, energy, mass, moon, orbit, lunar
Z + W_C	launch, shuttle, nasa, mission, moon, satellite, orbit, energy, mass, lunar
Z + W_D	shuttle, launch, nasa, mission, orbit, moon, satellite, lunar, mass, energy
Z	floor, door, said, people, azerbaijani, neighbors, apartment, like, saw, **dont**
Z + W_C	floor, azerbaijani, door, said, people, apartment, neighbors, like, saw, **dont**
Z + W_D	azerbaijani, floor, apartment, door, said, people, neighbors, saw, like, **building**
Z	join, **grant**, **foundation**, **suggested**, **previously**, discussions, **frequently**, **authors**, **positions**, **announced**
Z + W_C	discussions, **topic**, **suggested**, join, **mailing**, **responses**, **robert**, **lists**, **summary**, **received**
Z + W_D	join, discussions, **foundation**, **robert**, **mailing**, **lists**, **topic**, **grant**, **received**, **responses**
Z	pts, boston, van, pittsburgh, pp, san, **vancouver**, chicago, **la**, **st**
Z + W_C	pts, boston, van, pittsburgh, pp, san, **vancouver**, chicago, **buf**, **tor**
Z + W_D	pts, pittsburgh, van, boston, pp, chicago, **buf**, **tor**, san, **det**

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Limitations

☑ A2. Did you discuss any potential risks of your work?
Ethics

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

4, Appendix C, E

☑ B1. Did you cite the creators of artifacts you used?
Section 4

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
In our paper, datasets and software tools used to reproduce our results are open-sourced and available to all developers. All our datasets are available for the general public, the license and terms for use are available using the links provide in Appendix C. For all open-sourced software packages, we include a detailed list of the websites for Python libraries and the baseline code base in Appendix E, the license and terms for use can be easily found on the websites. Since our models are stored in our private repository, and the license will be disclosed when the code base becomes public.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Not applicable. Left blank.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Appendix C, Limitations

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

C ☑ Did you run computational experiments?

4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix E

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
4, and Appendix D

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
4

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix E

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.