Neural Topic Modeling by Incorporating Document Relationship Graph

Deyu Zhou[*]   **Xuemeng Hu**   **Rui Wang**

School of Computer Science and Engineering, Key Laboratory of Computer Network
and Information Integration, Ministry of Education, Southeast University, China
{d.zhou,xuemenghu,rui_wang}@seu.edu.cn

Abstract

Graph Neural Networks (GNNs) that capture the relationships between graph nodes via message passing have been a hot research direction in the natural language processing community. In this paper, we propose Graph Topic Model (GTM), a GNN based neural topic model that represents a corpus as a document relationship graph. Documents and words in the corpus become nodes in the graph and are connected based on document-word co-occurrences. By introducing the graph structure, the relationships between documents are established through their shared words and thus the topical representation of a document is enriched by aggregating information from its neighboring nodes using graph convolution. Extensive experiments on three datasets were conducted and the results demonstrate the effectiveness of the proposed approach.

1   Introduction

Probabilistic topic models (Blei, 2012) are tools for discovering main themes from large corpora. The popular Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and its variants (Lin and He, 2009; Zhao et al., 2010; Zhou et al., 2014) are effective in extracting coherent topics in an interpretable manner, but usually at the cost of designing sophisticated and model-specific learning algorithm. Recently, neural topic modeling that utilizes neural-network-based black-box inference has been the main research direction in this field. Notably, NVDM (Miao et al., 2016) employs variational autoencoder (VAE) (Kingma and Welling, 2013) to model topic inference and document generation. Specifically, NVDM consists of an encoder inferring topics from documents and a decoder generating documents from topics, where the latent topics are constrained by a Gaussian prior. Srivastava and Sutton (2017) argued that Dirichlet distribution is a more appropriate prior for topic modeling than Gaussian in NVDM and proposed ProdLDA that approximates the Dirichlet prior with logistic normal. There are also attempts that directly enforced a Dirichlet prior on the document topics. W-LDA (Nan et al., 2019) models topics in the Wasserstein autoencoders (Tolstikhin et al., 2017) framework and achieves distribution matching by minimizing their Maximum Mean Discrepancy (MMD) (Gretton et al., 2012), while adversarial topic model (Wang et al., 2019a,b, 2020) directly generates documents from the Dirichlet prior and such a process is adversarially trained with a discriminator under the framework of Generative Adversarial Network (GAN) (Goodfellow et al., 2014).

Recently, due to the effectiveness of Graph Neural Networks (GNNs) (Li et al., 2015; Kipf and Welling, 2016; Zhou et al., 2018) in embedding graph structures, there is a surge of interests of applying GNN to natural language processing tasks (Yasunaga et al., 2017; Song et al., 2018; Yao et al., 2019). For example, GraphBTM (Zhu et al., 2018) is a neural topic model that incorporates the graph representation of a document to capture biterm co-occurrences in the document. To construct the graph, a sliding window over the document is employed and all word pairs in the window are connected.

A limitation of GraphBTM is that only word relationships are considered while ignoring document relationships. Since a topic is possessed by a subset of documents in the corpus, we believe that the topical neighborhood of a document, i.e., documents with similar topics, would help determine the topics of a document. To this end, we propose Graph Topic Model (GTM), a neural topic model that a corpus is represented as a document relationship graph where documents and words in the corpus are nodes and they are connected based on

---

[*]Corresponding author.

document-word co-occurrences. In GTM, the topical representation of a document node is aggregated from its multi-hop neighborhood, including both document and word nodes, using Graph Convolutional Network (GCN) (Kipf and Welling, 2016). As GCN is able to capture high-order neighborhood relationships, GTM is essentially capable of modeling both word-word and doc-doc relationships. In specific, the relationships between relevant documents are established by their shared words, which is desirable for topic modeling as documents belonging to one topic typically have similar word distributions.

The main contributions of the paper are:

- We propose GTM, a novel topic model that incorporates document relationship graph to enrich document and word representations.

- We extensively experimented on three datasets and the results demonstrate the effectiveness of the proposed approach.

2 Graph Topic Model

2.1 Graph Representation of the Corpus

We represent the whole corpus $\mathcal{D}$ with an undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N}$ and $\mathcal{E}$ are nodes and edges in the graph respectively. To model both words and documents, each of them is represented as a node $n_i \in \mathcal{N}$, which gives rise to $N = V + D$ nodes in total, where V is the size of vocabulary $\mathcal{V}$ and D is the number of documents in corpus $\mathcal{D}$. An edge (n_i, n_j) indicates the relevance of node n_i and n_j, whose weight is determined by

$$A_{i,j} = \begin{cases} \text{TF-IDF}_{ij}, & i \in \mathcal{D} \text{ and } j \in \mathcal{V} \\ \text{TF-IDF}_{ji}, & i \in \mathcal{V} \text{ and } j \in \mathcal{D} \\ 1, & i = j \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

where $\boldsymbol{A}$ is the adjacency matrix of $\mathcal{G}$ and TF-IDF$_{ij}$ denotes the max-normalized TF-IDF (Term FrequencyInverse Document Frequency) weight of word j in document i. Besides self-connections, we only apply positive weights to edges between documents and words, while rely on the model to capture higher-order relationships, e.g. doc-doc and word-word relationships, by applying graph convolutions on graph $\mathcal{G}$.
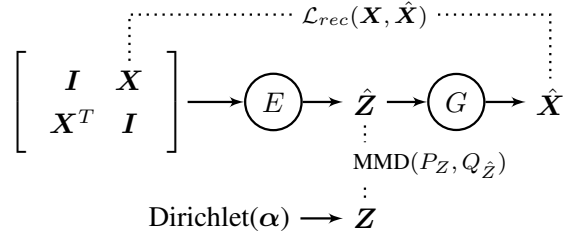

Figure 1: The framework of GTM. Circles denote neural networks. $\boldsymbol{X}$, $\boldsymbol{I}$, $\hat{\boldsymbol{Z}}$, $\hat{\boldsymbol{X}}$, $\boldsymbol{Z}$ are the TF-IDF matrix of the corpus, an identity matrix, latent topics of all documents, reconstructed word weights and topic distributions drawn from the Dirichlet prior respectively. $\mathcal{L}_{rec}(\boldsymbol{X}, \hat{\boldsymbol{X}})$ and MMD(P_Z, Q_Z) are training objectives.

2.2 Model Architecture

The proposed GTM consists of an encoder E and a decoder G. The framework is shown in Figure 1, and we detail the architecture in the following.

The encoder network E maps nodes in $\mathcal{G}$ to their topic distributions by iteratively applying graph convolution to the node features. Following (Kipf and Welling, 2016), the layer-wise propagation rule of the graph convolution at layer $l + 1 \in [1, L]$ is defined as

$$\boldsymbol{H}^{(l+1)} = \sigma(\boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{H}^{(l)} \boldsymbol{W}^{(l)}) \quad (2)$$

where $\boldsymbol{A} \in \mathbb{R}^{N \times N}$ is the adjacency matrix of $\mathcal{G}$, $D_{ii} = \sum_j A_{ij}$, $\boldsymbol{W}^{(l)} \in \mathbb{R}^{d^{(l)} \times d^{(l+1)}}$ is a layer-specific weight matrix where $d^{(l)}$ is the output size of layer l, and σ denotes an activation function that is LeakyReLU (Maas et al., 2013) in this paper. $\boldsymbol{H}^{(l)} \in \mathbb{R}^{N \times d^{(l)}}$ is the activations of all nodes at layer l and $\boldsymbol{H}_i^{(0)}$ is the embedding of node i.

At each encoder layer, what the graph convolution does is aggregating node features from a node's first-order neighborhood, which consequently enlarges the receptive field of the central node and enables the information propagation between relevant nodes. After successively applying L graph convolution layers, the encoding of a node essentially involves its Lth-order neighborhood. With $L \geq 2$, doc-doc and word-word relationships are naturally captured in the topic inference process.

We also add a batch normalization (Ioffe and Szegedy, 2015) after each graph convolution. After the graph encoding, a softmax is further applied to the node features of a document to produce a multinomial topic distribution $\hat{\boldsymbol{z}} \in \mathbb{R}^K$, where K is the topic number.

Based on the inferred topic distribution $\hat{z}$, the decoder network G tries to restore the original document representations. To achieve this goal, we employ a 2-layer MLP with LeakyReLU activation and batch normalization in the first layer. The output of the MLP decoder is then softmax-normalized to generate a word distribution $\hat{x} \in \mathbb{R}^V$.

The decoder is also used to interpret topics. In this case, we feed to the decoder an identity matrix $\boldsymbol{I} \in \mathbb{R}^{K \times K}$, and the decoder output $G(\boldsymbol{I})_i$ is the word distribution of the i-th topic.

2.3 Training Objective

Based on the Wasserstein Autoencoder (Tolstikhin et al., 2017) framework, the training objective of GTM is to minimize the document reconstruction loss when the latent topic space is constrained by a prior distribution. The reconstruction loss is defined as

$$\mathcal{L}_{rec}(\boldsymbol{X}, \hat{\boldsymbol{X}}) = -\mathbb{E}(\boldsymbol{x} \log \hat{\boldsymbol{x}}), \qquad (3)$$

where $\boldsymbol{x}$ denotes the TF-IDF of a document and $\hat{\boldsymbol{x}}$ is the reconstructed word distribution corresponding to $\boldsymbol{x}$. we use TF-IDF as the reconstruction target since TF-IDF basically preserves the relative importance of words and reduces some background noise that may hurt topic modeling, e.g., stop words.

We impose a Dirichlet prior, the conjugate prior of the multinomial distribution, to the latent topic distributions. Following W-LDA (Nan et al., 2019), we achieve this goal by minimizing the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012) between the distribution $Q_{\hat{Z}}$ of inferred topic distributions $\hat{z}$ and the Dirichlet prior P_Z from which we draw multinomial noises z:

$$\text{MMD}(P_Z, Q_{\hat{Z}}) = \frac{1}{m(m-1)} \sum_{i \neq j} k(\boldsymbol{z}^{(i)}, \boldsymbol{z}^{(j)}) +$$

$$\frac{1}{n(n-1)} \sum_{i \neq j} k(\hat{\boldsymbol{z}}^{(i)}, \hat{\boldsymbol{z}}^{(j)}) - \frac{2}{mn} \sum_{i,j} k(\boldsymbol{z}^{(i)}, \hat{\boldsymbol{z}}^{(j)}), \quad (4)$$

where m and n are the number of samples from Z and $\hat{Z}$ respectively (m and n are batch sizes and they are equal in our experiments), and $k \colon Z \times Z \to \mathbb{R}$ is the kernel function. We use the information diffusion kernel (Lebanon and Lafferty, 2003) as in W-LDA:

$$k(\boldsymbol{z}, \boldsymbol{z}') = \exp(-\arccos^2(\sum_{i=1}^{K} \sqrt{z_i z_i'})), \quad (5)$$

which is sensitive to points near the simplex boundary and thus more suitable for the sparse topic distributions.

3 Experiments

We evaluate our model on three datasets: 20Newsgroups consisting of 11,259 documents, Grolier consisting of 29,762 documents, and NYTimes consisting of 99,992 documents. We use the preprocessed 20Newsgroups of (Srivastava and Sutton, 2017), and preprocessed Grolier and NYTimes of (Wang et al., 2019a). We compare the performance of our model with LDA (Blei et al., 2003), NVDM (Miao et al., 2016), ProdLDA (Srivastava and Sutton, 2017), GraphBTM (Zhu et al., 2018), ATM (Wang et al., 2019a) and W-LDA (Nan et al., 2019) using topic coherence measures (Röder et al., 2015). To quantify the understandability of the extracted topics, a topic coherence measure aggregates the relatedness scores of the topic words (top-weighted words) of each topic, where the word relatedness scores are estimated based on word co-occurrence statistics on a large external corpus. For example, the NPMI coherence measure (Aletras and Stevenson, 2013) applies a sliding window of size 10 over the Wikipedia corpus to calculate NPMI (Bouma, 2009) for word pairs. We use three topic coherence measures in our experiments: C_A (Aletras and Stevenson, 2013), C_P (Röder et al., 2015), and NPMI. The topic coherence scores are calculated using Palmetto (Röder et al., 2015) [1].

Dataset	Model	C_A	C_P	NPMI
20Newsgroups	LDA	0.1769	0.2362	0.0524
	NVDM	0.1432	−0.2558	−0.0984
	ProdLDA	0.2155	0.1859	−0.0083
	GraphBTM	0.2195	0.2152	0.0082
	ATM	0.1720	0.1914	0.0207
	W-LDA	0.2065	0.2501	0.0400
	GTM	**0.2465**	**0.3451**	**0.0629**
Grolier	LDA	0.2009	0.1908	0.0498
	NVDM	0.1457	−0.1877	−0.0619
	ProdLDA	0.1734	−0.0374	−0.0193
	ATM	0.2189	0.2104	0.0582
	W-LDA	0.2354	0.2579	0.0725
	GTM	**0.2464**	**0.3251**	**0.0950**
NYTimes	LDA	0.2128	0.3083	0.0773
	NVDM	0.1342	−0.4131	−0.1437
	ProdLDA	0.1964	−0.0035	−0.0282
	ATM	0.2375	0.3568	0.0899
	W-LDA	0.2253	0.3352	0.0783
	GTM	**0.2443**	**0.3776**	**0.0911**

Table 1: Average topic coherence of 5 topic number settings (20, 30, 50, 75, 100). Bold values indicate the best performing model under the corresponding dataset/metric setting.
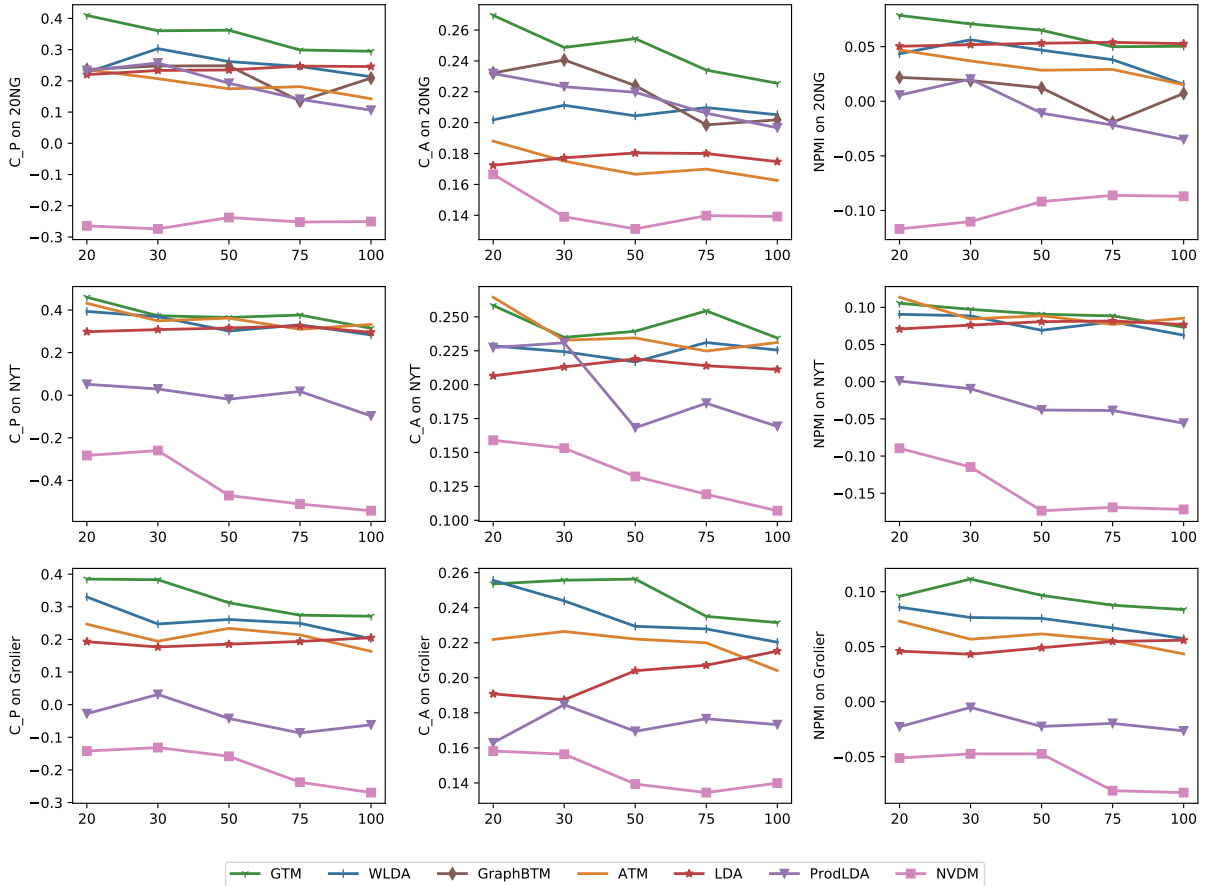
---

[1] https://github.com/AKSW/Palmetto

Figure 2: Topic coherence scores (C_P, C_A, NPMI) w.r.t. topic numbers on 20Newsgroups (20NG), NYTimes (NYT), and Grolier.

We use 2 graph convolution layers with output dimensions of 100 and K respectively in the encoder. The hidden size of the decoder is also set to 100. We use the RMSProp (Hinton et al., 2012) optimizer with a learning rate of 0.01 to train the model for 100 epochs. Since the training datasets scale up to 100K documents, i.e., 100K document nodes in the graph, it is hard to do batch training on a single GPU given the large memory requirements. We solve this issue by mini-batching the datasets and feeding to the model a subgraph consisting of 1000 document nodes and all word nodes at a training step, which results in efficient training (The training time increases almost linearly with the number of documents) and makes it possible to apply our model to even bigger datasets.

The topic coherence results on the three datasets are shown in Table 1, where each value is the average of 5 topic number settings: 20, 30, 50, 75, 100. From Table 1, we can observe that our proposed GTM is the best-performing model under all dataset/metric settings. W-LDA, ATM, LDA, and

GraphBTM alternately achieve the second-best but they are always under-performed compared to our model. As described in section 2, GTM is an extension to W-LDA with the main difference that GTM models topics in a larger context and incorporates more global information with the graph encoder. Therefore the improvements of GTM over W-LDA indicate the effectiveness of such information for topic modeling. We only experimented GraphBTM on 20Newsgroups because only 20Newsgroups preserves the sequential information that is necessary for GraphBTM to build graphs. GraphBTM performs well on the C_A metric, which is reasonable since C_A is a coherence measure based on a small sliding window of size 5 and consequently prefers models concentrating on a smaller context like GraphBTM. However, GraphBTM fails to achieve a high C_P or NPMI score, which uses a bigger window (70 and 10 respectively).

To explore how topic coherence results vary w.r.t. different topic numbers, we present in Figure 2 the topic coherence scores under different topic

Model	Topics
GTM	cancer medicine patient treatment medical disease md health hospital investigation
	satellite mission space launch lunar spacecraft shuttle orbit nasa flight
	car honda bmw engine ford saturn *dealer* turbo rear model
	ticket send mail price credit sale offer receive list customer
GraphBTM	cancer hus md medical health disease patient *mission* laboratory *culture*
	probe mission spacecraft lunar shuttle orbit nasa solar satellite space
	car bike *cop* road hit gas insurance *fbi guy lot*
	car buy *mouse scsi engine* card *audio pc windows faster*
	village turkish armenia azerbaijan troops militia greek lebanon armenian greece
W-LDA	msg food patient disease study science *one* treatment doctor scientific
	space launch nasa satellite ground mission shuttle *use* rocket orbit
	car *dog* road ride speed light drive bike *go* front
	condition sale offer shipping sell *excellent car speaker cd include*
LDA	*use* drug cause effect medical study disease patient doctor treatment
	space launch earth nasa mission system orbit satellite *design* moon
	car *buy price sale new* engine *offer* model *dealer*
	car buy sell price sale *new engine* offer *model* dealer

Table 2: Discovered topics that are most similar to 4 ground-truth categories (sci.med, sci.space, rec.autos, misc.forsale) on 20Newsgroups with topic number 50. Italics are manually labeled off-topic words.

numbers settings. It can be observed in Figure 2 that GTM enjoys the best overall performance, achieving the highest scores in most settings. LDA has a slightly higher NPMI score on 20Newsgroups dataset with 75 and 100 topics, nevertheless, GTM outperforms all baseline models with a relatively large margin on other settings of 20Newsgroups. NVDM is apparently the worst-performing model, while performances of models other than GTM and NVDM are not so consistent. Notably, W-LDA, GraphBTM, and LDA obtain the second-best overall C_P, C_A, and NPMI scores respectively. Another observation from Figure 2 is that GTM performs better on smaller topics, probably due to the fact that topics become more discriminative against each other when the topic number is small.

To gain an intuitive impression on the discovered topics, we present in Table 2 4 topics corresponding to 4 out of 20 ground-truth categories of 20Newsgroups. It can be observed that the topics discovered by GTM are more coherent and interpretable, containing few off-topic words. As a comparison, GraphBTM's rec.autos topic mixes up automobiles and criminals, W-LDA's misc.forsale topic is difficult to identify with too many off-topic words, while LDA can not distinguish between rec.autos and misc.forsale well thus recog-

nizes them as the same topic. It can be observed that GTM learns more discriminative topics by examining topic words from overlapping topics, e.g. rec.autos and misc.forsale.

4 Conclusion

We have introduced Graph Topic Model, a neural topic model that incorporates corpus-level neighboring context using graph convolutions to enrich document representations and facilitate the topic inference. Both quantitative and qualitative results are presented in the experiments to demonstrate the effectiveness of the proposed approach. In the future, we would like to extend GTM to corpora with explicit doc-doc interactions, e.g., scientific documents with citations or social media posts with user relationships. Replacing GCN in GTM with more advanced graph neural networks is another promising research direction.

References

Nikolaos Aletras and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*, pages 13–22, Potsdam, Germany. Association for Computational Linguistics.

David M. Blei. 2012. Probabilistic topic models. *Communications of the ACM*, 55(4):77.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

G. Bouma. 2009. Normalized (pointwise) mautual information in collocation extraction. In *From Form to Meaning: Processing Texts Automatically, Proceedings of the Biennial GSCL Conference 2009*, volume Normalized, pages 31–40, Tübingen.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723773.

Geoffrey Hinton, Nitish Srivastava, and Kevin Swersky. 2012. Neural networks for machine learning lecture 6a overview of mini-batch gradient descent.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 448–456, Lille, France. PMLR.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Guy Lebanon and John D. Lafferty. 2003. Information diffusion kernels. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 391–398. MIT Press.

Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard Zemel. 2015. Gated graph sequence neural networks. *arXiv preprint arXiv:1511.05493*.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 375–384, New York, NY, USA. ACM.

Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *ICML Workshop on Deep Learning for Audio, Speech and Language Processing*. Citeseer.

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, pages 1727–1736, New York, New York, USA. PMLR.

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.

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, pages 399–408, New York, NY, USA. ACM.

Linfeng Song, Yue Zhang, Zhiguo Wang, and Daniel Gildea. 2018. A graph-to-sequence model for AMR-to-text generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1616–1626, Melbourne, Australia. Association for Computational Linguistics.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Ilya Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schoelkopf. 2017. Wasserstein auto-encoders. *arXiv preprint arXiv:1711.01558*.

Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. 2020. Neural topic modeling with bidirectional adversarial training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 340–350, Online. Association for Computational Linguistics.

Rui Wang, Deyu Zhou, and Yulan He. 2019a. ATM: Adversarial-neural topic model. *Information Processing & Management*, 56(6):102098.

Rui Wang, Deyu Zhou, and Yulan He. 2019b. Open event extraction from online text using a generative adversarial network. 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 282–291, Hong Kong, China. Association for Computational Linguistics.

Liang Yao, Chengsheng Mao, and Yuan Luo. 2019. Graph convolutional networks for text classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:7370–7377.

Michihiro Yasunaga, Rui Zhang, Kshitij Meelu, Ayush Pareek, Krishnan Srinivasan, and Dragomir Radev. 2017. Graph-based neural multi-document summarization. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 452–462, Vancouver, Canada. Association for Computational Linguistics.

Xin Zhao, Jing Jiang, Hongfei Yan, and Xiaoming Li. 2010. Jointly modeling aspects and opinions with a MaxEnt-LDA hybrid. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 56–65, Cambridge, MA. Association for Computational Linguistics.

Deyu Zhou, Liangyu Chen, and Yulan He. 2014. A simple Bayesian modelling approach to event extraction from twitter. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 700–705, Baltimore, Maryland. Association for Computational Linguistics.

Jie Zhou, Ganqu Cui, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2018. Graph neural networks: A review of methods and applications. *arXiv preprint arXiv:1812.08434*.

Qile Zhu, Zheng Feng, and Xiaolin Li. 2018. GraphBTM: Graph enhanced autoencoded variational inference for biterm topic model. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4663–4672, Brussels, Belgium. Association for Computational Linguistics.