Topic Collapsing in Embedding Spaces: An Application to Education Text

Project Category: Natural Language Processing

Name: Jiner Zheng
SUNet ID: jzheng16
Graduate School of Education
Stanford University
jzheng16@stanford.edu

Name: Jon Ball
SUNet ID: jonball
Graduate School of Education
Stanford University
jonball@stanford.edu

Abstract

Neural topic models that use word embeddings reportedly outperform traditional Latent Dirichlet Allocation (LDA) topic models according to automated metrics such as topic coherence. (Dieng et al., 2020). But this finding is dubious. It begs the question of whether automated metrics for evaluating topics assess their actual usefulness. (Hoyle et al., 2021). Moreover, when both words and topics are represented as vectors in a multidimensional embedding space, they tend to collapse toward one another. This article demonstrates the "topic collapsing" problem in Dieng et al.'s implementation of the embedded topic model (ETM), with application to a novel text dataset in the domain of Education. We highlight aspects of the ETM's behavior that suggest drawbacks of topic modeling in embedding spaces, as well as a potential solution to the topic collapsing problem.

1 Introduction

Topic models are a popular unsupervised machine learning method for finding patterns of word co-occurrence in collections of documents. Blei et al., 2003 introduced the core statistical technique two decades ago, called "Latent dirichlet allocation" (LDA). More recently, with the popularity of neural methods for text analysis, researchers have attempted to combine Blei et al.'s LDA topic modeling algorithm with word embeddings. But difficulties are encountered in extending LDA to embedding spaces, where both topics and words coexist within a defined space. This feature of neural topic models results in topics "collapsing," meaning that topics start to repeat the same words as a neural topic model converges. In the limit, an overfitted neural topic model congeals into a topical blob.

Embedding spaces have properties that seem useful for topic modeling. Dieng et al., 2020 make the case that word embeddings capture information about infrequently occurring words in long-tailed vocabularies. Rare words still tend to cluster with similar words. Dieng et al. also argue that neural topic models are robust to functional words that convey little semantic information. These "stop words" tend to cluster together, which limits their inclusion in substantive topics. Intuitively, these properties of embedding spaces might influence topic quality.

Another potentially useful, but also fraught feature of word embeddings is that they can be pre-trained. A corpus vocabulary can be mapped to pre-trained GloVe (Pennington et al., 2014) or FastText (Bojanowski et al., 2017) static word embeddings, and then a neural topic model can be trained on the corpus using these prior representations. Without knowing what the dimensions in an embedding space actually represent, it may be erroneous to use pre-trained embeddings in this way. There is nonetheless an opportunity to draw on a rich prior semantics using neural topic models, which is not afforded by traditional LDA topic models.

But topic modeling in an embedding space requires that both words and topics be represented as vectors in that space. With a learning objective that minimizes the reconstruction error of a topic distribution α and word distribution β, topic and word distributions eventually collapse. This collapsing behavior depends on dataset size, embedding space dimensionality, and of course, number of iterations. A consistent trend when implementing Dieng et al.'s embedded topic model is that randomly initialized topics feature diverse words at first, but gradually become more homogenous. How then do we determine the point at which training is complete for a neural topic model?

2 Related Work

Blei et al., 2003 concisely described latent dirichlet allocation (LDA) as "a three-level hierarchical Bayesian model, in which each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities." The LDA topic model has received numerous adaptations (Blei and Lafferty, 2006, Mimno et al., 2009, Blei, 2012, Boyd-Graber et al., 2017). Recently, Dieng et al. "married" the LDA topic modeling algorithm with word embeddings (Mikolov et al., 2013) by computing the topic-word distribution as $\beta = \rho^{\top}\alpha_k$, where ρ is a matrix of word embeddings and α_k is a given topic's vector representation.

Dieng et al., 2020 are not the first to propose a neural variation of LDA. Kingma and Welling, 2013 first introduced the Variational Autoencoder (VAE), a neural network architecture for variational Bayesian modeling. Srivastava and Sutton, 2017 adapted VAE specifically for topic models, building on prior work by Dinh and Dumoulin, 2016, who initially identified a "component collapsing" problem. Numerous applications of VAE to topic modeling have been introduced (Miao et al., Burkhardt and Kramer, 2019, Zhao et al., 2020, Wu et al., 2020, Wu et al., 2021, Wang et al., 2022). Zhao et al., 2021 provide an overview.

Hoyle et al., 2021 unsettled the assumption that evaluative metrics such as topic coherence actually reflect topic quality. (Röder et al., 2015) In addition, Hoyle et al. cast doubt on established comparisons between different types of topic models, because text preprocessing pipelines are not standardized, and neither are metrics like coherence. We continue the trend of comparing different topic models' performance according to automated measures. Like Hoyle et al., we find that LDA topic models are still better in a qualitative sense, and also with regard to a measure we call "topic diversity." We show that the training behavior of LDA models and embedded topic models are diametrically opposed.

LDA topic models have been widely applied to answer research questions in the humanities and social sciences. (Hall et al., 2008, DiMaggio et al., 2013, Roberts et al., 2013, Hofstra et al., 2020). We contribute to the applied topic modeling literature by analyzing a new dataset in the domain of Education, which is described in the following section.

3 Dataset and Features

Our dataset contains metadata from English-language research articles published in Education journals using the open source software Open Journal Systems (OJS). The majority of these articles were published by editors in Indonesia, Brazil, and India. However, the dataset represents Education research published on a truly global scale. In total, the OJS data consist of 161,148 title-abstract pairs. We split the abstracts into sentences following Dieng et al.'s convention. We save each of these "documents" as a single line in a text file, meaning that each document in our corpus is either a title or a sentence describing an Education research article published using OJS. The resulting dataset is described in the two tables below.

Train	Test	Validation	**Total**
1,329,613	156,425	78,213	1,564,251

Table 1: Number of Documents in OJS Dataset

Vocab	Tokens	Max DF	Min DF
20,134	19.354M	0.7	30

Table 2: Vocabulary Size, Token Count, and Document Frequencies for OJS Dataset

In brief, our preprocessing pipeline can be described as lowercasing; removing all stop words, punctuation, and digits; and splitting on whitespace. In addition, we chose to re-implement the minimum and maximum document frequency values selected by Dieng et al.. When vectorizing our documents, we omitted all word types from our vocabulary that occurred in more than 70% of documents, or in fewer than 30 documents.

As Education researchers, we selected the OJS dataset because it would offer new information about the field of Education. The articles represented in the dataset are not widely indexed by American research services, and in fact, the scale of global OJS usage has just recently been documented. (Khanna et al., 2022) Applying topic models to this dataset revealed information about a body of Education research that remains largely unknown to professionals in the U.S. and Europe.

4 Methods

We followed the same conventions as Dieng et al., 2020 by fixing the number of topics for each of our topic models at $k = 50$ topics. We then trained pairs of LDA and embedded topic models (ETM) on identical subsets of the OJS dataset to show the opposite training behavior of each.

Latent dirichlet allocation (LDA): The LDA algorithm models each document in a corpus as a mixture of k topics, such that the mixture for that document is $\theta \sim$ Dirichlet(α_θ). The algorithm maximizes the marginal likelihood of a document $p(w|\alpha, \beta)$, which can be written as:

$$\int p(\theta|\alpha)(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta)p(w_n|z_n, \beta))d\theta \quad (1)$$

But inference over θ and z is intractable due to coupling between θ and β. (Dickey, 1983). An established method for approximation via variational inference introduces two free variational parameters: γ over θ and ϕ over z. (Jordan et al., 1999). The optimization problem is then to maximize an *evidence lower bound (ELBO)* for the marginal log likelihood $\mathcal{L}(\gamma, \phi|\alpha, \beta)$:

$$D_{KL}(q(\theta, z|\gamma, \phi)||p(\theta, z|w, \alpha, \beta)) - \log p(w|\alpha, \beta) \quad (2)$$

This enables a two step expectation maximization algorithm. First, the algorithm maximizes the lower bound of the log likelihood with respect to the variational parameters γ and ϕ. Then, with those parameters fixed, it maximizes the lower bound with respect to the model parameters α and β. These steps are repeated in alternation until the lower bound of the log likelihood converges. (Blei et al., 2003). For each of the k topics, the newly estimated β_k can then be normalized and its most likely words selected. This is the model output we are interested in: after training, we select the top 25 words belonging to each of 50 topics.

Embedded topic modeling (ETM): As with LDA, the ETM algorithm maximizes the log marginal likelihood of the documents:

$$\mathcal{L}(\alpha, \rho) = \sum_{d=1}^{D} \log p(w_d|\alpha, \rho) \quad (3)$$

where ρ is a matrix of word embeddings representing the corpus vocabulary and α is a matrix of topic embeddings. We approximate the marginal likelihood of a document with a draw from the logistic-normal distribution. The conditional distribution for each word is given by:

$$p(w_{d_n}|\delta_d, \alpha, \rho) = \sum_{k=1}^{K} \theta_{dk}\beta_{k,w_{dn}} \quad (4)$$

where a topic, or distribution over words, is induced by $\beta_{k,v} = \text{softmax}(\rho^\top \alpha_k)$.

Variational inference is again required. (Blei et al., 2017). We use the reparameterization trick (Kingma and Welling, 2013) and data subsampling (Hoffman et al., 2013), where $\mathcal{B}$ is a minibatch of documents drawn from the corpus D, to approximate the *ELBO* of the marginal log likelihood. Finally, we note that the KL divergence can be rewritten as a closed-form expression in terms of the diagonal covariance matrix Σ_d:

$$\mathrm{D}_{KL}(q(\delta_d; w_d, v)||p(\delta_d)) = \frac{1}{2}\{tr(\Sigma_d) + \mu_d^\top \mu_d - \log det(\Sigma_d) - K\} \quad (5)$$

As with LDA, the ETM algorithm alternates between fixing model parameters and fixing variational parameters, updating each in turn.

5 Results

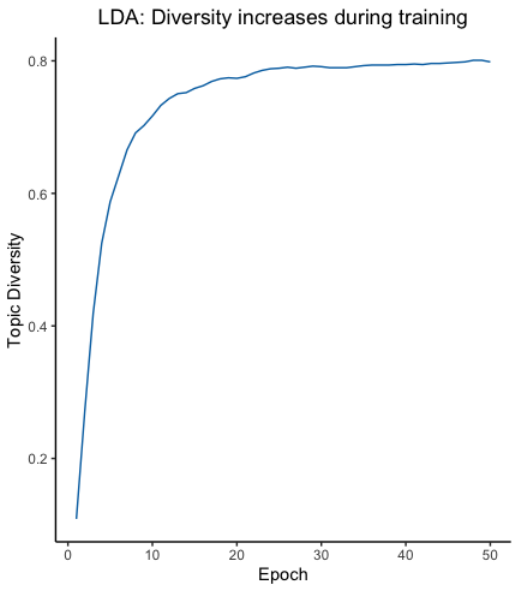
We compare LDA and embedded topic models on the basis of an intuitive metric: topic diversity. Topic diversity is calculated as the number of unique words in a $k \times$ *top_n* topic-word matrix, divided by the total number of words in the matrix. For all experiments described in this article, we used Dieng et al.'s defaults: $k = 50$ topics, 25 top words per topic ($\arg\max_{[\beta_{i...k}]} \in \mathbb{R}^{k \times 25}$). The intuition guiding topic diversity is that a sufficiently large document set should feature distinct topics. Distinctness depends on whether a topic features unique vocabulary: if several topics feature the same words, users may struggle to interpret them. Topic diversity is also well suited for describing the "topic collapsing" problem.

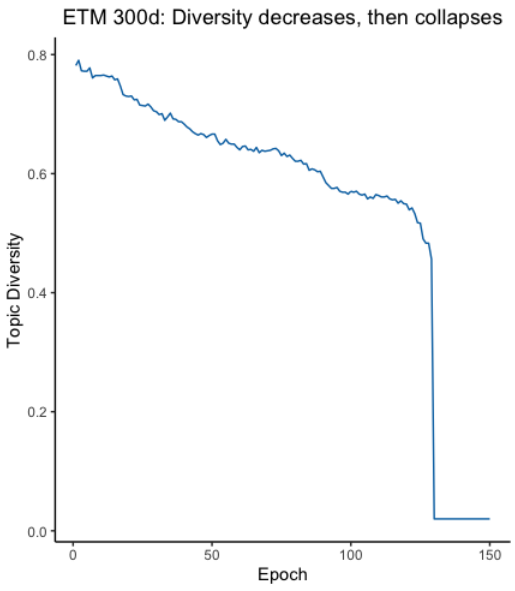
5.1 Experiment

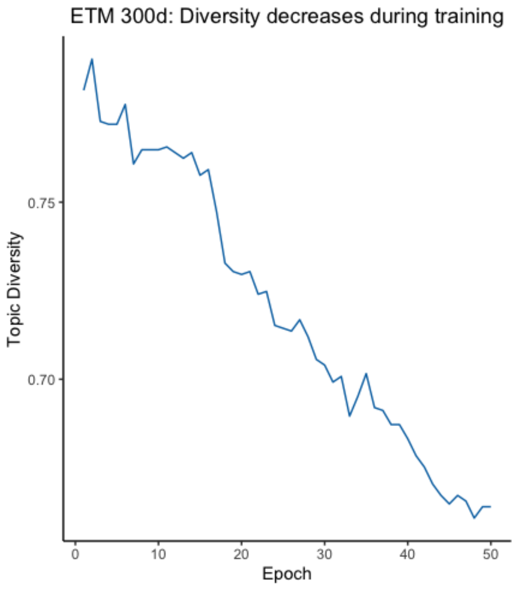
We compare the training behavior of a baseline LDA topic model against the training behavior of an ETM. The figures below plot topic diversity as a function of training epochs for our baseline LDA model and an ETM initialized with 300-dimensional static word embeddings, which were pre-trained on OJS corpus data using the FastText algorithm. (Bojanowski et al., 2017).

Whereas the LDA topic model displays a smooth ascent to more diverse topics over iterations, the ETM repeats more words across topics as training progresses. In the limit, the ETM repeats more words across topics until it collapses abruptly.

Topic collapsing behavior depends on dataset







and embedding size. When trained on small subsets of OJS documents, the LDA topic displays typical behavior. But the 300-dimensional ETM initially learns more diverse topics, then decreases in diversity until it collapses. 300-dimensional embeddings were chosen to demonstrate the embedded topic model's behavior in highly multidimensional spaces. We also calculated embedded topic model diversity curves in 200-, 100-, and 50-dimensional embedding spaces. The embedded topic model consistently collapses past a certain training threshold.

5.2 Discussion

An intuitive explanation for the ETM's behavior has to do with embedding spaces. Random initialization of topic embeddings in a highly multidimensional space produces diverse topics at the outset of training. But the gradual decrease in diversity holds over iterations and dataset sizes. With the 300-dimensional ETM, topics become less diverse until the model collapses at approximately 130 iterations. Perplexity also decreases, but this metric becomes incalculable past the point of collapse, because collapsed topics have no entropy.

Dieng et al.'s workaround for this problem is that they save the model state that minimizes perplexity during training. But with the 300-dimensional ETM, perplexity quickly converges and then perturbs slightly until the model collapses. Dieng et al.'s workaround results in a model with minimal perplexity score but topic diversity score 0.5544. Compared against the LDA topic model, which quickly converges to a minimum perplexity and then continues to increase in diversity, the ETM's behavior does not suggest a clear decision boundary for selecting a model over iterations.

Qualitatively, a 100-dimensional ETM delivers the most interpretable topics after 50 epochs of training. Based on our reading of both an LDA model and a 100-dimensional ETM trained for 50 epochs each, most topics are shared by both. This suggests the topics' validity as latent corpus variables. However, as the diversity curves suggest, we find that an LDA model delivers richer topics than an ETM after 50 epochs of training. Selected topics are presented below.

Some topics in the corpus are clearly unique to OJS. For example, a prominent topic in the OJS corpus relates to Islamic religious education and ethics. It features in articles written by Indonesian

learning	model	outcomes
student	based	achievement
cooperative	application	improve

Table 3: LDA: Learning Outcomes Topic

learning	student	outcomes
achievement	motivation	mathematics
cooperative	students	mastery

Table 4: ETM: Learning Outcomes Topic

problem	students	thinking
understanding	concept	skills
strategy	critical	solving

Table 5: LDA: Problem Solving Topic

problem	solving	questions
task	question	meaning
representation	tasks	errors

Table 6: ETM: Problem Solving Topic

authors, often about *pesantren*, or Islamic boarding schools. Another topic includes words such as development, tourism, and aid. With the exception of these two topics, the remaining 48 could be plausibly drawn from collections of documents about education in the U.S.

Two observations bear mentioning. First, the embedded topic model repeats plural forms of words in topics. Following Hoyle et al., 2021's suggestion, we did not lemmatize words. "Student" and "students" tend to co-occur more frequently in embedded topics than in LDA topics. This is likely because the word embeddings for singular and plural forms of words are nearly identical, especially when those embeddings are trained using sub-word information, as with FastText. LDA topic models seemingly better distinguish between different word forms, whereas embedded topic models seem to learn lemmas.

Second, the same property Dieng et al. identified as a strength of topic modeling in embedding spaces – that functionally similar words cluster together and do not intrude into other topics – may also be a weakness. The embedded topic models we trained learned a "countries and nations" topic from the OJS corpus. This topic is a cluster of country names: Malaysia, Kazakhstan, etc. It is likely that a similar topic would be present in any corpus embedding space. While this topic does provide useful information about OJS, it may not be ideal model behavior that an ETM returns such narrow clusters of functionally related nouns. When we did not remove stop words, for example, the ETM produced lists of pronouns and prepositions as topics. This raises a question for topic modeling: are clusters of related function words, or perhaps proper nouns and names, correctly identified as topics? Neural topic models may be especially sensitive to these clusters.

Finally, a potential solution to the topic collapsing problem might involve clustering regularization, such that topic embeddings are forced into distinct centers of separately aggregated word embedding clusters. Effective clustering regularization can support the joint optimization of topic and word embeddings and produce sparse soft-assignments of word embeddings to topics. This can potentially mitigate collapsing behavior because it would ensure the optimal transport of word embeddings to just one topic embedding, making topic embeddings more sparsely distributed and supporting more diverse semantics in the space. (Anonymous, 2023).

6 Conclusion

Neural topic models have potential because embedding spaces have numerous useful properties. However, there are tradeoffs when we model topics in embedding spaces, as opposed to modeling topics in the latent space of dirichlet allocation. In the limit, minimizing the reconstruction error between word and topic embeddings causes topics to collapse into one another. Clustering regularization is one potential solution to this topic collapsing problem. A deeper problem, and one less fully understood, is that word embeddings may also encode a specific semantics that causes neural topic models to subtly misrepresent the topics contained in a text corpus. Because topics are latent random variables which can only be approximated via variational inference, we have no real way of knowing whether neural topic models accurately generate the topics hidden in vast collections of documents. Future research promises to develop solutions and workarounds for the shortcomings of present neural topic models. In the process, such studies may develop our understanding of embedding spaces, as well as how words, topics, and meanings are represented within them.

Contributions

Jiner Zheng: LDA topic model implementation, visualization, literature review, report writing

Jon Ball: embedded topic model implementation, data collection, literature review, report writing

References

Anonymous. 2023. Neural topic modeling with embedding clustering regularization. In *Submitted to The Eleventh International Conference on Learning Representations*. Under review.

David M. Blei. 2012. Probabilistic topic models. *Commun. ACM*, 55(4):77–84.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120.

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

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jordan Boyd-Graber, Yuening Hu, David Mimno, et al. 2017. Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Sophie Burkhardt and Stefan Kramer. 2019. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *J. Mach. Learn. Res.*, 20(131):1–27.

James M Dickey. 1983. Multiple hypergeometric functions: Probabilistic interpretations and statistical uses. *Journal of the American Statistical Association*, 78(383):628–637.

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

Paul DiMaggio, Manish Nag, and David Blei. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of u.s. government arts funding. *Poetics*, 41(6):570–606. Topic Models and the Cultural Sciences.

Laurent Dinh and Vincent Dumoulin. 2016. Training neural bayesian nets.

David Hall, Dan Jurafsky, and Christopher D Manning. 2008. Studying the history of ideas using topic models. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 363–371.

Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. 2013. Stochastic variational inference. *Journal of Machine Learning Research*.

Bas Hofstra, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. 2020. The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, 117(17):9284–9291.

Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken?: the incoherence of coherence. *arXiv preprint arXiv:2107.02173*.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. 1999. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233.

Saurabh Khanna, Jon Ball, Juan Pablo Alperin, John Willinsky, et al. 2022. Recalibrating the scope of scholarly publishing: A modest step in a vast decolonization process.

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

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736. PMLR.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26.

David Mimno, Hanna Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. 2009. Polylingual topic models. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 880–889.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Edoardo M Airoldi, et al. 2013. The structural topic model and applied social science. In *Advances in neural information processing systems workshop* *on topic models: computation, application, and evaluation*, volume 4, pages 1–20. Harrahs and Harveys, Lake Tahoe.

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*, page 399–408. Association for Computing Machinery.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.

Dongsheng Wang, Dandan Guo, He Zhao, Huangjie Zheng, Korawat Tanwisuth, Bo Chen, and Mingyuan Zhou. 2022. Representing mixtures of word embeddings with mixtures of topic embeddings. *arXiv preprint arXiv:2203.01570*.

Xiaobao Wu, Chunping Li, and Yishu Miao. 2021. Discovering topics in long-tailed corpora with causal intervention. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 175–185.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Short text topic modeling with topic distribution quantization and negative sampling decoder. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1772–1782.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. 2020. Neural topic model via optimal transport. *arXiv preprint arXiv:2008.13537*.