EAI Endorsed Transactions
on Internet of Things
Research Article EAI.EU

Use of Neural Topic Models in conjunction with Word Embeddings to extract meaningful topics from short texts

Nassera HABBAT[1,*], Houda ANOUN[1], Larbi HASSOUNI[1] and Hicham NOURI[2]

[1]RITM Laboratory, CED ENSEM Ecole Superieure de Technologie Hassan II University, Casablanca, Morocco,
[2]Research Laboratory on New Economy and Development (LARNED), Faculty of Legal Economic and Social Sciences AIN SEBAA, Hassan II University, Casablanca, Morocco

Abstract

Unsupervised machine learning is utilized as a part of the process of topic modeling to discover dormant topics hidden within a large number of documents. The topic model can help with the comprehension, organization, and summarization of large amounts of text. Additionally, it can assist with the discovery of hidden topics that vary across different texts in a corpus. Traditional topic models like pLSA (probabilistic latent semantic analysis) and LDA suffer performance loss when applied to short-text analysis caused by the lack of word co-occurrence information in each short text. One technique being developed to solve this problem is pre-trained word embedding (PWE) with an external corpus used with topic models. These techniques are being developed to perform interpretable topic modeling on short texts. Deep neural networks (DNN) and deep generative models have recently advanced, allowing neural topic models (NTM) to achieve flexibility and efficiency in topic modeling. There have been few studies on neural-topic models with pre-trained word embedding for producing significant topics from short texts. An extensive study with five NTMs was accomplished to test the efficacy of additional PWE in generating comprehensible topics through experiments with different datasets in Arabic and French concerning Moroccan news published on Facebook pages. Several metrics, including topic coherence and topic diversity, are utilized in the process of evaluating the extracted topics. Our research shows that the topic coherence of short texts can be significantly improved using a word embedding with an external corpus.

Keywords: Neural Topic Models, Pre-training word embedding, Short text, Topic coherence.

Received on 30 July 2022, accepted on 29 September 2022, published on 30 September 2022

doi: 10.4108/eetiot.v8i3.2263

*Corresponding author. Email: nassera.habbat@gmail.com

1. Introduction

People are increasingly becoming emotionally attached to sharing information through diverse online social platforms, like Twitter, Facebook, webpages, etc., due to the fast development of information and communications technologies and extensive internet use. These messages sent via web and social networks include vital information about actual social trends and situations, opinions of people on various services and products, advertisements, government policy announcements, etc. To easily and quickly read through these huge numbers of messages and extract relevant information, an effective text processing technique is required. A topic modeling method is proven an efficient method for semantic understanding of textual data in traditional natural language processing (NLP). Conventional topic models [1], [2], like LDA [3] or pLSA [4] and their versions, are extremely effective at producing latent semantic structures from an unlabeled text and are popular in rapidly developing topic identification, comment summarization, classification of documents, and event tracking.

In comparison to the length of largely formal texts like scientific articles or newspapers, messages published on different social media are usually short. These short texts share the following major characteristics:

1. A limited number of words per document.
2. The use of unique and informal terminology.
3. Post length restrictions.
4. Word meanings and usage may differ based on the posting.
5. Inappropriate comments (or "spam").

The implementation of traditional topic models (TTM) for analysis of short text yields poor results caused by the absence of word co-occurrence information within every document of short text, which stems from the short text's characteristics mentioned above.

To address the issues of topic modeling by TTM on short text, neural topic models (NTMs) with interesting achievements have become available, as have significant advancements in word embedding that provide an efficient approach to understanding relations of semantic words from a large text, which can aid in the development of models for producing more comprehensible and coherent topics.

The current study investigates computationally efficient and simple techniques to enhance the comprehensibility of extracted topics from real-world short texts using NTMs. The hardest part of topic modeling for short texts is learning context information, and incorporating pre-trained word embedding into an NTM appears to be among the most effective methods of expressly enriching the content knowledge.

In summary, our contribution is as follows: An examination of the effectiveness of different NTMs in terms of the quality of generated topics, as tested by many metrics of topic coherence and topic diversity, with two pre-trained word embedding models.

In the next section, a short presentation of related works on neural topic models (NTM) is provided, followed by a brief definitions of neural topic models. Section 4 presents simulation experiments and results, and section 5 summarises this paper.

2. Related works

The most frequent neural topic models (NTMs) are based on the variational autoencoder (VAE) [5], a deep generative model, and amortised variational inferences (AVI) [6]. The following section describes the basic framework of NTMs based on VAE, in which inference and generative mechanisms are modelled by neural network-based encoders and decoders, respectively. Inference in NTMs is computationally easier than in traditional Bayesian probabilistic topic models (BPTM), their application is simplified by the abundance of advanced deep learning techniques, and consequently, NTMs are used easily with PWEs for the acquisition of prior knowledge. Divers VAE-based NTM types have been introduced. Neural Variational Latent Dirichlet Allocation (NVLDA) [7], The Neural Variational Document Model (NVDM) [8], Dirichlet Variational Autoencoder (DirVAE) [9], Dirichlet Variational Autoencoder topic model (DVAE) [10], iTM-VAE [11], and the Gaussian Softmax Model (GSM) [12] are a few examples. This is not an extensive list, and it is still increasing.

Just several researchers utilized NTMs instead of conventional topic models to extract meaningful, coherent, and understandable topics from short texts by integrating contextual and semantic information. An integration of NTM with either a memory network or a recurrent neural network (RNN) was used in [13], [14], in which topics developed by the NTM were used for classification by a memory network or an RNN. For both works, the NTM outperforms conventional topic models regarding topic coherence and classification task performance. Lin et al. [15] used Archimedean copulas to make distributions of multiple topics in a short text more distinct. However, Wu et al. [16] suggested a novel NTM with a quantization approach for topic-distribution, resulting in the best distributions, as well as a negative sampling decode, having to learn to reduce redundant topics. As a consequence, their proposed technique outperforms standard topic models.

Niu et al. [17] combined short texts into long texts or document and used document embedding to create word co-occurrence data. Zhao et al. [18] proposed a variational autoencoder topic model (VAETM) and its supervised variant (SVAETM) by mixing embedded representations of entities and words with an external dataset. To improve contextual information, Zhu et al. [19] presented a graph neural network as the NTM encoder, which receives a bi-term graph of words as input and gives as output the corpus's topic distribution. However, Feng et al. [20] presented a context-reinforced NTM based on the assumption of a few pertinent topics by each short text, with pre-trained word embedding informing the topic word distributions.

3. The proposed model

We outline in this section, the applied architecture and briefly discuss the various neural topic models that have been employed.

3.1. Global architecture

According to an analysis of current research on NTMs for analysis of short text, using auxiliary data from an outside corpus is one of the most common and successful ways to deal with short-text sparsity.

As seen in Figure 1, we used web scraping techniques (especially Request and BeautifulSoup Python libraries) to collect Posts from Facebook pages; then, those posts were pre-processed using NLP techniques. For pre-trained word embedding, we utilised GloVe [21] and Word2Vec [22]. For topic modelling, we compared five neural topic models that will be defined in the next part.

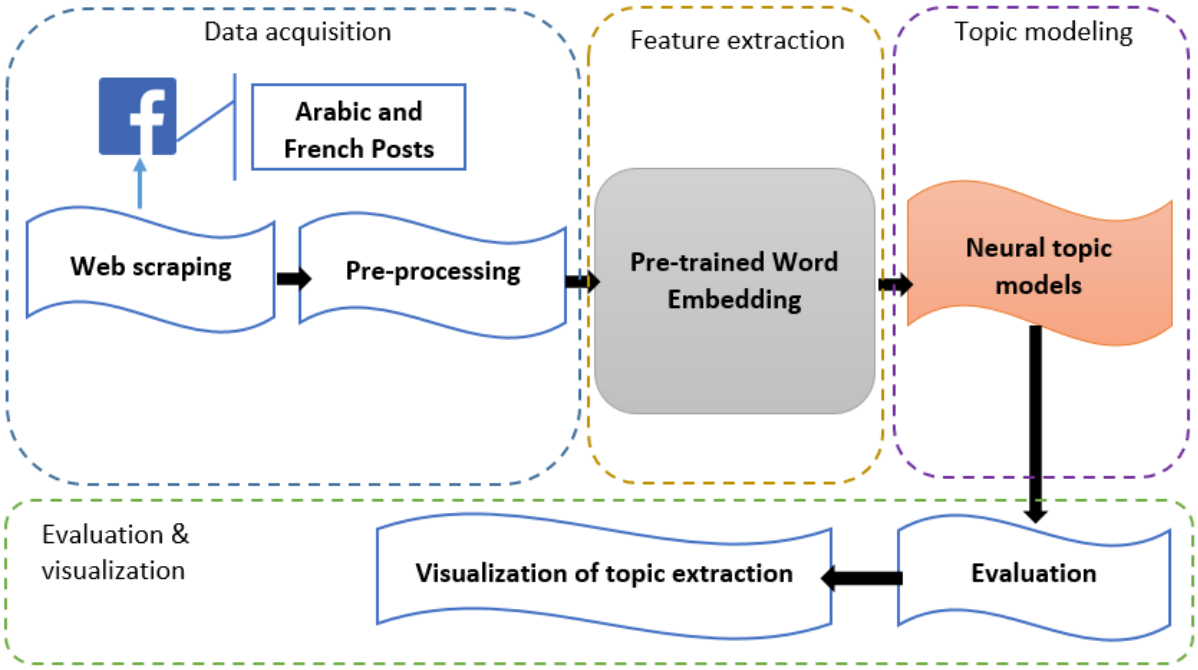

Figure 1. The global architecture.

3.2. Neural Topic Models for Analysis

We briefly describe in this section, the neural topic models utilized in this investigation. The significance of the notations used to describe models is provided in Table 1.

Table 1. List of used notations.

Indices:	
K	Number of topics $k \in \{1,\ldots,K\}$
L	Word embedding vectors dimension $l \in \{1,\ldots,L\}$
C	Number of classes $c \in \{1,\ldots,C\}$
Decision Variables:	
D	Set of documents
V	Set of lexicons, vocabularies
X	BoW matrix of all documents, $X \in \mathbb{R}_+^{\lvert V\rvert\times\lvert D\rvert}$
x_d	d's BoW representation vector, $x_d \in \mathbb{R}_+^{\lvert V\rvert}$
N	Number of words that occurred in document d
w_n	n-th word
Random Variables:	
$h^{(i)}$	i-th hidden layer's outputs
h	Gaussian random variables, $h \in \mathbb{R}^K$
z_n	latent topic for the n-th word
θ	Topic proportion vector, $\theta \in \mathbb{R}_+^K$
β	Topic–word distribution $\beta \in \mathbb{R}^{\lvert V\rvert\times K}$
α	Topic centroid vectors $\alpha \in \mathbb{R}^{L\times K}$
ρ	Word embedding vectors $\rho \in \mathbb{R}^{L\times\lvert V\rvert}$

Figure 2 depicts the generalized architecture of the neural topic models based on variational autoencoders (VAE). The encoder is the part of the network that generates θ by mapping the bag-of-words (BoW) input to a latent document-topic vector, and the decoder is the part of the network that takes θ and gives *p(x)* by mapping the vector of document-topic to a separate distribution

through the words in the vocabulary. They are named auto-encoders because the decoder attempts to rebuild the input's word distribution. In VAE, h is sampled using a Gaussian distribution, and θ is provided by transforming it.

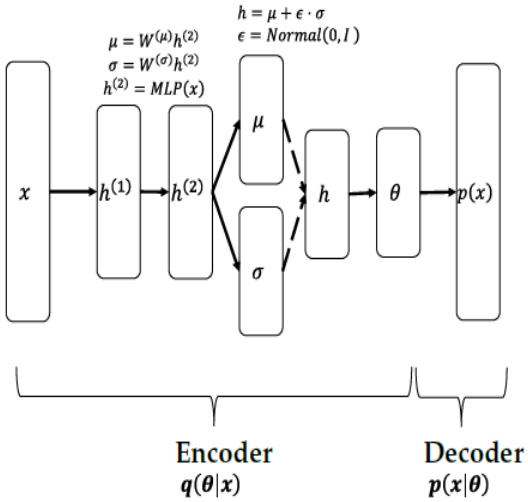

Figure 2. Representation of the model based on VAE.

NVDM

To our knowledge, Neural Variational Document Model or NVDM [8] is the first VAE-based document method proposed with a multilayer perceptron encoder. The pattern *h* from the distribution of Gaussian serves as the input for the decoder in this model, and variational inference is related to reducing KL divergence. NVDM is a general VAE, whereas the majority of subsequent NTMs regenerate h to handle θ as a vector of the topic proportion.

NVLDA

Another variant of NVDM is Neural Variational Latent Dirichlet Allocation or NVLDA [7], which utilizes Neural Variational Inference to replicate LDA. To transform z to θ in this case, the Softmax function is utilized. The Logistic-Normal distribution, used as a surrogate for the Dirichlet distribution, is the probability distribution that converts Gaussian distribution samples to the Softmax basis. Furthermore, the decoder is p(x) = softmax(β)·θ. This topic model, as opposed to the NVDM, in which both the topic proportions and the topic-word distribution are probability distributions. The following is the definition of the Logisitic-Normal distribution:

$$h \sim Normal(\mu, \sigma^2)$$
$$\theta = softmax(h)$$

ProdLDA

Product-of-Experts Latent Dirichlet Allocation or ProdLDA [7] is an extended form of NVLDA where the decoder is constructed using the expert model's product and the topics-word distribution is unnormalized.

WLDA

Wasserstein Latent Dirichlet Allocation or WLDA [23] is a Wasserstein auto-encoder-based topic model (WAE) (Figure 3). But while diverse probability distributions may be utilized for the prior distribution of θ, we used in this work the Dirichlet distribution, the most basic. The training based on GAN (Generative Adversarial Network) and MMD (Maximum Mean Discrepancy) are available in WAE, but MMD is utilized in WLDA due to the simplicity of training loss convergence.

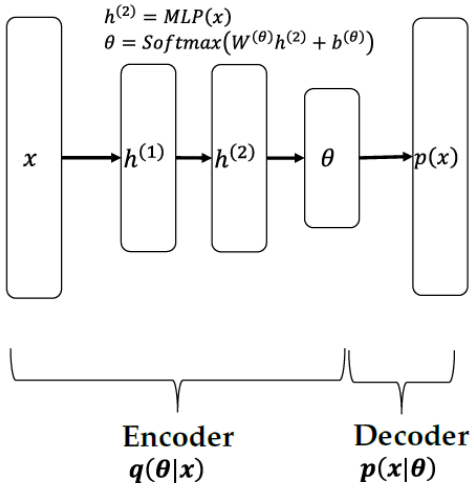

Figure 3. Representation of the model based on WAE.

Note that WAE generates θ directly using the Softmax function, so no sampling is needed.

NSTM

As with WLDA, Neural Sinkhorn Topic Model or NSTM [24] is trained to utilize optimal transport [42]. Because we suppose that q computes x into a low-dimensional latent space whilst still conserving adequate information about x, the Sinkhorn Algorithm calculates the optimised transport distance between x and θ. The loss function is the sum of the negative log-likelihood and the optimized transport distance.

4. Results from Simulation Experiments

The datasets, the evaluation measurement utilized in this work, and the results are presented in this part.

4.1. Datasets

We focused our analysis on Arabic and French Facebook posts produced by Moroccan news pages, and we chose four Facebook pages (Le Matin.ma, L'Economiste, Hespress, and Medi1TV). We collected approximately 81 000 posts written between November 11, 2021, and April 28, 2022. (Details are shown in Table 2).

Table 2. Description of the data.

Language	Facebook page	Number of posts
Arabic	**Hespress**	22 769
	Medi1TV	18 157
French	**Le Matin**	20 745
	L'Economiste	19 349
Total		**81 020**

4.2. Evaluation metrics

Multiple measures with two main directions have been introduced to assess the quality of the top-N words. The first is to determine whether the meanings of the top-N words are coherent with one another, which is known as topic coherence (TC). The second one is to assess the topic diversity (TD) or topic uniqueness of the top N words for every pair of topics. For topic coherence we used NPMI and WETC metrics described below.

NPMI

NPMI or Normalized Point-Wise Mutual Information measures a group of words' semantic coherence. It is calculated from the following equation and is regarded as having the strongest correlations with human evaluations.

$$NPMI = \sum_{i=1}^{k} \sum_{j=m+1}^{k} \frac{\log \frac{P(m_i, m_j)}{P(m_i)p(m_j)}}{-\log P(m_i, m_j)}$$

Where *m* is the top N words for a given topic.

WETC

Word Embeddings Topic Coherence or WETC denotes topic coherence based on word embeddings, and pair-wise WETC for a specific topic is computed as:

$$WETC_{Pll}(E^{(k)}) = \frac{1}{N(N-1)} \sum_{j=2}^{N} \sum_{i=1}^{j-1} \langle E_i^{(k)}, E_j^{(k)} \rangle$$

Where $\langle .,. \rangle$ represents the inner product Pretrained weights of word2vec/ Glove were used to calculate the WETC score, and $E^{(k)}$ is the sequence of Word2vec/Glove word embedding vector equivalent to the top N words for a specific topic k; $E_i^{(k)}$ implies and all vectors are adjusted as follows: $\|E_{i,:}^{(k)}\| = 1$, N is gotten as 10.

$WETC_c$ (centroid WETC) is computed as follows:

$$WETC_c(E^{(k)}) = \frac{1}{N} \sum_{n=1}^{N} E^{(k)} t$$

$$t = \frac{\alpha_{:,k}}{\|\alpha_{:,k}\|}$$

Topic Diversity (TD)

Topic diversity [25] is the proportion of distinct words in the top 25 words among the topics. A TD near 0 indicates a repetitive topic, while TD near 1 reflect more diverse topics.

We also employed an additional metric, inverted rank-biased overlap or InvertedRBO [26], which assesses disjointness around topics based on the top-N words and weighted according to word rankings. The higher this metric, the better.

4.3. Results analysis

The simulation experiments were conducted with several datasets, and the topic models' performance was assessed using topic coherence and topic diversity metrics.

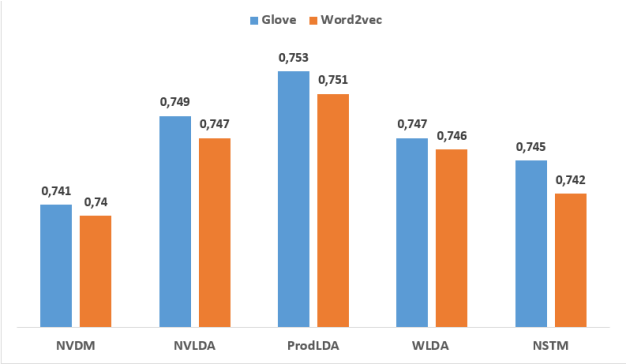

Figure 4. The NPMI of the compared models for the Arabic dataset.

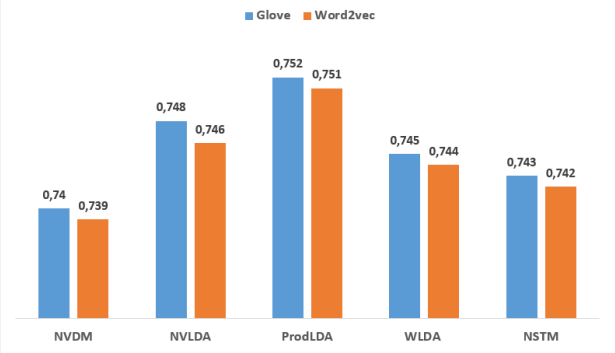

Figure 5. The NPMI of the compared models for the French dataset.

Figures 4 and 5 show the NPMI of various models, while Tables 3 and 4 show the detail results of the WETC topic coherence metric and topic diversity metrics for various NTMs and datasets, respectively. The values in bold are the best performances. As word embedding models, we used GloVe and Word2vec.

Table 3. The WETC and TD for different models on the Arabic dataset.

NTM	PWE	WETC	TD	InvertedRBO
NVDM	**Glove**	0.29	0.77	0.99
	Word2vec	0.24	0.75	0.99
NVLDA	**Glove**	0.42	0.85	**1.00**
	Word2vec	0.39	0.84	0.99
ProdLDA	**Glove**	**0.49**	**0.91**	**1.00**
	Word2vec	0.47	0.86	**1.00**
WLDA	**Glove**	0.38	0.84	**1.00**
	Word2vec	0.35	0.82	0.99
NSTM	**Glove**	0.34	0.81	**1.00**
	Word2vec	0.30	0.76	0.99

Table 4. The WETC and TD for different models on the French dataset.

NTM	PWE	WETC	TD	InvertedRBO
NVDM	**Glove**	0.30	0.76	0.99
	Word2vec	0.26	0.75	0.99
NVLDA	**Glove**	0.41	0.84	**1.00**
	Word2vec	0.40	0.82	**1.00**
ProdLDA	**Glove**	**0.48**	**0.90**	**1.00**
	Word2vec	0.46	0.87	**1.00**
WLDA	**Glove**	0.37	0.85	**1.00**
	Word2vec	0.34	0.83	0.99
NSTM	**Glove**	0.34	0.82	**1.00**
	Word2vec	0.31	0.77	0.99

For many datasets, ProdLDA + Glove has the highest TC with 0.753 on the Arabic dataset and 0.752 on the French dataset. When it comes to TD metrics, the InvertedRBO value is almost always the best for most models. This shows that there is enough variety in most situations.

According to the findings, the TC metric differs significantly according to the used word embedding model. This finding indicates that the effectiveness of the word embeddings may have a substantial effect on topic model training.

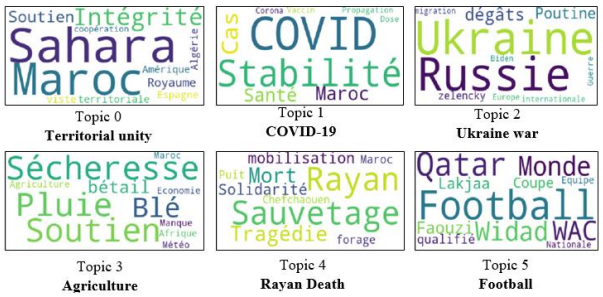

Figure 6. The extracted topics from the French dataset.

Figure 6 shows the word cloud of the six relevant generated using ProdLDA + Glove on French datasets.

5. Summary and future work

Through social networking sites, short-text data are becoming more and more common in the real world. Every day, it's also becoming more important to figure out what these short messages mean. Because short texts aren't as long as long texts or documents, they don't have as much information about how words are used together. This makes it hard for popular topic model techniques to create topics that make sense and are easy to understand.

Using word embeddings that have already been trained in neural topic models is an effective method for rapidly enhancing the quality of the generated topics, as evaluated by topic diversity and topic coherence in our experiments, which show that using ProdLDA with Glove as PWE gives the best performance and more coherent topics.

As Future direction, we aim to investigate the use of contextualized word embedding based on the transformers like BERT with neural topic models.

References

[1] N. Habbat, H. Anoun, et L. Hassouni, « Topic Modeling and Sentiment Analysis with LDA and NMF on Moroccan Tweets », in *Innovations in Smart Cities Applications Volume 4*, Cham, 2021, p. 147-161.

[2] N. Habbat, H. Anoun, et L. Hassouni, « Sentiment Analysis and Topic Modeling on Arabic Twitter Data during Covid-19 Pandemic », *Indones. J. Innov. Appl. Sci. IJIAS*, vol. 2, nº 1, p. 60-67, févr. 2022, doi: 10.47540/ijias.v2i1.432.

[3] D. M. Blei, A. Y. Ng, et M. I. Jordan, « Latent Dirichlet Allocation », *J Mach Learn Res*, vol. 3, nº null, p. 993-1022, mars 2003.
[4] T. Hofmann, « Unsupervised Learning by Probabilistic Latent Semantic Analysis », p. 20.
[5] D. P. Kingma et M. Welling, « Auto-Encoding Variational Bayes », *ArXiv13126114 Cs Stat*, mai 2014, Consulté le: 10 mars 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1312.6114
[6] D. J. Rezende, S. Mohamed, et D. Wierstra, « Stochastic Backpropagation and Approximate Inference in Deep Generative Models », *ArXiv14014082 Cs Stat*, mai 2014, Consulté le: 16 mars 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1401.4082
[7] A. Srivastava et C. Sutton, « Autoencoding Variational Inference For Topic Models », *ArXiv170301488 Stat*, mars 2017, Consulté le: 12 janvier 2021. [En ligne]. Disponible sur: http://arxiv.org/abs/1703.01488
[8] Y. Miao, L. Yu, et P. Blunsom, « Neural Variational Inference for Text Processing », *ArXiv151106038 Cs Stat*, juin 2016, Consulté le: 16 mars 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1511.06038
[9] W. Joo, W. Lee, S. Park, et I.-C. Moon, « Dirichlet Variational Autoencoder », *ArXiv190102739 Cs Stat*, janv. 2019, Consulté le: 16 mars 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1901.02739
[10] S. Burkhardt et S. Kramer, « Decoupling Sparsity and Smoothness in the Dirichlet Variational Autoencoder Topic Model », p. 27.
[11] X. Ning, Y. Zheng, Z. Jiang, Y. Wang, H. Yang, et J. Huang, « Nonparametric Topic Modeling with Neural Inference », *ArXiv180606583 Cs*, juin 2018, Consulté le: 16 mars 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1806.06583
[12] Y. Miao, E. Grefenstette, et P. Blunsom, « Discovering Discrete Latent Topics with Neural Variational Inference », *ArXiv170600359 Cs*, mai 2018, Consulté le: 16 mars 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1706.00359
[13] X. Wang et Y. YANG, « Neural Topic Model with Attention for Supervised Learning », in *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, août 2020, vol. 108, p. 1147-1156. [En ligne]. Disponible sur: https://proceedings.mlr.press/v108/wang20c.html
[14] J. Zeng, J. Li, Y. Song, C. Gao, M. R. Lyu, et I. King, « Topic Memory Networks for Short Text Classification ». arXiv, 10 septembre 2018. Consulté le: 26 juillet 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1809.03664
[15] L. Lin, H. Jiang, et Y. Rao, « Copula Guided Neural Topic Modelling for Short Texts », in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, New York, NY, USA: Association for Computing Machinery, 2020, p. 1773-1776. [En ligne]. Disponible sur: https://doi.org/10.1145/3397271.3401245
[16] X. Wu, C. Li, Y. Zhu, et Y. Miao, « 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)*, Online, nov. 2020, p. 1772-1782. doi: 10.18653/v1/2020.emnlp-main.138.
[17] Y. Niu, H. Zhang, et J. Li, « A Nested Chinese Restaurant Topic Model for Short Texts with Document Embeddings », *Appl. Sci.*, vol. 11, nº 18, 2021, doi: 10.3390/app11188708.
[18] X. Zhao, D. Wang, Z. Zhao, W. Liu, C. Lu, et F. Zhuang, « A neural topic model with word vectors and entity vectors for short texts », *Inf. Process. Manag.*, vol. 58, nº 2, p. 102455, mars 2021, doi: 10.1016/j.ipm.2020.102455.
[19] Q. Zhu, Z. Feng, et X. Li, « GraphBTM: Graph Enhanced Autoencoded Variational Inference for Biterm Topic Model », in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, oct. 2018, p. 4663-4672. doi: 10.18653/v1/D18-1495.
[20] J. Feng, Z. Zhang, C. Ding, Y. Rao, et H. Xie, « Context Reinforced Neural Topic Modeling over Short Texts ». arXiv, 11 août 2020. Consulté le: 26 juillet 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/2008.04545
[21] J. Pennington, R. Socher, et C. Manning, « Glove: Global Vectors for Word Representation », in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, 2014, p. 1532-1543. doi: 10.3115/v1/D14-1162.
[22] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, et J. Dean, « Distributed Representations of Words and Phrases and their Compositionality », *ArXiv13104546 Cs Stat*, oct. 2013, Consulté le: 6 mars 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1310.4546
[23] F. Nan, R. Ding, R. Nallapati, et B. Xiang, « Topic Modeling with Wasserstein Autoencoders », in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, juill. 2019, p. 6345-6381. doi: 10.18653/v1/P19-1640.
[24] H. Zhao, D. Phung, V. Huynh, T. Le, et W. Buntine, « Neural Topic Model via Optimal Transport », 2021. [En ligne]. Disponible sur: https://openreview.net/forum?id=Oos98K9Lv-k
[25] A. B. Dieng, F. J. R. Ruiz, et D. M. Blei, « Topic Modeling in Embedding Spaces ». arXiv, 7 juillet 2019. Consulté le: 10 juin 2022. [En ligne]. Disponible sur: http://arxiv.org/abs/1907.04907
[26] G. Carbone et G. Sarti, « ETC-NLG: End-to-end Topic-Conditioned Natural Language Generation », *Ital. J. Comput. Linguist.*, vol. 6, p. 61-77, déc. 2020, doi: 10.4000/ijcol.728.