IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received November 30, 2020, accepted December 14, 2020, date of publication December 23, 2020, date of current version January 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3046607

Topic-Document Inference With the Gumbel-Softmax Distribution

AMIT KUMAR[1,2], **(Graduate Student Member, IEEE), NAZANIN ESMAILI**[1], **AND MASSIMO PICCARDI**[1], **(Senior Member, IEEE)**

[1]School of Electrical and Data Engineering, FEIT, University of Technology Sydney, Sydney, NSW 2007, Australia
[2]Food Agility CRC Ltd., Ultimo, NSW 2007, Australia

Corresponding author: Amit Kumar (amit270980@gmail.com)

This work was supported in part by the Food Agility CRC Ltd., through the Commonwealth Government CRC Program.

ABSTRACT Topic modeling is an important application of natural language processing (NLP) that can automatically identify the set of main topics of a given, typically large, collection of documents. In addition to identifying the main topics in the given collection, topic modeling infers which combination of topics is addressed by each individual document (the so-called topic-document inference), which can be useful for their classification and organization. However, the distributional assumptions for this inference are typically restricted to the Dirichlet family which can limit the performance of the model. For this reason, in this paper we propose modeling the topic-document inference with the Gumbel-Softmax distribution, a distribution recently introduced to expand differentiability in deep networks. To set up a performing system, the proposed approach integrates Gumbel-Softmax topic-document inference in a state-of-the-art topic model based on a deep variational autoencoder. Experimental results over two probing datasets show that the proposed approach has been able to outperform the original deep variational autoencoder and other popular topic models in terms of test-set perplexity and two topic coherence measures.

INDEX TERMS Topic models, topic-document inference, variational autoencoders, Gumbel-Softmax distribution, deep neural networks.

I. INTRODUCTION

Unstructured textual data are growing by the day in the form of news, press releases, blogs, social media posts and others. The possibility for humans to annotate such documents is limited since manual annotation is labor-intensive and time-consuming. Therefore, there is an urgent and widespread need for automated, unsupervised analysis tools that can provide an understanding of such data and work at scale [7].

Topic modeling is an unsupervised, probabilistic approach of natural language processing (NLP) that is capable of discovering the main topics of large amounts of unstructured text, and presenting them to a user in succinct and comprehensible forms. It has established a strong reputation as a useful text analytics technique and has found application in fields ranging from business and finance to healthcare and scientific corpora analysis [2], [4], [20], [21], [27], [29], [32]. In topic modeling, a topic is typically represented by the set of its most-frequent words. For instance, a topic such as "cricket" may be represented by words such as "innings", "stump", "wicket" and all the other typical terminology of cricket commentaries. As a more sobering example, a topic such as "pandemic" may be represented by words such as "infection", "intensive care", "death", "recovery" and so forth. In more general terms, a topic can be seen as a probability distribution over the words of an available vocabulary, where the words that are distinctive for that topic are characterized by the highest probabilities.

Topic modeling is able to parse a whole corpus of documents and identify the most common topics "shared" by these documents. Simultaneously, it is able to determine what proportion of topics is addressed by each individual document. The existing approaches for topic modeling are predominantly based on non-negative matrix factorization and probabilistic inference, and the most famous is undoubtedly the latent Dirichlet allocation (LDA) of Blei, Ng and Jordan [3]. In this approach and many of its derivatives, the topic proportions of the individual documents are modeled using the Dirichlet distribution which is a convenient conjugate prior for the topic frequencies. However, limiting

The associate editor coordinating the review of this manuscript and approving it for publication was Xin Luo.

the models to this assumption may be restrictive, since other distributions over the topic proportions may be able to achieve better performance figures for the derived topic models.

For this reason, in this paper we propose modeling the topic proportions of the individual documents using the Gumbel-Softmax distribution [9], [18]. This distribution has been recently introduced to expand the applicability of backpropagation in deep learning models with latent categorical variables, where it is used to replace non-differentiable, categorical samples with ''soft'' samples from a differentiable transformation. The main expected advantage of using this distribution for topic modeling is that it can effectively control the sparsity of its samples by a pseudo-temperature hyperparameter, and can thus be able to control the expected number of topics of each individual document during the so-called topic-document inference. To set up a performing system, we have integrated this distribution into the sampling step of a state-of-the-art topic model, the autoencoding variational inference for topic models (AVITM) of Srivastava and Sutton [28].

Experiments have been carried out on two challenging text datasets: the popular 20 Newsgroups dataset [14], consisting of 18,846 user-posted documents from newsgroups, and the recent, large-scale COVID-19 news dataset,[1] aggregated by AYLIEN using their news API on more than 400 different sources. The experimental results show that the proposed topic-document inference approach has been able to achieve higher topic coherence and lower perplexity than all the other compared approaches.

The rest of this paper is organized as follows: Section II presents the related work. Section III presents the proposed model, preceding it with a concise review of LDA and a state-of-the-art variational topic model. Section IV describes the experiments, and presents and discusses the results. Section V concludes the paper.

II. RELATED WORK

Topic modeling is unarguably one of the most researched areas of natural language processing. Its aim is to find concise descriptors for a typically-large ($> 10,000$ documents) given corpus and for its individual documents. This is generally achieved by introducing a set of latent variables, known as the ''topics'', which are shared across the corpus and describe it, while simultaneously determining the proportions of the topics in each document. The input to topic modeling is typically a simplified representation of the documents in the corpus known as the term-document matrix. Topic modeling has found application in a large number of areas including news [29], social media [1], [21] finance [4], [21], healthcare [2], [27], [32] and many others.

Among the many techniques proposed over the years, latent semantic indexing (LSI, also known as latent semantic analysis, or LSA) is credited as the first explicit topic model [5]. It consists of the factorization of the term-document matrix in a low-rank latent space by means of a singular value decomposition. To more clearly explain this factorization, which will also be useful for the remainder of the paper, let us introduce the following notations: V is the size of the given vocabulary, D is the number of documents in the given corpus, K is the number of topics chosen to describe the corpus, and W is the term-document matrix, of $V \times D$ size. The LSI factorization can then be expressed as:

$$W \approx \beta\theta \tag{1}$$

where β is a $V \times K$ matrix usually referred to as the term-topic matrix, and θ is a $K \times D$ matrix referred to as the topic-document matrix. The values for β and θ can be obtained by applying singular value decomposition to W, and incorporating the resulting eigenvalues into either of the other two factors. This ensures that $\beta\theta$ is the best possible approximation of W in a least-square sense. For this factorization to be of any practical utility, the chosen number of topics, K, must satisfy $K \ll D$. However, since K is typically chosen in a range such as [20, 100] and the corpora are large, this condition is always easily met. Among various uses, the LSI factorization can be used to compare, cluster and classify documents (e.g. [10]); to extract the top words of each topic; and even to compare and cluster words.

Probabilistic latent semantic analysis (pLSA, or, analogously, pLSI) [8] has overlaid a probabilistic interpretation to the LSI factorization: the first factor, the term-topic matrix, is interpreted as the probability of a word, w, in a given topic, t, while the second factor, the topic-document matrix, is interpreted as the probability of a topic, t, in a given document, d. Both probabilities are modeled as multinomial distributions. The computation of the factorization is similar to that of LSI, but the elements of the factor matrices must all belong to interval [0, 1], and the relevant columns and rows must abide by a sum-to-one constraint (the simplex domain). The multinomial distributions of the term-topic matrix, $p(w|t)$, are concisely called the ''topics'', as they express how probable it is that any of the words in the given vocabulary will appear in text from a given topic. The multinomial distributions in the topic-document matrix, $p(t|d)$, are called the ''topic vectors'' and express the mixture of topics covered by a given document. A highly popular generalization of pLSA called latent Dirichlet allocation (LDA) adds prior probabilities to both the topics and the topic vectors in the form of Dirichlet distributions [3]. Since the Dirichlet distribution is conjugate to the multinomial, the posterior probabilities can be computed analytically, allowing for efficient inference algorithms. We review LDA in detail in Section III-A. LDA has also spawned a large number of extensions and variants, including hierarchical versions [11], [17], sequential versions [26], class-supervised versions [26], sparse versions [22], [30], [34], and many others.

In recent years, neural topic models have come into the spotlight by combining the advantages of deep neural networks and LDA. Deep models based on variational autoencoders (VAEs) such as [13], [19], [28], [31] have proved

[1] https://aylien.com/resources/datasets/coronavirus-dataset

effective at automatic discovery of the latent topics in the corpus, and deep models based on CNNs have been used for topic-based document classification and non-negative matrix factorization [16], [33]. Recently, Srivastava and Sutton [28] have proposed a topic model that joins the properties of LDA with the strong representational power of a deep variational autoencoder. This approach has proved to clearly outperform LDA both quantitatively and qualitatively, and can be regarded as one of the current state-of the-art approaches. In addition, various deep topic models have been proposed based on generative adversarial networks (GANs). Among them, [6] uses a denoising autoencoder to implement the discriminator network, under the expectation that the discriminator should achieve a small reconstruction error on the documents in the corpus, while a large reconstruction error on the synthetic documents generated by the generator network. The main aim of this GAN-based topic model is to provide effective topic vectors for document classification [6]. However, it can also be used for extracting the top words of the topics, and vector representations for the words.

III. AUTOENCODING VARIATIONAL INFERENCE FOR TOPIC MODELS

In this section, we present the proposed methodology, preceded by an overview of latent Dirichlet allocation and variational autoencoders for topic modeling.

A. LATENT DIRICHLET ALLOCATION

Latent Dirichlet allocation (LDA), proposed by Blei, Ng and Jordan in 2003 [3], is probably the reference model for the field of topic modeling. To briefly describe it hereafter, let us introduce the following notations:

- $w_{d,n}$ is the n-th word in the d-th document in the corpus. By "word" we mean a categorical value in the corpus' vocabulary (essentially, an index). The size of the vocabulary is noted as V. Wherever unambiguous, we omit the document index for brevity.
- w_d is the set of all the words in document d (again, where possible, we omit the document index).
- Each word, $w_{d,n}$, is assigned to a corresponding topic, $z_{d,n}$. A topic, too, is a categorical variable taking values in a set of $1 \ldots K$ possible values (NB: the topics are "nameless", but can be later assigned meaningful names with a post-analysis). This correspondence means that, for example, a word such as "bat" can be assigned to topic "mammals" in one instance and "cricket" in another.

The model makes the following distributional assumptions:

- The topic variables for a given document are independently and identically distributed according to a multinomial distribution, $\text{Mult}(z_{d,n}|\theta_d)$, parametrized by a K-dimensional probability vector, θ_d.
- At its turn, vector θ_d is distributed according to a Dirichlet distribution, $\text{Dir}(\theta_d|\alpha)$, parametrized by a K-dimensional integer vector, α, shared by the whole corpus. (The conjugacy between the multinomial and Dirichlet eases the computation of the required posteriors.)
- The words in the corpus are distributed according to a set of K multinomial distributions, parametrized by K corresponding V-dimensional probability vectors, $\beta = \beta_1, \ldots \beta_K$. Each word in a given document is independently distributed according to one of these distributions, indexed by its topic variable, as in $\text{Mult}(w_{d,n}|\beta_{z_{d,n}})$.

All these assumptions can be concisely noted in a "generative" model, that is a model that allows sampling an entire synthetic corpus from these distributions:

$$\begin{aligned} \forall d &= 1 \ldots D : \\ \theta_d &\sim \text{Dir}(\theta|\alpha) \\ \forall n &= 1 \ldots N : \\ z_n &\sim \text{Mult}(z_n|\theta_d) \\ w_n &\sim \text{Mult}(w_n|\beta_{z_n}) \end{aligned} \tag{2}$$

which also corresponds to the following factorization:

$$\begin{aligned} &p(w_n, z_n, \theta_d|\alpha, \beta) \\ &\qquad = \text{Mult}(w_n|\beta_{z_n})\text{Mult}(z_n|\theta_d)\text{Dir}(\theta_d|\alpha) \end{aligned} \tag{3}$$

Since both w_n and z_n are multinomially distributed, it is also possible to dispose of z_n altogether by marginalizing it analytically. In this case, the generative model simplifies to:

$$\begin{aligned} \forall d &= 1 \ldots D : \\ \theta_d &\sim \text{Dir}(\theta|\alpha) \\ \forall n &= 1 \ldots N : \\ w_n &\sim \text{Mult}(w_n|\beta\theta_d) \end{aligned} \tag{4}$$

where with $\beta\theta_d$ we have noted the product between $V \times K$ matrix β and $K \times 1$ vector θ_d. The corresponding factorized probability is:

$$p(w_n, \theta_d|\alpha, \beta) = \text{Mult}(w_n|\beta\theta_d)\text{Dir}(\theta_d|\alpha) \tag{5}$$

and the probability for all the words in a document can be simply expressed as:

$$p(w, \theta_d|\alpha, \beta) = \prod_{n=1}^{N} p(w_n, \theta_d|\alpha, \beta) \tag{6}$$

The inference problem for this model consists of maximizing (6) by estimating θ_d, β and α over a given training corpus of documents. In essence, answering these questions: what is the distribution of words in each of these topics? ($\beta = \beta_1, \ldots \beta_K$); what are the proportions of the topics in each of these documents? ($\theta = \theta_1, \ldots \theta_D$); and what are the proportions of the topics across the whole corpus? (α). For new/test documents given after training is complete, β and α are kept unchanged and only their topic vectors are inferred.

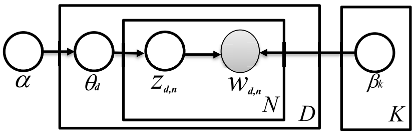

FIGURE 1. The graphical model of LDA. The meaning of the notations is as follows: α denotes the parameter vector for the Dirichlet prior over the topic vectors (i.e. the topic proportions per document), unique for the corpus. θ_d is the topic vector of the d-th document, sampled from Dir($\theta_d|\alpha$). For each document, N topics, $z_{d,n}$, are then sampled from Mult($z_{d,n}|\theta_d$). Finally, the corresponding N words, $w_{d,n}$ are sampled from a multinomial distribution over the vocabulary, Mult($w_{d,n}|\beta_{z_{d,n}}$); its parameter vector, $\beta_{z_{d,n}}$, is chosen from a set of K parameter vectors, $\beta = \{\beta_1, \ldots \beta_k \ldots \beta_K\}$, based on the value of topic $z_{d,n}$.

B. VARIATIONAL AUTOENCODERS FOR TOPIC MODELING

Since the ascendance of deep learning, a fresh wave of models best known as deep generative models (DGM) have come into existence, fundamentally a blend of deep neural nets, generative models and Bayesian inference. Among them, variational autoencoders (VAEs) have proved very effective for models that contain latent variables (in our case, the topics) [12]. VAEs are able to efficiently maximize the log-likelihood of the observed data even when this function is not directly optimizable, making them widely applicable in all fields of big data including, among others, signal processing, computer vision, natural language processing and transactional data analytics.

A VAE is essentially a generalization of a traditional autoencoder, which is a neural network consisting of two subnetworks: an encoder and a decoder. The encoder receives a multidimensional measurement in input, and outputs a latent representation for it; the decoder receives the latent representation in input, and outputs a "reconstruction" of the original measurement. Through this process, the model is able to generate latent representations and reconstructed measurements which are often more useful than the original measurements in downstream tasks of pattern recognition (e.g. [23]).

A variational autoencoder is a probabilistic extension of an autoencoder where both the measurement and the latent representation are treated as random variables, and therefore the encoder and the decoder are treated as probability distributions. The "reconstruction" of the original measurement is meant in a probabilistic manner in terms of log-likelihood maximization. In the case of our topic model, the aim of the VAE is to maximize the log-likelihood of the words of each document:

$$p(w|\alpha, \beta) = \int_\theta p(w, \theta|\alpha, \beta) d\theta \tag{7}$$

However, the above objective is too complex to be maximized directly, and therefore the VAE establishes an approachable lower bound for the log-likelihood known as the Evidence Lower Bound, or ELBO, and sets to maximize it [12]. In the case of the topic model, the ELBO has the following form:

$$\mathcal{L}(w|\alpha, \beta) = \mathbb{E}_{q(\theta|w)}\big[\log p(w|\theta, \beta)\big] \\ -D_{\mathrm{KL}}(q(\theta|w) \| p(\theta|\alpha)) \tag{8}$$

The terms in (8) have the following meaning: 1) $q(\theta|w)$ is an estimator for the probability of the topic proportions for a given document (represented by its words, w) and is known as the "encoder"; 2) $\log p(w|\theta, \beta)$ is the log-probability of the given document given its topic proportions and is known as the "decoder"; 3) $\mathbb{E}_{q(\theta|w)}\big[\log p(w|\theta, \beta)\big]$ is the expectation of this quantity over $q(\theta|w)$ and is known as the "reconstruction term"; 4) $p(\theta|\alpha)$ is a learnable prior probability for the topic proportions that is shared by the entire corpus. The rationale for (8) is twofold: first, it is a proven lower bound for (7), that is the target of the maximization; second, it consists of a trade-off between two terms that can be interpreted intuitively: the model is rewarded for either improving the reconstruction term, or for keeping the encoder close to the prior.

Srivastava and Sutton in [28] have proposed a VAE for topic modeling (AVITM) that leverages a Laplace approximation of the usual Dirichlet prior to permit its integration into the autoencoder. In AVITM, both the prior and the encoder are modeled as logistic normal distributions: the prior is modeled as $p(\theta|\alpha) = \mathcal{LN}(\theta|\mu(\alpha), \Sigma(\alpha))$, and the encoder is modeled as $q(\theta|w) = \mathcal{LN}(\theta|f_\mu(\phi, w), f_\Sigma(\phi, w))$, where ϕ are the internal parameters of two neural networks that predict the mean and covariance of the encoder, respectively. The expectation in (8) is computed by sampling $q(\theta|w)$, which in turn is performed through reparametrization. The decoder takes the following form:

$$p(w|\theta, \beta) = \mathrm{Mult}(w|\sigma(\beta)\theta) \tag{9}$$

where $\sigma()$ is the softmax operator and the word distributions are parametrized in the softmax basis rather than the simplex to remove unnecessary constraints during backpropagation. The authors have also proposed a second, heuristic version of the decoder, called ProdLDA, that performs the product before the softmax:

$$p(w|\theta, \beta) = \mathrm{Mult}(w|\sigma(\beta\theta)) \tag{10}$$

As shown in [28], both AVITM and ProdLDA have outperformed a number of compared topic model approaches by large margins, and can be regarded as state-of-the-art approaches for this task.

C. THE PROPOSED APPROACH: VAE TOPIC MODELS WITH THE GUMBEL-SOFTMAX

The Gumbel-Softmax distribution, co-credited to [18] and [9], has channeled much attention from the deep learning community in recent years. This distribution models "soft" categorical variables (categorical variables that are not restricted to have one-hot values) and has been introduced to circumvent issues related to backpropagation in models with latent categorical variables. Many deep learning models (prominently, variational autoencoders and generative adversarial networks, or GANs) need to sample from distributions, and sampling is a non-differentiable operation that breaks the backpropagation chain. The Gumbel-Softmax distribution is an alternative to the multinomial distribution that allows sampling of quasi-categorical variables and is differentiable via reparametrization. Given a multinomial distribution, $\mathrm{Mult}(z|\theta)$, with K possible values, samples from the corresponding Gumbel-Softmax distribution, $\mathcal{GS}(\tilde{z}|\theta, \tau)$, can be obtained as:

$$\tilde{z} = \sigma\big([\log\theta - \log(-\log u)]/\tau\big)$$
$$u \sim \mathcal{U}(0, 1)^K \tag{11}$$

where u is a vector of K random variables each sampled from the uniform distribution over (0, 1), and τ is a hyperparameter (referred to as "temperature") that controls the sparsity of $\tilde{z}$ (the lower τ, the more the samples resembles one-hot values; the larger, the more the samples become uniform). Note that the sampled distribution is fixed and does not need gradient updates, and the functions in (11) are all differentiable.

To take advantage of its properties, we propose sampling the topic vector from a Gumbel-Softmax distribution. The modified decoder (nicknamed *AVITM-GS*) becomes:

$$p(w|\theta, \beta) = \mathrm{Mult}(w|\sigma(\beta)\tilde{z}), \quad \tilde{z} \sim \mathcal{GS}(\theta, \tau) \tag{12}$$

and in the case of ProdLDA (*ProdLDA-GS*) it becomes:

$$p(w|\theta, \beta) = \mathrm{Mult}(w|\sigma(\beta\tilde{z}), \quad \tilde{z} \sim \mathcal{GS}(\theta, \tau) \tag{13}$$

Please note that the number of trainable parameters is the same as in the original decoders, with the exception of the scalar hyperparameter τ that we can use to control the sparsity of the inferred topic vectors.

IV. EXPERIMENTS AND RESULTS

A. DATASETS

As datasets for the experiments, we have used the popular 20 Newsgroups dataset (a de-facto benchmark for the field) and a 500K-document subset of AYLIEN's recently released COVID-19 news dataset. 20 Newsgroups consists of 18,846 news documents posted by users, split over 11,314 as training set and 7,532 as test set. The average length of these documents is 311 words. To be consistent with the experiments carried out in [28], we have used the preprocessed version publicly released by the authors[2] which uses a vocabulary of 1,995 words. The COVID-19 news dataset is a dataset aggregated by AYLIEN using their News Intelligence Platform from November 2019 to July 2020 from approximately 440 different sources. For our experiments, we have used the first 500K documents (over 7 GB of uncompressed text) split over 400K as training set and 100K as test set since this size could still be managed by a PC with 16 GB of RAM. The documents were preprocessed with tokenization, stopword elimination, stemming and lemmatization, and encoded with a vocabulary formed by the most-frequent 5,000 unique words.

B. EXPERIMENTAL SET-UP

To probe the comparative performance of the proposed approach, we have integrated it in both AVITM and ProdLDA, and compared these versions with the original versions. In the following, we refer to them as AVITM-GS and ProdLDA-GS, respectively. We have also included LDA and LSI from Gensim [24] in the comparison as baselines, and the GAN-based topic model from [6] that we refer to as GANTM in the following. As learning rate for the variational autoencoders, we have used the rather standard value of 0.001. Any other hyperparameters were left to their default values. For the temperature of the Gumbel-Softmax distribution, τ, we have carried out a preliminary sensitivity analysis and chosen to run experiments with $\tau \in [1.5-2.5]$ in steps of 0.25. This range corresponds to moderately-sparse to dense topic vectors. As number of topics, we have used both 50 and 100 topics for both datasets. We have also initially carried out multiple runs per model, and realised that the performance did not vary significantly ($< 0.5\%$ in all cases). Therefore, in Section IV-C we report results from single runs of each model.

As an unsupervised technique, the performance evaluation of a topic model is non-trivial. For our work, we have used two common measures:

- *perplexity over the test set*: the perplexity of a model over a set S is defined as: $\mathrm{perplexity}(S) = \exp(-\mathcal{L}(S)/(\text{number of tokens in } S))$. In the general case, $\mathcal{L}$ denotes the log-likelihood of the data, but for the variational methods (all except LSI and GANTM in our case), it is given by the ELBO in (8). The perplexity is a measure of the "poorness of fit" of the model on the data (the lower, the better) and, as such, it is important that it is measured over an independent test set for realistic generalization.
- *topic coherence*: topic coherence quantifies the coherence of a topic by measuring how often its top K words co-occur within a text window that slides across the documents (the higher the co-occurrence, the better). Since this measure is not uniquely defined, we report both the normalized pointwise mutual information (`coher-NMPI`) [15] and the C_V coherence (`coher-Cv`) [25] from their Gensim implementation. The coherence is typically measured on the training set itself since this guarantees the presence of all the top words. For the experiments, K has been set to 10. For the variational methods, the top words per topic have been selected as those with highest probability in the term-topic matrix. For LSI, they have been selected as those with highest weight in the term-topic matrix (which is not normalized to probability values). For GANTM, they have been selected as those with highest weight in the discriminator's decoder network (equivalent to the term-topic matrix of LSI).

[2] Available at: https://github.com/akashgit/autoencoding_vi_for_topic_models.

TABLE 1. Results with 50 topics on 20 Newsgroups.

Measure/Model	LDA	LSI	GANTM	ProdLDA	AVITM	ProdLDA-GS	AVITM-GS
Perplexity	*2389.6*	—	—	1159.9	1133.0	1136.6	**1110.6**
Coher-NPMI	-2.346	-0.062	-0.234	0.111	0.117	**0.148**	0.104
Coher-Cv	-0.053	0.294	0.247	0.751	0.704	**0.806**	0.638

TABLE 2. Results with 100 topics on 20 Newsgroups.

Measure/Model	LDA	LSI	GANTM	ProdLDA	AVITM	ProdLDA-GS	AVITM-GS
Perplexity	*4857.1*	—	—	1147.1	1128.0	1136.1	**1111.4**
Coher-NPMI	-0.063	-0.071	-0.223	0.114	0.085	**0.117**	0.079
Coher-Cv	0.296	0.267	0.259	0.742	0.650	**0.763**	0.616

TABLE 3. Results with 50 topics on COVID-19.

Measure/Model	LDA	LSI	ProdLDA	AVITM	ProdLDA-GS	AVITM-GS
Perplexity	*1130.0*	—	2178.7	1909.0	1957.7	**1850.5**
Coher-NPMI	0.086	-0.008	0.076	**0.180**	0.170	0.175
Coher-Cv	0.589	0.310	0.682	0.760	**0.787**	0.744

TABLE 4. Results with 100 topics on COVID-19.

Measure/Model	LDA	LSI	ProdLDA	AVITM	ProdLDA-GS	AVITM-GS
Perplexity	*1119.2*	—	2251.7	1904.3	**1855.2**	1855.7
Coher-NPMI	0.090	-0.017	0.049	**0.177**	0.174	0.158
Coher-Cv	0.581	0.271	0.652	0.736	**0.765**	0.700

Given their significantly different nature, some disagreement in model ranking between perplexity and topic coherence is to be expected. Perplexity is, essentially, a measure of fit of the model, while topic coherence is a measure of quality of the extracted topics and may better reflect the user's perception of performance. For this reason, for comparing the models we resort to a majority criterion, with emphasis on the topic coherence.

C. RESULTS

Tables 1 and 2 report the results over the 20 Newsgroups dataset for 50 and 100 topics, respectively. In terms of test-set perplexity, it is evident that the proposed approach has been able to improve over the original variational autoencoder, both for ProdLDA and AVITM. In these and the following tables, we report the perplexity also for LDA, but the scale of its ELBO is not directly comparable with that of the autoencoder techniques; for this reason, its values are marked in italics and not commented further. For LSI and GANTM, the perplexity is simply not available since they are not probabilistic models. In terms of coherence, ProdLDA-GS has been able to achieve significantly higher values than all the other techniques in both coherence metrics. In addition, the two topic model baselines, LDA and LSI, and the GANTM model have scored significantly lower values of topic coherence than all the variational autoencoder approaches. Overall, ProdLDA-GS has achieved the best performance in 4 cases out of 6 (combined number of topics/metrics) and can be regarded as the best-performing technique overall.

TABLE 5. Results for ProdLDA-GS (50 topics, 20 Newsgroups) with varying temperature hyperparameter, τ.

Measure/τ	10^{-5}	1.5	1.75	2.0	2.25	2.5	10
Perplexity	1131.7	1180.0	1161.1	1145.4	1136.6	1124.7	**1099.7**
Coher-NPMI	-0.224	0.126	0.131	0.125	**0.148**	**0.148**	0.010
Coher-Cv	NaN	0.788	0.780	0.785	**0.806**	0.799	0.638

In turn, Tables 3 and 4 report the results over the COVID-19 dataset for 50 and 100 topics, respectively. In terms of test-set perplexity, the proposed approach has again been able to improve over the original variational autoencoders. In terms of coherence, the original AVITM has achieved the highest values for `coher-NPMI`, while ProdLDA-GS has achieved the highest values for `coher-Cv`. Again, all the variational autoencoder approaches have scored significantly higher coherence values than both the LDA and LSI baselines. GANTM generated an out-of-memory error while training over larger training sets, and is therefore not reported. Overall, ProdLDA-GS has achieved the best performance in 3 cases out of 6 and may still be regarded as the best-performing overall.

As expected, the choice of the temperature hyperparameter, τ, in the Gumbel-Softmax distribution has a major impact on the performance as it substantially changes the shape of the samples (from almost one-hot to almost uniform). Since the coherence measures are to be computed on the training set, it is legitimate to choose the value of τ that empirically maximizes them. Conversely, the perplexity is a test-set measure and the optimal τ should be chosen on the training set or a separate validation set. In all cases, the different measures may be maximized by different values of τ, and a trade-off between them is required. To illustrate this dependence, Table 5 shows the results with varying τ for ProdLDA-GS with 50 topics on 20 Newsgroups. With

TABLE 6. Examples of topics extracted from the COVID-19 dataset (50 topics).

LDA:	itali countri franc europ european spain italian germani measur lockdown new york citi state cuomo san governor mayor francisco andrew south korea japan africa countri north tokyo korean japanes brazil
ProdLDA:	rub sampl mer nasal patient symptom cough lung genet molecular diamond passeng disembark repatri dock princess liner hubei aboard cruis trophi leagu europa juventus hudson champion coach footbal munich albert
ProdLDA-GS:	symptom cough respiratori patient ill nose hospit infect doctor sneez democrat biden sander republican trump voter vote senat sen nomin crude barrel oil opec investor output price brent bpd yield

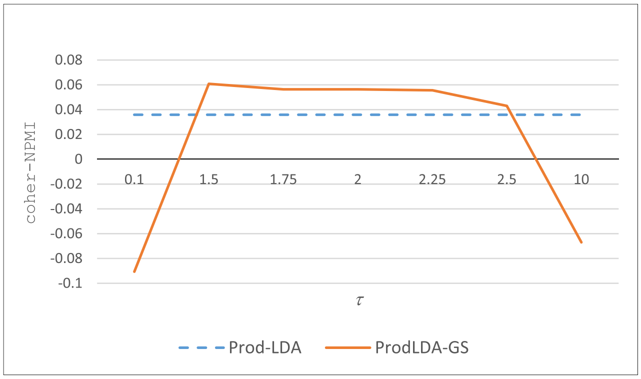

FIGURE 2. Comparison of `coher-NPMI` on the test set for ProdLDA and ProdLDA-GS (50 topics, 20 Newsgroups) with varying temperature hyperparameter, τ.

$\tau = 10^{-5}$ (almost one-hot samples), the model has achieved a very low coherence. At the other end of the spectrum, with $\tau = 10$ (almost uniform samples), the coherence has been again very low. The equal-best `coher-NPMI` coherence values have been achieved with $\tau = 2.25$ and 2.5, and the best value for `coher-Cv` has been achieved with $\tau = 2.25$, so we have used these results for the comparison in Table 1. To further evaluate the model's quality with varying τ, we have also measured the topic coherence (`coher-NPMI`) of ProdLDA-GS over the test set, using ProdLDA as the reference. Figure 2 shows that τ has played a key role also for this measure: for $\tau \in [1.5 - 2.5]$, the topic coherence of ProdLDA-GS has been invariably higher than that of Prod-LDA, while it has noticeably deteriorated for more "extreme" values (0.1, 10).

In terms of qualitative analysis of the extracted topics, all approaches seem to have performed well overall. The extracted topics are presented to the user as the lists of their $K = 10$ top words, and such lists must appear informative and coherent. Examples for LDA, ProdLDA and ProdLDA-GS from the COVID-19 topic models are displayed in Table 6. For LDA, the first example clearly addresses the lockdown measures taken by various European countries; the second names New York State Governor Andrew Cuomo and the mayor of San Francisco, but fails to include the "reason" for their mention; and the last is simply a list of countries, again with no explicit mention of the COVID outbreak. For ProdLDA, the first example refers to COVID symptoms and testing (word "mer" is the stemmed version of "MERS"); the second refers to the case of the Diamond Princess cruise ship; and the last addresses football news from the observation period. For ProdLDA-GS, the first example clearly refers to COVID symptoms and the risk of infection for the doctors; the second to the recent US presidential primaries, which were held during the observation period; and the last to economic news. Their lists of top words seem very consistent and descriptive. A possible limitation of both ProdLDA and ProdLDA-GS, and possibly of all autoencoding methods which are based on sampling, is the presence of a number of repeated topics. However, it should be easy to prune them post-hoc.

V. CONCLUSION

This paper has presented an approach for topic modeling based on the Gumbel-Softmax distribution and variational autoencoders. During the step of topic-document inference, the topic proportions of the current document are sampled in the autoencoder from a Gumbel-Softmax distribution with appropriate temperature. The samples are then used to mix either the topic distributions (AVITM-GS) or their logits (ProdLDA-GS). To validate the proposed approach, experiments have been carried out on two challenging datasets, the well-known 20 Newsgroups and a recently-released, large-scale COVID-19 news dataset. The experimental results have shown that the proposed approach has been able to outperform the original variational autoencoders and two significant baselines in terms of topic coherence, and achieve the best trade-off across two coherence metrics and the test-set perplexity. In addition, a qualitative analysis of the extracted topics has shown that they appear informative and consistent. In the near future, we plan to extend our research to other distributional models and reparametrization approaches.

ACKNOWLEDGMENT

The CRC Program supports industry-led collaborations between industry, researchers, and the community.

REFERENCES

[1] D. Alvarez-Melis and M. Saveski, "Topic modeling in Twitter: Aggregating tweets by conversations," in *Proc. 10th Int. Conf. Web Social Media*. AAAI Press, 2016, pp. 519–522.

[2] C. Arnold, S. El-Saden, A. Bui, and R. Taira, "Clinical case-based retrieval using latent topic analysis," in *Proc. AMIA Annu. Symp.*, 2010, pp. 26–30, 2010.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[4] M. Cecchini, H. Aytug, G. J. Koehler, and P. Pathak, "Making words work: Using financial text as a predictor of financial events," *Decis. Support Syst.*, vol. 50, no. 1, pp. 164–175, Dec. 2010.

[5] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *J. Amer. Soc. Inf. Sci.*, vol. 41, no. 6, pp. 391–407, 1990.

[6] J. Glover, "Modeling documents with Generative Adversarial Networks," in *Proc. NIPS Workshop Adversarial Training*, 2016, pp. 1–7.

[7] J. David Hand, "Text mining: Classification, clustering, and applications," *Int. Stat. Rev.*, vol. 78, pp. 134–135, Jan. 2010.

[8] T. Hofmann, "Probabilistic latent semantic analysis," in *Proc. 15th Conf. Uncertainty Artif. Intell. (UAI)*, 1999, pp. 289–296.

[9] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with Gumbel-Softmax," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–12.

[10] X. Kang, F. Ren, and Y. Wu, "Exploring latent semantic information for textual emotion recognition in blog articles," *IEEE/CAA J. Automatica Sinica*, vol. 5, no. 1, pp. 204–216, Jan. 2018.

[11] H. Kim, B. Drake, A. Endert, and H. Park, "Architext: Interactive hierarchical topic modeling," *IEEE Trans. Visualizat. Comput. Graph.*, to be published.

[12] P. Diederik Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–14.

[13] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Found. Trends Mach. Learn.*, vol. 12, no. 4, pp. 307–392, 2019.

[14] K. Lang, "Newsweeder: Learning to filter netnews," in *Proc. 12th Int. Conf. Mach. Learn. (ICML)*, 1995, pp. 331–339.

[15] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2014, pp. 530–539.

[16] Q. Lian, W. Yan, X. Zhang, and S. Chen, "Single image rain removal using image decomposition and a dense network," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 6, pp. 1428–1437, Nov. 2019.

[17] T. Liu, N. L. Zhang, and P. Chen, "Hierarchical latent tree analysis for topic detection," *CoRR*, vol. 8725, pp. 256–272, 2014.

[18] C. Maddison, A. Mnih, and Y. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," 2016, *arXiv:1611.00712*. [Online]. Available: https://arxiv.org/abs/1611.00712

[19] Y. Miao, E. Grefenstette, and P. Blunsom, "Discovering discrete latent topics with neural variational inference," in *Proc. 34th Int. Conf. Mach. Learn. (ICML)*, 2017, pp. 2410–2419.

[20] A. Murakami, P. Thompson, S. Hunston, and D. Vajn, "'What is this corpus about?': Using topic modelling to explore a specialised corpus," *Corpora*, vol. 12, no. 2, pp. 243–277, Aug. 2017.

[21] T. H. Nguyen and K. Shirai, "Topic modeling based sentiment analysis on social media for stock market prediction," in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Lang. Process. (Long Papers)*, vol. 1, 2015, pp. 1354–1364.

[22] M. Peng, Q. Xie, Y. Zhang, H. Wang, X. Zhang, J. Huang, and G. Tian, "Neural sparse topical coding," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics (Long Papers)*, vol. 1, 2018, pp. 2332–2340.

[23] E. Principi, D. Rossetti, S. Squartini, and F. Piazza, "Unsupervised electric motor fault detection by using deep autoencoders," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 2, pp. 441–451, Mar. 2019.

[24] R. Rehurek and P. Sojka, "Software framework for topic modelling with large corpora," in *Proc. LREC Workshop New Challenges NLP Frameworks*, 2010, pp. 45–50.

[25] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proc. 8th ACM Int. Conf. Web Search Data Mining WSDM*, 2015, pp. 399–408.

[26] F. Rodrigues, M. Lourenco, B. Ribeiro, and F. C. Pereira, "Learning supervised topic models for classification and regression from crowds," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2409–2422, Dec. 2017.

[27] E. Sarioglu, H.-A. Choi, and K. Yadav, "Clinical report classification using natural language processing and topic modeling," in *Proc. 11th Int. Conf. Mach. Learn. Appl.*, Dec. 2012, pp. 204–209.

[28] A. Srivastava and A. Charles Sutton, "Autoencoding variational inference for topic models," in *Proc. 5th Int. Conf. Learn. Represent., (ICLR)*, 2017, pp. 1–12.

[29] G. Xu, Y. Meng, Z. Chen, X. Qiu, C. Wang, and H. Yao, "Research on topic detection and tracking for online news texts," *IEEE Access*, vol. 7, pp. 58407–58418, 2019.

[30] A. Zhang, J. Zhu, and B. Zhang, "Sparse online topic models," in *Proc. 22nd Int. Conf. World Wide Web WWW*, 2013, pp. 1489–1500.

[31] H. Zhang, B. Chen, Y. Cong, D. Guo, H. Liu, and M. Zhou, "Deep autoencoding topic model with scalable hybrid Bayesian inference," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Jun. 19, 2020, doi: 10.1109/TPAMI.2020.3003660.

[32] R. Zhang, S. Pakhomov, S. Gladding, M. Aylward, E. Borman-Shoap, and G. Melton, "Automated assessment of medical training evaluation text," *AMIA Annu. Symp. Proc.*, vol. 2012, pp. 1459–1468, 2012.

[33] Y. Zhang, B. Xu, and T. Zhao, "Convolutional multi-head self-attention on memory for aspect sentiment classification," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 4, pp. 1038–1044, Jul. 2020.

[34] J. Zhu and P. E. Xing, "Sparse topical coding," in *Proc. 27th Conf. Uncertainty Artif. Intell. (UAI)*, 2011, pp. 831–838.

AMIT KUMAR (Graduate Student Member, IEEE) is currently pursuing the Ph.D. degree in natural language processing (NLP) with the Faculty of Engineering and Information Technology, University of Technology Sydney, Ultimo, NSW, Australia. His research interests include machine learning, deep learning, data science, artificial intelligence, and deep generative models.

NAZANIN ESMAILI received the B.Sc. and M.B.A. degrees from the Sharif University of Technology, and the Ph.D. degree from the University of Pittsburgh. She is currently an Associate with the Faculty of Engineering and Information Technology, University of Technology Sydney. She is also an affiliated Research Associate with the Healthcare Systems Engineering Institute (HSyE), Northeastern University, Boston. She has been a Postdoctoral Research Associate with Northeastern University and a Research Associate with the Global Big Data Technologies Centre, University of Technology Sydney. Her research interests include in the areas of data analytics, mathematical modeling, and machine learning.

MASSIMO PICCARDI (Senior Member, IEEE) received the M.Eng. and Ph.D. degrees from the University of Bologna, Bologna, Italy, in 1991 and 1995, respectively. He is currently a Full Professor in computer systems with the University of Technology Sydney, Ultimo, NSW, Australia. He has coauthored more than 150 articles in these areas. His research interests include natural language processing, computer vision, and pattern recognition. He is a member of the Computer and Systems, Man, and Cybernetics Society and the International Association for Pattern Recognition, and serves as an Associate Editor for the IEEE Transactions on Big Data.

• • •