Title page

Efficient Integration of Generative Topic Models Into Discriminative Classifiers Using Robust Probabilistic Kernels

First author:
Koffi Eddy Ihou
Concordia Institute for Information Systems Engineering,
Concordia University,
Montreal QC H3G 1M8, Canada
E-mail: k_ihou@encs.concordia.ca

Second author ***(corresponding author)***:
Nizar Bouguila
Concordia Institute for Information Systems Engineering,
Concordia University,
Montreal QC H3G 1M8, Canada
E-mail: nizar.bouguila@concordia.ca

Third author:
Wassim Bouachir
Computer science, TÉLUQ University,
Montreal QC H2T 2C8, Canada
E-mail: wassim.bouachir@teluq.ca

Abstract

A direct implementation of supervised topic modeling using a Naive Bayes classifier is mainly characterized by the formulation of robust generative topic models that utilize prior distributions such as Dirichlet in LDA (latent Dirichlet allocation), where the classification ultimately follows the Bayes theorem. Though, in large scale applications, SVM (support vector machine) seems to outperform Naive Bayes. In this paper, we propose a classification framework that combines the flexibility of the generative topic models and the strong performance of the SVM. We therefore present a generative-discriminative collapsed variational Bayes technique for text documents and visual classification. Our collapsed variational Bayes topic model implements simultaneously two different and asymmetric conjugate priors within the same generative process as it specifically draws the document and corpus parameters using both GD (generalized Dirichlet) and BL (Beta-Liouville) distributions. Each of these flexible priors generalizes the Dirichlet in LDA. The proposed hybrid model results in a much improved inference that contributes to more accurate estimates, coherent (topic) generative features, a robust formulation of probabilistic kernels, and a much improved classification rate. Experiments in image and text documents classification show the merits of the proposed approach.

Keywords: generative topic models; discriminative approach; Beta-Liouville; generalized Dirichlet; probabilistic kernels; image and text documents classification

Acknowledgements

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

Noname manuscript No.
(will be inserted by the editor)

Efficient Integration of Generative Topic Models Into Discriminative Classifiers Using Robust Probabilistic Kernels

Received: date / Accepted: date

Abstract We propose an alternative to the generative classifier that usually models both the class conditionals and class priors separately, and then uses the Bayes theorem to compute the posterior distribution of classes given the training set as a decision boundary. Because SVM (support vector machine) is not a probabilistic framework, it is really difficult to implement a direct posterior distribution-based discriminative classifier. As SVM lacks in full Bayesian analysis, we propose a hybrid (generative-discriminative) technique where the generative topic features from a Bayesian learning are fed to the SVM. The standard LDA (latent Dirichlet allocation) topic model with its Dirichlet (Dir) prior could be defined as Dir-Dir topic model to characterize the Dirichlet placed on the document and corpus parameters. With very flexible conjugate priors to the multinomials such as GD (generalized-Dirichlet) and BL (Beta-Liouville) in our proposed approach, we define two new topic models: the BL-GD and GD-BL. We take advantage of the geometric interpretation of our generative topic (latent) models that associate a K-dimensional manifold (K is the size of the topics) embedded into a V-dimensional feature space (word simplex) where V is the vocabulary size. Under this structure, the low dimensional topic simplex (the subspace) defines a document as a single point on its manifold and associates each document with a single probability. The SVM, with its kernel trick, performs on these documents probabilities in classification where it utilizes the maximum marging learning approach as a decision boundary. The key note is that points or documents that are close to each other on the manifold must belong to the same class. Experimental results with text documents and images show the merits of the proposed framework.

Keywords hybrid (generative-discriminative) models · support vector machine · conjugate priors · Beta-Liouville · generalized Dirichlet · probabilistic kernels · documents classification

1 Introduction

Machine learning and AI (artificial intelligence) have been responsible for a wide variety of applications such as object detection and recognition, information retrieval, and natural language understanding and processing. These are very hot topics in the research community. Though, object categorization has always received a particular attention from researchers in the area of computer vision due to the emergence of multimedia datasets (texts, images, videos, sounds, etc) as they are increasingly becoming very complex and difficult to handle. Building models that could fully represent or describe the intrinsic characteristics in these collections of data while allowing easy classification has always been one of the top objectives and challenging tasks in machine learning. In general, object classification can be divided in two main groups in the literature: the generative approach and the discriminative scheme [1].

These two techniques can be formulated as follows: using for instance (for now) the variable Υ as the class label and χ as the observed data in class Υ, the discriminative approach will directly model the posterior distribution $p(\Upsilon/\chi)$ or estimate a function h such that $h(\chi) = \Upsilon$, from the observed data [2,3,1]. On the other hand, generative techniques will model both the prior distribution $p(\Upsilon)$ and the class conditional (likelihood

Address(es) of author(s) should be given

function) $p(\chi/\Upsilon)$ separately, which is equivalent to modeling the joint distribution $p(\chi, \Upsilon)$ before estimating the posterior $p(\Upsilon/\chi)$ of the class given the training set using Bayes theorem as a decision boundary. [4,5,1,3]. A real life analogy to these definitions would be to determine for instance, the type of music someone is currently listening (song). In this scenario, the generative approach will obviously learn about each music type (such as classical, jazz, country, electronic, etc.) before indicating to which type of music this particular song belongs. A discriminative method takes a much simpler and faster approach: it does not learn any of these music types. It will only focus on showing differences between the types of musics (similarities or dissimilarities). Consequently, discriminative techniques do not learn the very details about models of different classes while generative approaches do. Discriminative methods go directly to the point and often do not require lot of computational ressources as in the case of generative schemes. This simplicity and robustness (superior performance) in the discriminative approaches have often attracted many researchers [6,2,3, 1] since their asymptotic error is even lower than the one found in generative approaches [2]. However, generative schemes are still being implemented in many machine learning environments for their usefullness and popularity [6–8,4,5,9–15]. This is because generative approaches (while requiring prior information [16]) learn about the additional details about their models which can be useful in a case of occlusion and missing data. Discriminative techniques on the other hand do not have such flexibility when facing missing data or occlusions. Generative techniques can compute marginals from the joint distributions. This is useful in applications such as outlier detection or novelty detection where the model detects efficiently new data that carry low probability and therefore very difficult to predict accurately [17]. Importantly, during the learning process for instance, generative approaches have ability to handle many (thousands) object categories better than discriminative classifiers [1]. Moreover, following the work in [2], generative schemes have also proved to outperform discriminative methods in a binary classification problem with small number of training samples. For instance, the SVM despite its discriminative power in classification is not a probabilistic approach, and it does not provide posterior distributions. Posterior distributions are important in Bayesian analysis because they provide the tool to make optimal decisions in machine learning (for instance when combining models, minimizing risk, determining a rejection criteria that minimizes misclassification rate, etc. [17]). Therefore, their absence makes it difficult to implement a Bayesian learning in SVM. In contrast, generative schemes benefit from a Bayesian analysis. These characteristics illustrate the strengths and capabilities of each approach. As they carry complementary advantages, it has been suggested to merge the two methods, so that their integration guarantees improvement in performance in automatic object classification. It led to the emergence of hybrid (generative-discriminative) models [6,18–20]. Particularly, for SVM, as today's machine learning techniques carry a strong emphasis on Bayesian paradigm, combining generative models with the SVM classifier remains an essentiel step to allow this classifier to implicitly take advantage of the Bayesian learning. This has been the work of researchers such as [6] who successfully showed the flexibility of the hybrid generative-discriminative with mixtures models where the discriminative classifier is the SVM. The SVM heavily relies on efficient kernel formulation in order to provide robust classification. With the high complexity in the datasets and models, standard kernels such as linear, polynomial, Gaussian RBF (radial basis function) are very restrictive in terms of performance. Furthermore, despite the flexibility of the well-known Fisher kernel [21], it often lacks in preserving the nonlinearity induced by the generative model [22]. This is an example of the necessity to utilize appropriate kernels for better results in the hybrid, generative-discriminative models [6]. The introduction of the Fisher kernel has been immediately followed by the work of other researchers such as [23] and [24] who were able to combine generative features to SVM using the Kullback-Leibler kernel and the TOP kernel derived from Tangent vectors Of Posterior log-odds (TOP), respectively.

It is also noteworthy that recent development in the generative architecture has witnessed the emergence of topic models [25–30] such as LDA (latent Dirichlet Allocation) [31,32]. Originally implemented for text document modeling and analysis within the BoW (bag of words), the LDA topic model is currently dominating the area of computer vision with interesting applications related to image categorization [32], sentiment and behaviour analysis [33], text analysis through the social CQA (Community Answering Questions) platform [34], videos analysis [5], and 3D object modeling [35] for retrieval systems. One of the successes of topic modeling is the introduction of intermediate representations within the bag of words called topics. They are low dimensional subspace representations such that documents are now described as mixtures of topics while topics are defined as distributions over the vocabulary words. This provides a hierarchical description of documents with the observed data. Though, the limitation of the Dirichlet-based topic models due to the Dirichlet

(Dir) prior [14,15,5,4] prompted the use of other flexible priors such as GD and BL. These conjugate priors led to some improvement in generative topic models as they provide robust inferences along with efficient generative processes [14,15,5,4,36]. In addition, the collapsed representation proposed in [37] for batch processing has shown improvement in the generative topic models implementation. However, little work has been done in the literature to connect the generative topic model to the SVM classifier to take advantage of its superior discriminative property based on maximum margin learning as a decision boundary. In the generative stage, the topic features must be generated and then in the discriminative stage, the topics are then fed to the SVM which performs the classification. This constitutes our main objective. The generative stage which learns the topics requires an efficient inference capable of delivering heterogeneous topic features. Though, many probabilistic topic models usually implement standard variational Bayes approaches. Variational Bayes [38,39, 14,15,40], despite their deterministic nature are very limited when it comes to characterize dependency bewtwen topic components, for instance to allow a better compression of the topic features, which is essential for the performance of our SVM classifier. In the generative stage, our proposed approach ultimately implements two robust generative topic models using asymmetric BL and GD in the collapsed space of latent variables. The superiority of the collapsed variational Bayes (CVB) inference in topic modeling is enhanced by the use of these two specific conjugate priors to the multinomials. Normally, using these two priors leads to four topic models: the BL-BL topic model, the GD-GD-topic model, the GD-BL topic model, and finally the BL-GD topic model. The first two topic models here (GD-GD and BL-BL) have been already implemented in our previous work within the CVB inference [4,5,41] and they represent the direct extensions to the Dir based-CVB-LDA [42]. The last two topic models (GD-BL and BL-GD) are the ones that are subjects of implementation in this paper. Importantly, they also carry the CVB inference; and they represent the generative stage in the formation of our hybrid (generative-discriminative) model. As the generative topic features must be fed into the SVM classifier using powerful kernel functions that operate in distribution space, we therefore provide to the SVM, a collection of nonlinear probabilistic kernels (such as Jensen-Shannon kernel, symmetric Kullback-Leibler divergence kernel, Bhattacharyyaa kernel, Renyi kernel, etc.) to cope with data processing in distribution space while allowing an improved classification rate as we induce the space with the CGS (collapsed Gibbs sampler) that operates within the variational Bayesian inference [42]. It samples from the variational distribution in the collapsed space. The CVB corrects the bias in VB due to its CGS and the VB fixes the deterministic limitation of CGS [42]. Due to CVB, our generative topic features are robust, accurate and efficient [42,37,4,5]. The contribution in our proposed hybrid framework is as it follows:

- With CVB inference using asymmetric GD and BL priors simultaneously, we obtained the BL-GD and GD-BL topic models that produce heterogeneous topic features in the generative stage
- SVM is not a probabilistic model; however, we successfully use the kernel trick formulation to make it operate on documents represented as topic features which are probability distributions; SVM now assigns a class label to a previously unseen document based on its topic distribution using its maximum margin framework.

Experimental results in image and text document classification show the efficiency of the proposed approach in comparison to its major competitors.
This paper is structured as follows: section 2 illustrates the background and related work. Section 3 presents the new approach while section 4 covers the experiments and results in several applications. And finally, section 5 emphasizes on some future work and provides a conclusion.

2 Related work and background

In general, low performance in traditional machine learning techniques in applications such as object categorization [43,44,28] have led to the emergence of hybrid models especially generative-discriminative methods. This type of hybrid framework is often a combination of two stages: the generative stage which produces the features, and the discriminative stage which performs the classification using the features produced by the generative stage [6]. It is noteworthy that the complexity and characteristics in data representation often dictate the model to implement. For instance, in the past, Gaussian data dominated model learning; however, recently, the emergence of multimedia data causes many processing systems to work with count data especially text documents [31,32,37,5,4,7,45,6]. Using the same analogy to modeling techniques, we can observe that in machine learning literature, generative models such as GMM (Gaussian Mixture Models) and HMM (Hidden Markov Models) were very specific to Gaussian data. Despite their strong assumption on parameters (as parametric distributions), these models have often received a lot of attention in the research community because of their simplicity in learning and estimation; most importantly as their functionalities were very well understood in data science [46]. So, the recent proliferation of count data led to the introduction of other generative models such as Beta-Liouville mixtures, generalized Dirichlet mixtures [8,5,4], Dirichlet process mixtures [47,7,48], and finally topic models considered as a new class of generative approaches [31,32,36,15,40,4, 5].

Two main groups define topic models [26,49] in the literature. We have probabilistic models (PLSA (probabilistic latent semantic analysis) and LDA) and non-probabilistic topic models such as latent semantic analysis (LSA), matrix factorization, and non-negative matrix factorization (NNMF) [50,25]. The early success of probabilistic models especially LDA has led to other extensions to enhance the flexibility of LDA. They represent LDA-based topic models. Methods such as Patchinko Allocation topic model [51], correlated topic model [52–54], supervised topic model [55,20,56–58], dynamic topic model [59,29,60,61], hierarchical topic model [62], spherical topic model [63], all characterize these alternatives provided to the LDA architecture. Currently, within the framework of LDA-based topic models, the advancement of social media platforms [64] and online services such as Q&A (questions and answers) [34] communities are having some serious impacts on extensions such as dynamic topic model [65,66,64,29,60], correlated topic model [53,52], supervised topic model, and online topic model schemes [67–70,41]. Current topic models also provide improvement in semantic analysis [30,44,71,72, 29] to enhance coherence in the topics estimated and the relationship between documents [73]. Some current hot topics in research (within topic modeling framework) include social network analysis, bioinformatics [74], emotion, sentiment analysis [65,75,66], and information retrieval [76,35]. It is important to notice that the generative setting, through the BoW representation including its derivates and topic models, have provided tremendous success in computer vision for object learning and categorization [32,77–79,4,5]. Typical to generative techniques, probabilistic topic models use extensively prior information with distributions such as Dirichlet, Beta-Liouville, and generalized Dirichlet [1, 31,42,80,81,7,8].

Particularly, the immediate success of the well-known topic models such as PLSI (probabilistic latent semantic indexing) or PLSA(probability latent semantic analysis) [82] and LDA in text document processing and analysis has been well received in the research community; especially, with the tremendeous contributions of LDA in both text and visual document annotation and categorization [83]. As a parametric model and a generative probabilistic technique initially implemented for topic discovery in large document collections [84], LDA [31] characterizes documents as mixtures of topics while the topics are themselves mixtures over the vocabulary words. By observing the LDA architecture, we can conclude that a very important attribute of topic models (PLSI [82] and HDP(hierarchical Dirichlet process) [85]) is their ability to operate on distribution space where their topic structures (latent variables) are defined as distributions summarizing the characteristics of the documents. They produce multinomial distributions over the topics given the data.

There has been a huge interest in providing extensions within the generative topic model framework by utilizing the flexibility of operating in distribution space. For instance, the work of [86] successfully builds a nonparametric topic model by replacing the document multinomial mixture model in LDA with the kernel density estimator. It is a way of solving the discretization problems related to the clustering and quantization processes during the codebook formation in topic model. It provides a framework that implicitly works on continuous feature space rather than discrete features space in topic modeling. Furthermore, authors in [35] propose a multitopic model with a model selection criteria that solves the problem of predefining a fixed number of topics for 3D object retrieval using the Kullback-Leibler divergence between 3D objects distributions within the BoW. Therefore, improving the characteristics of generative topic models coupled with the possibilities offered by working with distributions became subjects of discussions in the research community as well for tasks such as classification. The motivations include the possibility of carrying potential properties of generative topic models into discriminative classifiers to boost classification performance. This is because a recent development in discriminative setting through kernels formulation also allows SVM to perform with input features that are fully represented as distributions [87]. As a result, due to the success of LDA, recent works in machine learning and computer vision are able to provide extensions that combine LDA with discriminative classifiers [87,88,83]. For instance, authors in [87] provide a way of extracting latent features from probabilistic topic models in distribution space. The features are then used by the SVM for classification. Their topic model (LDA) is implemented within a Bayesian nonparametric setting using HDP (hierarchical Dirichlet process) for model selection. It leads to a topic model kernel that is robust for classification with the SVM. The work in [88] implements kernel topic models where it provides an extension to topic models by replacing the document mixture weights with Gaussian distributions leading to a Bayesian inference based on latent Gaussian. As a Gaussian process latent variable model, the technique is a combination of Gaussian process regression and LDA topic model in a nonparametric setting. In addition, authors in [83] were able to successfully perform classification on high spatial resolution remote sensing images using the LDA topic model with a kernel-based SVM that utilizes a combination of RBF or Gaussian kernels. In [4,5], the authors provide alternatives to the LDA topic model [31] and LGDA (latent generalized Dirichlet allocation) [14] in a classification framework where they combine unsupervised learning (for the topics estimation) to a supervised technique by implementing generative classifiers for the topics similar to the work in [32]. An online version of the Naive Bayes classifier has been proposed within topic modeling environment by the same authors in [41]. In supervised learning, there have been extensive works using hybrid (generative-discriminative) models.

Hybrids in general are able to demonstrate that the performance in discriminative models using SVM always depends on the characteristic of the generative features (data) and the choice of the kernels used [1,46]. Standard kernels such as Gaussian, linear, polynomial were heavily utilized in the past in classification problems with success. This is the case of hybrids that implement for instance GMM or HMM into discriminative classifiers (SVM) using standard kernels [46] with excellent results in object categorization. The complexity in today's data and models characteristics are requesting a new generation of kernels that can cope with the challenge added to the fact that there is a huge interest in working with distributions nowadays. This automatically leads to the introduction of probabilistic kernels. Their flexibility allows a better generalization of the SVM. The SMM (support measure machine) [89] and latent SMM [90] are true examples that illustrate this generalization capabilities of the SVM: they currently represent one of the state-of-the-art techniques for object classification using distributions within the BoW framework in discriminative settings. However, originally, the Fisher kernels proposed by Jaakola and Haussler [21] catalyzed the emergence of kernels for probabilistic generative models used in discriminative classifiers today. Another kernel is the TOP kernel derived from the Tangent vectors Of Posterior log-odds. These two kernels (Fisher and TOP) were successfully used in DNA (Deoxyribonucleic acid) and protein sequence analysis (classification) [24,21].

Other recent hybrids as they exhibit the flexibility of their generative models (based on Beta-Liouville and generalized Dirichlet mixtures) in discriminative classifiers (SVM) have reported similar success in image categorization [6,91] while using probabilistic kernels. Despite the major contributions shown by previous and some recent schemes, they still carry some limitations. For instance, as we emphasize on topic models in this paper, the Dirichlet conjugate prior often affects LDA's performance for positively correlated data. This is because it has a very restricted covariance structure compared to GD and BL that are more flexible [14,40, 5,4]. Many topic models in the literature are LDA-based. This could have a negative impact on the generative process and inferences [4,5] in Dirichlet-based topic models such as LDA. In addition, the possibility of using topic models in discriminative classifiers has created many extensions within the nonparametric setting to account for efficient model selection and processing. However, working with nonparametric models could be very challenging as they require operation or modeling in infinite dimensional spaces. For instance, in [88], the kernel topic model implemented is a Gaussian process latent variable model based on LDA. It has a very complex inference as the framework is not analytically tractable. Similar challenges are noticed in inferences in the work of [87] with the implementation of the HDP in LDA model as it sets the number of latent factors or topics into infinite. Furthermore, the SMM and latent SMM [89,90] have also provided insights on the possibilty of using the concept of distributions within the discriminative platform itself. They have good mathematical foundations and formulations about the space that could allow such implementations as they apply their work in the RKHS (reproducing kernel Hilbert space) that is equipped with an embedding kernel and inner product; however, these techniques in overall could be very complex and require knowledge of vector spaces such as Hilbert spaces which are generalizations of the Euclidean space in finite or infinite dimension. These methods are not hybrids of the type generative-discriminative. They are dedicated discriminative classifiers working indirectly (implicitly) on distributions by using standard kernels in the RKHS [89, 90] where the probability distributions are represented as mean embeddings [89]. Because these methods operate on standard kernels, nonlinear probabilistic kernels such as symmetric KL (Kullback-Leibler) divergence could not be defined directly on the RKHS because of the inner product operation on the Hilbert space.

As we consider all the different characteristics within previous methods that include generative models, (especially topic models), discriminative approaches using SVM, kernels, and hybrid (generative-discriminative) techniques, we propose, in this paper, an extension in topic modeling framework using finite mixtures, similar to LDA. We especially implement a new approach

(hybrid method) that integrates the flexibility of our generative topic model into a powerful discriminative classifier (SVM). It is equipped with well-defined nonlinear probabilistic kernels that allow analysis in distribution space using empirical likelihood (EL) framework for generative topic models in SVM. Within our proposed approach, the use of EL provides distribution estimations. Importantly, with a combination of two different priors (asymmetric GD and BL) used simultaneously within the same generative process, our proposed method introduces a collapsed representation through the collapsed variational Bayesian inference that allows estimation of exact posteriors and easy access to convergence. Most previous generative topic models are either variational-based inference [92] techniques (provide convergence, but posterior estimations are often not exact [38,39,42]) or collapsed Gibbs sampling-based methods (posterior distributions estimations are exact; however, they suffer from convergence [42]). In contrast, our proposed generative topic model is obtained from a combination of two inferences: VB and CGS. It follows the work in [42] which introduced the CVB (collapsed variational Bayesian inference) for LDA. It is one of the state-of-the-art inferences in topic modeling. Though, because of the limitations of the Dirichlet distribution in LDA [14,15,40], we provide alternatives with the use of Beta-Liouville and generalized Dirichlet conjugate priors.
The generative topic model in our proposed approach, because, based on LDA, automatically introduces hierarchies in the observed data with the use of topics as intermediate representations. So, the topic representation in our proposed method could be for instance an alternative to generative models using Beta-Liouville mixtures and generalized Dirichlet mixtures [8]. Using our proposed framework with nonlinear probabilistic kernels, we obtain a system that finally gives us tools to represent any object or document as a distribution parameterized by two mean variables: the document-topic parameter and the topic-word parameter.

3 Proposed Approach

We implement a classification framework where the classifier, the SVM, gets its features from our generative topic models which simultaneously use asymmetric BL and GD as conjugate priors to the multinomial distributions. Documents (images, texts) are first represented as distributions using characteristics of our generative topic models, and then they are presented to the support vector machine for classification. This setting ultimately constitutes our generative-discriminative method that utilizes nonlinear probabilistic kernels. The generative topic models implemented in this paper follow the graphical representation previously proposed in [5, 4,32,41] for object classification using intermediate representations such as topics [32] as shown in Fig.1. Based on the LDA architecture [31], the extensions (generative topic models) we are also providing in this paper are a result of sampling documents and corpus parameters using asymmetric GD and BL priors, simultaneously. In this scenario, the proposed generative process uniquely offers the possibility to either 1) draw the documents parameters from the BL while the corpus parameters are sampled from GD or 2) sample the documents parameters from GD while the corpus parameters are drawn from the BL distribution. This leads to the implementation of two topic models in our proposed generative framework.

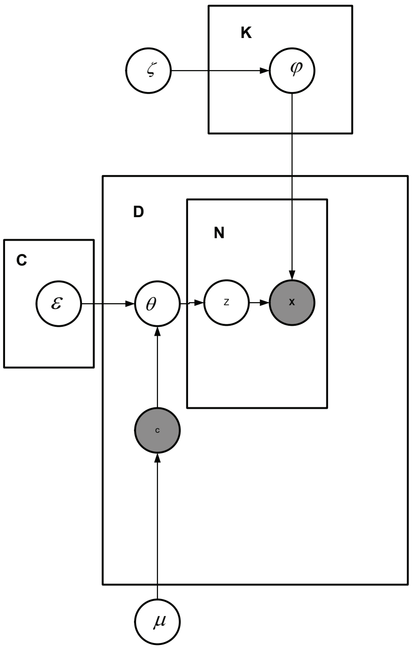

Fig. 1 Generative stage using topic (latent) graphical model. The shaded circle denotes observed variables $\boldsymbol{x}$ and the class c

.

3.0.1 Research objectives

Many techniques related to classification using the hybrids, generative topic models-discriminative methods do not always fulfill the following requirements: 1) the flexibility and the structure (symmetric or asymmetric) of the prior 2) the robustness of the generative process including inference techniques and 3) the choice of kernels. In a supervised topic modeling, these characteristics and requirements are intimately related to each other [93]. However, many hybrid techniques using topic models are just partially robust because they lack some of these essential requirements. In our proposed method, we are mainly implementing a system of integration that takes into account each of these requirements where we provide a combination of much capable and flexible priors (than the Dirichlet) that first helps improving the generative process and inferences. A much improved inference technique is essential for an accurate parameter estimation that increases the coherence and robustness of our generative features and kernel functions formulation. This is the essence of our hybrid model as we formulate a complete framework where we combine two different and flexible priors (BL and GD) within the collapsed variational inference that enables robust generative features for our kernel machine. In addition, the flexibility of our priors and inferences allow us to handle with efficiency inter and intraclass variation problems due to the ability of our method to deal with correlation and semantic analysis effectively. And this includes the possibility of working with a variety of datasets. Our proposed method in its hybrid setting guarantees the best generative topic model and the best discriminative method as we also believe that the SVM is the appropriate candidate in large scale processing compared to the standard Naive Bayes classifier widely used in classification framework that implements topic models [32].

3.0.2 Beta-Liouville and generalized Dirichlet distributions

The generalized Dirichlet (GD) distribution was already introduced and defined in [5,4,15]. In this paper, we also present the Beta-Liouville (BL) distribution (another flexible conjugate prior with a more versatile covariance structure) [40,15,6,8]. Compared to LDA [31], both priors (GD and BL) are now replacing the Dirichlet distribution in topic modeling. We also emphasize on the use of asymmetric priors compared to symmetric ones as they have a direct impact on the robustness of the generative topic models [93].

In a $(K+1)$-dimensional space, the BL distribution with parameters $\varepsilon = (\alpha_1, ..., \alpha_K, \alpha, \beta)$ also written as BL(ε) could be defined as:

$$p(\mathbf{P}|\varepsilon) = \frac{\Gamma\left(\sum_{k=1}^{K}\alpha_k\right)\Gamma\left(\alpha+\beta\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta\right)} \times \prod_{k=1}^{K}\frac{P_k^{\alpha_k-1}}{\Gamma\left(\alpha_k\right)}\left(\sum_{k=1}^{K}P_k\right)^{\alpha-\sum_{k=1}^{K}\alpha_k}\left(1-\sum_{k=1}^{K}P_k\right)^{\beta-1} \tag{1}$$

where $\mathbf{P} = (P_1, ..., P_K)$ is a K-dimensional random variable. Using the notion of conjugate prior to the multinomial, if $\mathbf{P} = (P_1, ..., P_K)$ follows a Beta-Liouville distribution with parameter θ while the vector of counts $\mathbf{X_i} = (X_1, ..., X_{D+1})$ is drawn from a multinomial distribution with parameter $\mathbf{P}$, then the posterior distribution $p(\mathbf{P}|\varepsilon, \mathbf{X_i})$ is also a Beta-Liouville. It therefore leads to the following updates in the posterior distribution $p(\mathbf{P}|\varepsilon, \mathbf{X_i})$.

$$\begin{cases} \alpha'_k = \alpha_k + X_k \\ \alpha' = \alpha + \sum_{k=1}^{K} X_k \\ \beta' = \beta + X_{K+1} \end{cases} \tag{2}$$

As previously mentioned, the implementation of our proposed approach using two conjugate and asymmetric priors (BL and GD) simultaneously, leads to two generative topic models: the first model draws the document parameter from GD while the corpus is sampled from the BL. In addition, it uses the collapsed variational inference (CVB), that is one of the state-of-the-art inference techniques in topic modeling [5,4,42]. We call it the CVB-GD-BL-based topic model or *topic model I*. On the other hand, similarly, the second method uses BL for the document parameter and GD for the corpus parameter within the CVB inference leading to CVB-BL-GD-based topic model. This is *topic model II*.

3.0.3 Generative Processes

LDA is recognized as the simplest topic model where each document is a mixture of K topics in different proportions. Documents while being maintaining K topics in different proportions must belong to same class. This to characterize the observe data. Though in our proposed approach, the BL and GD priors replace the Dir distribution. For instance, in the GD-BL topic model, the generative process is now expressed as follows:

1-We draw topics from $\varphi_k \in$ BL(ζ) for $k \in \{1, 2, 3, ..., K\}$ where $\zeta = (\lambda_{kv}, ..., \lambda_{kV}, \lambda, \eta)$

2-We draw each document $j \in \{1, ..., D\}$

(a) Draw topic proportions $\theta^c \sim GD(\varepsilon)$ where $\varepsilon = (\alpha_{c1}, \beta_{c1}, ..., \alpha_{cK}, \beta_{cK})$ and $c \in \{1, 2, ..., C\}$

(b) For each word $x \in \{1, ..., N\}$
i) Draw topic assignments
$z_{jn} \sim$ Multinomial (θ_d^c)
ii) Draw word
$x_{jn}|z_{jn}, \varphi_k \sim$ Multinomial $(\varphi_{k z_{jn}})$

We could therefore provide a generative of the BL-GD-topic model as well following the same scheme.

3.1 CVB-GD-BL-based topic model

Using concepts such as patches for images [5,4] (similar to words for text analysis) within the BoW, we implicitly elaborate on document representation as visual features in topic modeling framework. In contrast to the standard Naive Bayes classifier for topic modeling, we simply implement in our proposed approach an improved supervised topic model that uses SVM in single-label classification problems. One major contribution is that our proposed method is ultimately done with (a combination of) better priors that provide much flexible generative processes leading to robust inferences and generative features for our kernel functions formulation. In this framework, we can use the variable $\mathcal{X}$ and W interchangeably to denote the collection of words or patches (visual words) in a document or object within the BoW.

3.1.1 Bayesian inference using asymmetric GD and BL priors

From the work presented in [5,4], the generative equation in the fully collapsed space is given by:

$$p(\mathcal{X}, z|\varepsilon, \zeta, c) = \int_\theta \int_\varphi p(\mathcal{X}, z, \theta, \varphi|\varepsilon, \zeta) d\varphi d\theta \tag{3}$$

Due to the prior conjugacy between both GD and BL with respect to the multinomial distribution, Eq.3 becomes easy to compute as it is now expressed as a product of Gamma functions. As a result, the generative equation of the proposed model in the collapsed space of latent variables is:

$$p(\mathcal{X}, z|\varepsilon, \zeta, c) = \prod_{j=1}^{D} \left[\prod_{i=1}^{K} \frac{\Gamma(\alpha_{ci} + \beta_{ci})}{\Gamma(\alpha_{ci})\,\Gamma(\beta_i)} \right] \times \left[\prod_{i=1}^{K} \frac{\Gamma(\alpha'_{ci})\,\Gamma(\beta'_{ci})}{\Gamma(\alpha'_{ci} + \beta'_{ci})} \right] \left[\prod_{i=1}^{K} \frac{\Gamma\left(\sum_{r=1}^{V} \lambda_r\right) \Gamma(\lambda + \eta)}{\Gamma(\lambda)\,\Gamma(\eta) \prod_{r=1}^{V} \Gamma(\lambda_r)} \right] \times \left[\frac{\Gamma(\lambda')\,\Gamma(\eta') \prod_{r=1}^{V} \Gamma(\lambda'_r)}{\Gamma\left(\sum_{r=1}^{V} \lambda'_r\right) \Gamma(\lambda' + \eta')} \right] \tag{4}$$

The equation provided by the joint $p(\mathcal{X}, z|c, \varepsilon, \zeta)$ finally shows some updates due to the multinomial distributions. In the document-topic update in class c, we have:

$$\alpha'_{ci} = \alpha_{ci} + N^i_{j(.)} \qquad \beta'_{ci} = \beta_{ci} + \sum_{l=i+1}^{K+1} N^l_{j(.)} \tag{5}$$

In the topic-word update, it shows:

$$\begin{cases} \lambda'_r = \lambda_r + N^i_{(.),r} \\ \lambda' = \lambda + \sum_{r=1}^{V} N^i_{(.),r} \\ \eta' = \eta + N^i_{(.),V+1} \end{cases} \tag{6}$$

From this point, performing a Bayesian inference in the fully collapsed space is equivalent to approximating the conditional distribution of the latent variable $p(z|\mathcal{X}, \varepsilon, \zeta)$. By integrating out the parameters, the collapsed Gibbs sampler's equation is obtained as an expectation expression:

$$p(z_{ij} = k|\mathcal{X}, \varepsilon, \zeta, c) = E_{p(z^{-ij}|\mathcal{X}, \varepsilon, \zeta, c)}[p(z_{ij} = k|z^{-ij}, \mathcal{X}, \varepsilon, \zeta, c)] \tag{7}$$

such that:

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, \varepsilon, \zeta, c) \propto \left[\frac{(N^{-ij}_{jk.} + \alpha_{ck})(\beta_{ck} + \sum_{l=k+1}^{K+1} N^{-ij}_{jl.})}{(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} N^{-ij}_{jl.})} \right] \times \left[\frac{(\lambda + \sum_{r=1}^{V} N^{-ij}_{.kr_{ij}})}{(\lambda + \eta + \sum_{r=1}^{V+1} N^{-ij}_{.kr_{ij}})} \right] \times \left[\frac{(\lambda_v + N^{-ij}_{.kv_{ij}})(\eta + N^{-ij}_{.k(V+1)_{ij}})}{(\sum_{r=1}^{V} N^{-ij}_{.kr_{ij}} + \lambda_r)} \right] \tag{8}$$

Normalizing the distribution above leads to a posterior probability defined as:

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, \varepsilon, \zeta, c) = \frac{A(k)}{\sum_{k'=1}^{K} A(k')} \tag{9}$$

such that:

$$A(k) = \left[\frac{(N^{-ij}_{jk.} + \alpha_{ck})(\beta_{ck} + \sum_{l=k+1}^{K+1} N^{-ij}_{jl.})}{(\alpha_k + \beta_{ck} + \sum_{l=k}^{K+1} N^{-ij}_{jl.})} \right] \times \left[\frac{(\lambda + \sum_{r=1}^{V} N^{-ij}_{.kr_{ij}})}{(\lambda + \eta + \sum_{r=1}^{V+1} N^{-ij}_{.kr_{ij}})} \right] \times \left[\frac{(\lambda_v + N^{-ij}_{.kv_{ij}})(\eta + N^{-ij}_{.k(V+1)_{ij}})}{(\sum_{r=1}^{V} N^{-ij}_{.kr_{ij}} + \lambda_r)} \right] \tag{10}$$

3.1.2 CVB inference with asymmetric priors

In general, the main goal in Bayesian inference is the estimation of models hidden variables (models parameters and latent variables). This is equivalent to computing the joint posterior distribution $p(z, \theta, \varphi|\mathcal{X}, \varepsilon, \zeta, c)$. Though, the posterior distribution in topic modeling framework is often intractable because the denominator of the posterior equation, the normalizing factor, is not tractable. This normalizing factor is the marginal likelihood. Therefore, inference techniques such as VB and CGS from MCMC (Markov chain Monte Carlo) are often used for hidden variables estimations. The collapsed variational Bayesian inference implemented in our proposed approach is essentially a VB in the collapsed space of latent variables induced by the CGS (Eqs. 7 to 10). As usual, performing VB inference is equivalent to introducing a set of variational distributions (exponentinal family) $\hat{Q}(z, \theta, \varphi)$ that minimize the Kullback-Leibler divergence (KL) between the joint variational distribution $\hat{Q}(z, \theta, \varphi)$ and the true joint posterior distribution $p(z, \theta, \varphi|\mathcal{X}, \varepsilon, \zeta, c)$. The scheme also introduces a lower bound (evidence lower bound or ELBO) to the log marginal likelihood $\log p(\mathcal{X}|\varepsilon, \zeta, c)$. And maximizing the ELBO is equivalent to minimizing the KL$(\hat{Q}(z, \theta, \varphi)||\ p(z, \theta, \varphi|\mathcal{X}, \varepsilon, \zeta, c))$. The lower bound (ELBO) to the log marginal likelihood can be considered as an upper bound (negative ELBO) to the negative log marginal likelihood. So instead of maximizing the ELBO, we could minimize the negative ELBO. This negative ELBO is a functional acting on the joint variational posterior distribution following the work in [42]. It is called variational free energy ($\tilde{\mathscr{F}}(\tilde{Q})$) in the joint space and $\hat{\mathscr{F}}(\hat{Q})$ in the collapsed space).

In CVB, minimizing the variational free energy with respect to $\hat{Q}(\theta, \varphi|z)$ and then with respect to $\hat{Q}(z_{ij}|\hat{\psi}_{ij})$ leads to $\hat{\mathscr{F}}(\hat{Q}(z))$ such that:

$$
\begin{aligned}
\hat{\mathscr{F}}(\hat{Q}(z)) \triangleq \min_{\hat{Q}(\theta,\varphi|z)} \hat{\mathscr{F}}(\hat{Q}(z)\hat{Q}(\theta, \varphi|z)) = \\
\mathbb{E}_{\hat{Q}(z)}[-\log p(\mathcal{X}, z|\varepsilon, \zeta)] - \mathscr{H}(\hat{Q}(z))
\end{aligned}
\tag{11}
$$

Following the work in [5,4,41,37,42] and using Eqs. 8, 9, and 10, the minimization of the functional $\hat{\mathscr{F}}(\hat{Q}(z))$ in Eq. 11 with respect to the variational distribution $\hat{\psi}_{ijk}$ finally gives the following CVB update equation using the Gaussian approximation:

$$
\begin{aligned}
\hat{\psi}_{ijk} = \hat{Q}(z_{ij} = k) \propto \Bigg\{ \Big(\alpha_{ck} + \mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}]\Big) \\
\times \Big(\lambda + \mathbb{E}_{\hat{Q}}[N_{.k.}^{-ij}]\Big) \\
\times \left(\beta_{ck} + \sum_{l=k+1}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jl.}^{-ij}]\right) \\
\times \Big(\lambda_v + \mathbb{E}_{\hat{Q}}[N_{.kx_{ij}}^{-ij}]\Big) \\
\times \Big(\eta + \mathbb{E}_{\hat{Q}}[N_{.k(V+1)_{ij}}^{-ij}]\Big) \\
\times \left(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jk.}]\right)^{-1} \\
\times \left(\lambda + \eta + \sum_{r=1}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kr_{ij}}^{-ij}]\right)^{-1} \\
\times \left(\sum_{r=1}^{V} \lambda_r + \mathbb{E}_{\hat{Q}}[N_{.k.}^{-ij}]\right)^{-1} \times \mathbb{G} \Bigg\}
\end{aligned}
\tag{12}
$$

such that:

$$
\begin{aligned}
\mathbb{G} = \exp\left(-\frac{Var_{\hat{Q}}(N_{jk.}^{-ij})}{2(\alpha_{ck} + \mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}])^2}\right) \\
\times \exp\left(-\frac{Var_{\hat{Q}}(N_{.k.}^{-ij})}{2(\lambda + \mathbb{E}_{\hat{Q}}[N_{.k.}^{-ij}])^2}\right) \\
\times \exp\left(-\frac{Var_{\hat{Q}}(\sum_{l=k}^{K+1} N_{jl.}^{-ij})}{2(\beta_{ck} + \sum_{l=k+1}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jk.}])^2}\right) \\
\times \exp\left(-\frac{Var_{\hat{Q}}(N_{.kx_{ij}}^{-ij})}{2(\lambda_v + \mathbb{E}_{\hat{Q}}[N_{.kxij}^{-ij}])^2}\right) \\
\times \exp\left(-\frac{Var_{\hat{Q}}(N_{.k(V+1)_{ij}}^{-ij})}{2(\eta + \mathbb{E}_{\hat{Q}}[N_{.k(V+1)_{ij}}^{-ij}])^2}\right) \\
\times \exp\left(\frac{Var_{\hat{Q}}(\sum_{l=k}^{K+1} N_{jk.}^{-ij})}{2(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}])^2}\right) \\
\times \exp\left(\frac{Var_{\hat{Q}}(\sum_{r=1}^{V+1} N_{.kr_{ij}}^{-ij})}{2(\eta + \lambda + \sum_{r=1}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kr_{ij}}^{-ij}])^2}\right) \\
\times \exp\left(\frac{Var_{\hat{Q}}(N_{.k.}^{-ij})}{2(\mathbb{E}_{\hat{Q}}[N_{.k.}^{-ij}] + \sum_{r=1}^{V} \mathbb{E}_{\hat{Q}}[\lambda_r])^2}\right)
\end{aligned}
\tag{13}
$$

where:
$\mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}] = \sum_{i' \neq i} \hat{\psi}_{i'jk}$; $\mathbb{E}_{\hat{Q}}[N_{jk.}^{-ij}] = \sum_{i' \neq i} \hat{\psi}_{ijk}(1 - \hat{\psi}_{i'jk})$ in a class. The superscript $-ij$ means all the words except the word ij. It is important to notice that this update equation in CVB is the result from implementing a topic model (CVB-GD-BL-based topic model) where the document and corpus parameters are drawn from asymmetric GD and BL, respectively.

3.1.3 Predictive distributions from the CVB-based topic model

After the sampling process reaches a stationary distribution (convergence), the model parameters that have been initially marginalized out in the fully collapsed space are now estimated. For large samples [37,42], the document predictive distribution in our proposed CVB-GD-BL topic model is therefore given by:

$$\hat{\theta}^{c}_{jk} = \frac{\left(\alpha_{ck} + \mathbb{E}_{\hat{Q}}[N_{jk.}]\right)\left(\beta_{ck} + \sum_{l=k+1}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jl.}]\right)}{\left(\alpha_{ck} + \beta_{ck} + \sum_{l=k}^{K+1} \mathbb{E}_{\hat{Q}}[N_{jl.}]\right)} \tag{14}$$

Conditional on the topic k, the predictive distribution of the words φ_{kv} is:

$$\hat{\varphi}_{kv} = \left(\frac{(\lambda + \mathbb{E}_{\hat{Q}}[N_{.k.}])(\lambda_v + \mathbb{E}_{\hat{Q}}[N_{.kx_{ij}}])}{(\lambda + \eta + \sum_{r=1}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kr_{ij}}])}\right) \times \left(\frac{(\eta + \mathbb{E}_{\hat{Q}}[N_{.k(V+1)_{ij}}])}{(\mathbb{E}_{\hat{Q}}[N_{.k.}] + \sum_{r=1}^{V} \lambda_r)}\right) \tag{15}$$

Following estimation of the predictive distributions (model parameters), the empirical log likelihood could be computed since it is defined as:

$$p(X_j|\theta^c_j, \varphi, \varepsilon, \zeta) = \prod_{ij} \sum_{k} \hat{\theta}^c_{jk} \hat{\varphi}_{kx} \tag{16}$$

following the work in [94][42][36] such that where the following expected counts, $\mathbb{E}_{\hat{Q}}[N_{j..}]$, $\mathbb{E}_{\hat{Q}}[N_{.k.}]$, $\mathbb{E}_{\hat{Q}}[N_{.k(V+1)_{ij}}]$, $\mathbb{E}_{\hat{Q}}[N_{jk.}]$, $\mathbb{E}_{\hat{Q}}[N_{.kx_{ij}}]$, and $\mathbb{E}_{\hat{Q}}[N_{.kr_{ij}}]$ of the unseen document are obtained from the CVB-GD-BL sampling process. The parameters of the unseen document are then used to predict its likelihood. The EL implemented in this paper ultimately follows the work in [36,42,5].

3.2 The CVB-BL-GD-based topic model

Using the framework in [5,4,41,42] and the derivations obtained from subsection 3.1 in this paper, the generative equation in the collapsed space (for M documents and K topics) in our second proposed topic model is:

$$p(\mathcal{X}, z|\varepsilon, \zeta, c) = \prod_{j=1}^{M} \left[\frac{\Gamma\left(\sum_{i=1}^{K} \alpha_{ci}\right)\Gamma\left(\alpha_c + \beta_c\right)}{\Gamma\left(\alpha\right)\Gamma\left(\beta_c\right)\prod_{i=1}^{K}\Gamma\left(\alpha_{ci}\right)}\right] \times \left[\frac{\Gamma\left(\alpha'_c\right)\Gamma\left(\beta'_c\right)\prod_{i=1}^{K}\Gamma\left(\alpha'_i\right)}{\Gamma\left(\sum_{i=1}^{K}\alpha'_i\right)\Gamma\left(\alpha'_c + \beta'_c\right)}\right] \times \prod_{j=1}^{M}\left[\prod_{i=1}^{K}\frac{\Gamma\left(\lambda_r + \eta_r\right)}{\Gamma\left(\lambda_r\right)\Gamma\left(\eta_r\right)}\prod_{i=1}^{K}\frac{\Gamma\left(\lambda'_r\right)\Gamma\left(\eta'_r\right)}{\Gamma\left(\lambda'_r + \eta'_r\right)}\right] \tag{17}$$

where the document-topic update in a class is:

$$\begin{cases} \alpha'_{ci} = \alpha_{ci} + N^{i}_{j,(.)} \\ \alpha'_c = \alpha_c + \sum_{i=1}^{K} N^{i}_{j,(.)} \\ \beta'_c = \beta_c + N^{K+1}_{j,(.)} \end{cases} \tag{18}$$

The topic-word update is:

$$\lambda'_r = \lambda_r + N^{i}_{(.),r} \qquad \eta'_r = \eta_r + \sum_{d=v+1}^{V+1} N^{i}_{(.)d} \tag{19}$$

In the collapsed space, as we integrate out the parameters, the collapsed Gibbs sampler's equation is computed as follows:

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, \varepsilon, \zeta, c) \propto [(\alpha_{ck} + N_{jk.})] \times \left[\frac{(\lambda_v + N_{.kv_{ij}})(\eta_v + \sum_{d=v+1}^{V+1} N_{.kd_{ij}})}{(\lambda_v + \eta_v + \sum_{d=v}^{V+1} N_{.kd_{ij}})}\right] \tag{20}$$

So, normalizing the distribution above provides a posterior probability defined as:

$$p(z_{ij} = k|z^{-ij}, \mathcal{X}, \varepsilon, \zeta) = \frac{\mathcal{A}(k)}{\sum_{k'=1}^{K} \mathcal{A}(k')} \tag{21}$$

such that:

$$\mathcal{A}(k) = (\alpha_{ck} + N_{jk.})\frac{(\lambda_v + N_{.kv_{ij}})(\eta_v + \sum_{d=v+1}^{V+1} N_{.kd_{ij}})}{(\lambda_v + \eta_v + \sum_{d=v}^{V+1} N_{.kd_{ij}})} \tag{22}$$

Following similar steps in subsection 3.1, we reach the final variational update for the CVB-based framework in the second generative topic model where we use the BL and GD for document and corpus parameters, respectively:

$$\begin{aligned} \hat{\psi}_{ijk} = \hat{Q}(z_{ij} = k) \propto \\ \Bigg\{ \left[\left(\alpha_{ck} + \mathbb{E}_{\hat{Q}}[N^{-ij}_{jk.}]\right)\right] \\ \times \left[\frac{\left(\lambda_\nu + \mathbb{E}_{\hat{Q}}[N^{-ij}_{.k\nu_{ij}}]\right)\left(\eta_\nu + \sum_{d=\nu+1}^{V+1} \mathbb{E}_{\hat{Q}}[N^{-ij}_{.kd_{ij}}]\right)}{\left(\lambda_\nu + \eta_\nu + \sum_{d=\nu}^{V+1} \mathbb{E}_{\hat{Q}}[N^{-ij}_{.kd_{ij}}]\right)}\right] \\ \times \exp\left(-\frac{Var_{\hat{Q}}\left(N^{-ij}_{jk.}\right)}{2(\alpha_k + \mathbb{E}_{\hat{Q}}[N^{-ij}_{jk.}])^2}\right) \\ \times \exp\left(-\frac{Var_{\hat{Q}}\left(N^{-ij}_{.k\nu_{ij}}\right)}{2(\lambda_\nu + \mathbb{E}_{\hat{Q}}[N^{-ij}_{.k\nu_{ij}}])^2}\right) \\ \times \exp\left(-\frac{Var_{\hat{Q}}\left(\sum_{d=\nu+1}^{V+1} N^{-ij}_{.kd_{ij}}\right)}{2\left(\eta_\nu + \sum_{d=\nu+1}^{V+1} \mathbb{E}_{\hat{Q}}[N^{-ij}_{.kd_{ij}}]\right)^2}\right) \\ \times \exp\left(\frac{Var_{\hat{Q}}\left(\sum_{d=\nu}^{V+1} N^{-ij}_{.kd_{ij}}\right)}{2\left(\lambda_\nu + \eta_\nu + \sum_{d=\nu}^{V+1} \mathbb{E}_{\hat{Q}}[N^{-ij}_{.kd_{ij}}]\right)^2}\right)\Bigg\} \end{aligned} \tag{23}$$

The parameters estimates for the topic model is as follows:

$$\hat{\theta}^c_{jk} = \frac{\left(\alpha_{ck} + \mathbb{E}_{\hat{Q}}[N_{jk.}]\right)}{\left(\mathbb{E}_{\hat{Q}}[N_{j..}] + \sum_{i=1}^{K} \alpha_{ci}\right)} \tag{24}$$

The predictive distribution of the words φ_{kw} is:

$$\hat{\varphi}_{kv} = \frac{\left(\lambda_v + \mathbb{E}_{\hat{Q}}[N_{.kv_{ij}}]\right)\left(\eta_v + \sum_{d=v+1}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kd_{ij}}]\right)}{\left(\lambda_v + \eta_v + \sum_{d=v}^{V+1} \mathbb{E}_{\hat{Q}}[N_{.kd_{ij}}]\right)} \tag{25}$$

3.3 Discriminative framework: SVM, kernels, and discrete distributions

A probability kernel is defined as the mapping $\mathscr{K} : \mathscr{P} \times \mathscr{P} \rightarrow \mathbb{R}$ with $\mathscr{P}$ defined as the space of probability distributions [95].
For instance, let $\mathcal{X}_i = \{x_{i1}, x_{i2}, ..., x_{iM}\}$ and $\mathcal{X}_j = \{x_{j1}, x_{j2}, ..., x_{jM}\}$ be two sequences of vectors for two multimedia objects X_i and X_j, respectively. Then, each object is associated with its probablility density function $p(x|\Omega_i)$ and $q(x|\Omega_j)$, respectively. These are parametric distributions such that Ω_i is the parameter for object X_i while Ω_j is the parameter for object X_j. When implementing topic models especially in computer vision, the bag of visual words scheme leads to the discretization of the continuous visual features space as we perform clustering and quantization methods for the elaboration of the codebook [86,87]. This discretization causes the reformulation of the kernels using discrete distributions instead of PDFs (probability density functions).
For our generative topic model framework in the collapsed space, we recover the parameters through sampling process of the topic assignments $\boldsymbol{z}$. Let Ω be defined as $\Omega = \{\theta^c, \varphi\}$ such that θ^c is $1 \times K$ vector (document-topic parameter) and φ is a $K \times V$ (word-topic) parameter for the corpus such that its entries are φ_{kv} from $\varphi = \{\varphi_{kv}\}$. Let $\hat{\Omega}$ be the estimate of Ω such that $\hat{\Omega} = \{\hat{\theta}^c, \hat{\varphi}\}$. With $\hat{\Omega}$, we can efficiently represent the PMF (probability mass function) of each document X_j. In other words the SVM carries the generative predictive distributions for each document obtained by marginalizing out the topic model parameters. With documents in the generative stage equipped with probabilities (Eq.16), we can define the different probabilistic kernels in the following section. Let P and Q be two distributions defined on the space Δ such that $p(x)$ and $q(x)$ represent the densities of P and Q, respectively. For our supervised topic model framework using SVM, we replace the kernel formulation in the standard (original) feature space by the one in the distribution space that accounts for topic generative features as shown in:

$$\mathscr{K}(\mathcal{X}_i, \mathcal{X}_j) \Rightarrow \mathscr{K}(P, Q) \tag{26}$$

There have been many ways of characterizing the generative structure (features) in topic models. For instance authors in [35] in 3D object retrieval system use the LDA document topic proportions θ and the KL divergence to compute the distance between two 3D objects. In their work, the topic proportions θ represent an object. However, in [67], authors implement the Jensen-Shannon divergence by considering the topics φ_k themseleves to evaluate the change in topics between two successive time slices. Similar choice is suggested in [70] where authors define the topic as a vector of probabilities over the space of words and then formulate the KL divergence between two topic distributions to assess their dissimilarity.
As topic mixtures are parameterized by θ^c while the topics themselves are parameterized by φ, we decide in our proposed approach to parameterize each document with both θ^c and φ. This representation is in line with the definition of a document in topic modeling which is described as a mixture over topics where each topic is a mixture over the vocabulary. Therefore, each discrete document multinomial distribution fed to the SVM could be described by Eq.16 as a PMF parameterized by θ^c_j and φ .

3.3.1 The Kullback-Leibler kernel

Based on information divergence measures (where the measure here is the KL divergence), this probabilistic kernel computes the dissimilarity between two probability density functions $p(x|\Omega_i)$ and $q(x|\Omega_j)$ defined on the support (space) Δ:

$$D_{KL}(P, Q) = \int_{\Delta} p(x|\Omega_i) \log\left(\frac{p(x|\Omega_i)}{q(x|\Omega_j)}\right) dx \tag{27}$$

The KL divergence between P and Q $(KL(P||Q))$ could be seen as the additional amount of bits needed to encode samples from P distribution using a Q distribution-based code [70,17].
From the KL divergence measure, we can evaluate the symmetric KL divergence as:

$$\begin{aligned} D_{SKL}(P, Q) = &\int_{\Delta} p(x|\Omega_i) \log\left(\frac{p(x|\Omega_i)}{q(x|\Omega_j)}\right) dx \\ &+ \int_{\Delta} q(x|\Omega_j) \log\left(\frac{q(x|\Omega_j)}{p(x|\Omega_i)}\right) dx \end{aligned} \tag{28}$$

For discrete probability distributions $P(x)$ and $Q(x)$, we can reformulate $D_{KL}(P,Q)$ over the support $\varDelta$ as:

$$\begin{aligned} D_{SKL}(P,Q) = &\sum_{x\in\varDelta} P(x)\log\left(\frac{P(x)}{Q(x)}\right) \\ &+ \sum_{x\in\varDelta} Q(x)\log\left(\frac{Q(x)}{P(x)}\right) \end{aligned} \tag{29}$$

Once the symmetric KL divergence measure is defined, the KL kernel [46] is estimated by exponentiating the symmetric KL divergence.

$$\mathscr{K}(X_i,X_j) \Rightarrow \mathscr{K}\,(P,Q)) \Rightarrow \exp\left(D_{SKL}(P,Q)\right) \tag{30}$$

3.3.2 The Jensen-Shannon kernel

It is based on Jensen-Shannon (JS) divergence [96] as it measures the similarity between two distributions. The JS divergence between distributions P and Q is defined as:

$$JS(P||Q) = \mathscr{H}[\upsilon P+(1-\upsilon)Q]-\upsilon\mathscr{H}[P]-(1-\upsilon)\mathscr{H}[Q] \tag{31}$$

with υ a parameter and $\mathscr{H}[P]$ the Shannon entropy of P over the space $\varDelta$ is $\mathscr{H}[P] = -\int_{\varDelta} p(x|\varOmega_i)\log p(x|\varOmega_i)dx$ such that p is the density of distribution P. A discrete formulation of the Shannon entropy is:

$$\mathscr{H}[P] = -\sum_{x\in\varDelta} P(x)\log P(x) \tag{32}$$

The Jensen-Shannon kernel is obtained by exponentiating the JS divergence.

$$\mathscr{K}_{JS}(P,Q) = \exp(-aJS(P||Q)) \tag{33}$$

The JS could also be formulated using the KL by setting $g(x) = \frac{1}{2}p(x) + \frac{1}{2}q(x)$ with $\upsilon = 1/2$

$$JS(P||Q) = \frac{1}{2}KL(P||G) + \frac{1}{2}KL(Q||G) \tag{34}$$

3.3.3 The Bhattacharyya kernel

It is a member of the probability product kernel (PPK) family [97] that is defined as:

$$\mathscr{K}_\rho(P,Q) = \sum_{x\in\varDelta} P(x)^\rho Q(x)^\rho \tag{35}$$

such that ρ is a parameter. Following the formulation in Eq.35, we can define the Bhattacharyya kernel [98] as a PPK at $\rho = \frac{1}{2}$:

$$\mathscr{K}_{\frac{1}{2}}(P,Q) = \sum_{x\in\varDelta} \sqrt{P(x)}\sqrt{Q(x)} \tag{36}$$

However, when $\rho = 1$, the PPK becomes the expected likelihood kernel, also called the correlation kernel as it measures the corelation between two distributions such that:

$$\mathscr{K}_1(P,Q) = \sum_{x\in\varDelta} P(x)Q(x) \tag{37}$$

Because it is related to traditional linear kernels, the correlation kernel is called probabilistic linear kernel [95].

3.3.4 The Renyi kernel

Straight from the Shannon entropy theory, the Renyi kernel is based on the Renyi divergence measure of order σ:

$$D_\sigma(P||Q) = \frac{1}{\sigma-1}\log\sum_{x\in\varDelta} P(x)^\sigma Q(x)^{1-\sigma} \tag{38}$$

where $\sigma > 0$ and $\sigma \neq 0$
By exponentiating the symmetric Renyi divergence, it leads to the Renyi kernel that is defined as: $\mathscr{K}_R(P,Q) = \exp\{-a(D_\sigma(p(x|\varOmega_i)||q(x|\varOmega_j)) + D_\sigma(q(x|\varOmega_j)||p(x|\varOmega_i))\}$ where $a > 0$.

$$\begin{aligned} \mathscr{K}_R(P,Q) = &\left[\log\sum_{x\in\varDelta} P(x)^\sigma Q(x)^{1-\sigma}\right] \\ &\times\left[\log\sum_{x\in\varDelta} Q(x)^\sigma P(x)^{1-\sigma}\right]^{\frac{a}{1-\sigma}} \end{aligned} \tag{39}$$

The Renyi divergence is a generalization of the KL divergence, and both are identical when $\sigma \to 1$. In addition, the Renyi kernel becomes a PPK when $a = \frac{1-\sigma}{2}$. It also a Bhattacharyya kernel for $\sigma = \frac{1}{2}$.

3.4 Time and memory complexities

The time and memory complexities have been presented in many topic model publications [23,99,36,42,5,17]. Though the work of [99] provided the most extensive details about time and memory complexities when processing large collections under LDA. Following the work in [99], for D documents containing each N words from a vocabulary of size V, in a particular class c, we obtain a $D\times V$ matrix where $NN0$ is the total number of nonzero elements in this document-word (sparse) matrix. During the formation of K topics, it involves placing a $K+1$-dimensional variational distribution on every word leading to a $K\times NN0$ matrix. The parameter estimation provided the predictive document-topic distribution θ_j^c of size $K\times D$ and the topic-word predictive distribution of size $K\times V$. CVB-LDA carries

Table 1 Models time complexities

Models	Time complexity
L	$O(\xi \times 2 \times K \times NN0)$
S	$O(M^3)$
L+S	$O(\xi \times 2 \times K \times NN0) + O(M^3)$
I	$O(\xi \times 2 \times (K'') \times NN0)$
II	$O(\xi \times 2 \times (K'') \times NN0)$
I + S	$O(\xi \times 2 \times (K') \times NN0) + O(M^3)$
II + S	$O(\xi \times 2 \times (K'') \times NN0) + O(M^3)$

Table 2 Models memory complexities

Models	Memory complexity
L	$O(K \times 2 \times (V + D) + NN0)$
S	$O(M^2)$
L+S	$O(K \times 2 \times (V + D) + NN0) + O(M^3)$
I	$O(K' \times 2 \times (V' + D) + NN0)$
II	$O(K'' \times 2 \times (V'' + D) + NN0)$
I + S	$O(K' \times 2 \times (V' + D) + NN0) + O(M^2)$
II + S	$O(K'' \times 2 \times (V'' + D) + NN0) + O(M^2)$

$K \times NN0$ matrix along with two copies $\hat{\theta}$ and $\hat{\varphi}$ one for the inference and the second one for the correction factor using the variance. This leads to a time complexity of $O(\xi \times 2 \times K \times NN0)$ where ξ is the extra cost for the exponential correction factor. The brute space complexity is around $O(K \times 2 \times (V + D) + NN0)$

In SVM we carry M documents of size $1 \times K$ for each class. Let's call M the documents/topics pairs during the training stage. The time complexity of SVM is $O(M^3)$ while the memory complexity is $O(M^2)$ where $M \leq D$. Though due to the flexibility of GD and BL in pruning out irrelevant topics, we usually obtain $K' \leq K$, $K'' \leq K$ and $V' \leq V$, $V'' \leq V$ under BL and GD. Therefore, the memory and time complexities are improved. For instance, in CVB-GD-BL(*topic Model I*), we have this memory complexity below $O(\xi \times 2 \times (K' \times NN0)$ $O(K' \times 2 \times (V' + D) + NN0)$. We could obtain the memory and time complexities of topic *topic Model II* just by using K" and V'' which the reduced versions of the vocabulary and topics.
LDA does not have ability for to retain the most relevant topics due to its Dirichlet prior. It leads to $O(\xi \times 2 \times K \times NN0)$ $O(K \times 2 \times (V + D) + NN0)$ for LDA as shown in [99]. Tables 1 and 2 recapitulate complexiy of the proposed approach compared to the standard LDA.

In these tables L=LDA, S=SVM, I= *topic Model I*, II=*topic Model II*. We can see that under the time and memory complexities, the LDA is slower and uses a lot of memory than our proposed models. we can also observe that the proposed techniques perform almost equally with their reduced number of topics and vocabulary. In the worst case, our GD-BL and BL-GD topic models will have the same time complexity as LDA. However, those are very flexible topic models that execute many tasks at the same time including semantic analysis between word and between topics. This suggests they execute each task faster than LDA. LDA does not perform in topic correlation. So, it is slower than our proposed models [5]. Tables 1 and 2 shows how the topic correlation analysis improve the time and memory complexities.

4 Experimental results

We show the robustness of our proposed approach by selecting some real world and challenging applications in image and text classification. Our framework provides the generative topic models which are then used in SVM. The SVM operates on a series of kernels (in distribution space) such as Bhattacharyya kernel (BK), symmetric Kullback-Leibler divergence kernel (KLDK), the Renyi kernel (RK), the Jensen-Shannon kernel (JSK), and the Expected likelihood kernel (ELK). In our SVM implementation in this paper, as we are dealing with a multiclass problem, we select the one-versus-all technique for the training set modeling: that is, the class with the largest positive score will ultimately win the class label. In addition, an 8 fold-cross validation scheme has been implemented to account for the estimation of the design parameters within the SVM.
Using the collapsed Gibbs sampling method and the empirical likelihood scheme, each document distribution is evaluated over the finite set of topics. As we are demonstrating the performance of our proposed approach compared to previous topic models illustrated in Table 5 using probabilistic kernels, we also include cases where we compare our proposed topics models to SVMs operating in the original feature space using standard kernels such as linear and RBF. Consequently, we include the performance of our proposed topic models with a linear kernel-based SVM for the text document dataset.

4.1 Implementation

This implementation concerns the generative stage where we construct the topic distributions to be utilized by the SVM. In this implementation, we are using the collapsed Gibbs sampling method in variational Bayes inference. The variational update equation is similar to the update equation in the standard CGS. The difference is that here we sample from the variational distribution instead of sampling from the true posterior distribution. Immediately, to deal with the digamma functions, we can reset the variational update equation using [100] work. The main idea is to compute the variational model parameters θ_j^c and φ_k using the CVB algorithm which implements this variant of CGS. To do this, we set a number of iterations such that at each iteration we sample a topic for each of the N words in the corpus. We use the variational expected count variables (the variational statistics). We use these statistics to estimate the topic model parameters at the generative stage. The framework requires an initial setting of the variational expected count variables along with the model hyper-parameters. We usually set them randomly. Though, many times for the BL hyperparameters, we could also provide initializations in this way: within a class, at the document level we choose $\alpha_{jk} = \frac{1}{k}$ where k $\in\{1, 2, ..., K\}$. We also set α_j such that $\alpha_j \leq \sum_{k=1}^{K} \alpha_{jk}$ or $\alpha_j \geq \sum_{k=1}^{K} \alpha_{jk}$. Then, we choose β_j within the same scale as α_j. At the corpus level for BL, we repeat the same process by setting values for λ_{kv} with $v \in \{1, 2, ..., V\}$ and λ and η For the GD hyperparameters at the document level $\alpha_{jk} = \frac{t}{i}$ with $i \in \{1, 2, ..., K\}$ and $\beta_{ji} = \frac{1}{K+i}$ with $i \in \{1, 2, ..., K\}$. At the corpus level we also repeat the same process with λ_{iv} and η_{iv} with $v \in \{1, 2, ..., V\}$. We initialize the number of topics along with the maximum number of iterations. We also randomly initialize the topic assignment associated to each word in the class in the latent $\boldsymbol{z}$ (N-dimensional random variable) associated to each document j. The main expected counts in the sampling process include $\mathbb{E}_q[N_{jk}]$ the number of words assigned to topic k in document j, $\mathbb{E}_q[N_{j(K+1)}]$ the total number of words in topic $K + 1$ in document j, $\mathbb{E}_q[N_{kv}]$ the number of times the vth word in the vocabulary is assigned to topic k, $\mathbb{E}_q[N_{k(V+1)}]$ the number of times the $(V + 1)th$ word in the vocabulary is assigned to topic k, $\mathbb{E}_q[N_k]$ the the number of times any word is assigned to a topic k $\mathbb{E}_q[N_{j(K+1)}]$ the total number of words in topic $K + 1$ in document j. In the document which is a collection of vocabulary words w organized as count data, we associate each word to its initial count (frequency count). In CVB-based CGS algorithm, as shown in Eq. 12 and 23, we remove the current topic assignment from these update equations by decreasing the count associated to the current assignment. We compute the probability of each topic assignment using Eq. 12 and 23 leading to a discrete distribution, a K-dimensional variational distribution associated to every word. We sample from this distribution of latent topic assignments and choose a topic that is returned to vector z where it updates the counts. In other words, the appropriate counts are increased. At the covergence, we collect the latent variables z, the variational statistics which allow us to compute the predictive distributions for the document paramter (document-topic), θ_j^c and corpus (topic-work) parameter φ_k.

4.2 Text document classification

4.2.1 Preprocessing and methodology

In this paper, we chose a challenging text classification problem using our proposed hybrid technique. For this work, we selected the Yahoo! Answers topic classification dataset. This dataset has been constructed from the original Yahoo! Answers corpus which is a vast collection of text documents containing around $4,483,032$ questions and their corresponding answers (in .csv format). This current dataset has been used in a text classification problem by Zhang et al. in [101]. In fact, the Yahoo! Answers topic classification dataset has 10 main categories (Table 3) where the total training set is about $1,400,000$ samples ($140,000$ samples per category). The testing set contains $60,000$ samples (6000 samples per category). The dataset has a 4 column text layout where the first column carries the class labels of each text document. The second and third columns provide questions while the last column shows the best answers to those questions. In our case, in this particular text classification problem, we are interested in documents containing answers and their corresponding classes from the corpus layout.

Though, in this experiment, we did not use the whole dataset as we utilized only a subset of the data that consists of 6000 samples per category, so 60000 samples in total. This is mainly due to poor initializations which were slowing down the sampling process. We reduced the size to speed up the sampling process.

As usual for text document data, the collections are initially unstructured or noisy as they carry a lot of unwanted materials. Consequently, in the preprocessing stage, we cleaned up the data by removing irrelevant items such as stop words and punctuations through MATLAB. In each class, 90% of the dataset have been assigned to the training set while the remaining is the testing set. The training set obtained is then used to construct the bag of words from the tokenized documents. Further preprocessing steps have been implemented to remove infrequent words in BoW model (for instance, words that appear less than two times in documents). In addition, empty documents have been also removed from the training data. The characteristics of the training set following the BoW framework is summarized in Table 4 which shows the frequency count data represented in a matrix form, the total number of documents, the vocabulary size, and the total number of words in the corpus.

The frequency count data (training set) is then used by our algorithm where we learn documents topics first: this is the generative stage. It is important to mention that our text data using the BoW framework is really sparse due to the large size of the vocabulary. We proceeded with a sparse-based data representation for efficient storage management in this batch processing.
For the generative framework, we finally obtain the optimal number of topics at $K = 60$. Once our generative topic model is built, we represent each document as a topic distribution. We, in fact, constructed two generative topic models: the *topic model I* or CVB-GD-BL-based topic model and *topic model II* or CVB-BL-GD-based topic model, all presented in section 3. With these topic models, we estimated the predictive distributions that allow us to express the document distributions using Eq.16. The topic distribution are then used by the SVM classifier to perform document categorization with probabilistic kernels. The representation of each document as probability distribution is fully detailed in subsection 3.3 in this paper. Importantly, there are no clustering and quantization steps for text documents during the BoW formation. These steps only occur when dealing with images in feature representation. Text documents naturally decompose into bag of words. This ultimately summarizes our generative-discriminative approach for text document classification.

4.2.2 Results

Initially, the BoW representation of the data shows a very sparse data, and the first samples used for modeling did not yield good approximates. It means there is a need to provide more discriminative features that facilitate classification. As we increase the size of the documents and the number of latent factors or topics as shown in Figs. 2, 3, and 4, we saw an immediate improvement in the estimates for the topics and the distribution over the topics. The improvement in the estimates not only shows the characteristics of each document, but also exhibits differences between these documents by observing the topic and distribution structures.
The experimental results from our proposed approach with this text dataset using the different probabilistic kernels utilized in this paper are summarized in Table 5. These results show that our two generative (topic) models implemented, (*topic model I* and *topic model II*) have provided satisfactory performance with SVM framework compared to their major competitors (such as LDA, CVB-LDA, CVB-LGDA, and LGDA) in this text document classification. So, the hybrids obtained from the proposed topic models outperformed their competitors under these probabilistic kernels.

All the hybrids in this text document classification have successfully provided good results with the expected likelihood kernel (ELK) which is a linear probabilistic kernel. Under the ELK-based SVM, the *topic model II* had the highest accuracy (68.53%). In overall, results from hybrids using *topic model II* were slightly improved compared to hybrids from *topic model I.* In this experiment, the linear kernel was able to outperform nonlinear kernels in text document classification. We think that linear probabilistic kernels could be seen as alternatives to nonlinear probabilistic kernels in text document classification. Though, the JSK-based SVM coupled with *topic model II* remains the best performer among nonlinear probabilistic kernel models.
This ultimately demonstrates the robustness of probabilistic linear kernels in text document classification. However, both topic models proposed in our work outpformed an SVM-based classifier using traditional and standard linear kernel (in the original feature space). The classification accuracy with topic Model I is 58.43%. It is 56.38% with topic Model II, and 54.41% with SVM. Finally, these performances ultimately illustrate the importance of documents representation in distribution space. And this starts from providing an optimal number of topics from our generative models from which distributions are built over the topics structure. The ability to summarize documents (initially represented in 8805 dimensional feature space in this paper due to the size of the vocabulary within the BoW as shown in Table 4) using efficient and very few low dimensional features such as topics is an ideal framework for memory space management in databases. From documents with initially 8805 features, we obtain at the end $K = 70$ topics from the generative model to represent documents new features in distribution space.

Table 3 10 Categories for text documents

Text Document Categories	Class label
Society & Culture	1
Science & Mathematics	2
Health	3
Education & Reference	4
Computers & Internet	5
Sports	6
Business & Finance	7
Entertainment & Music	8
Family & Relationship	9
Politics & Government	10

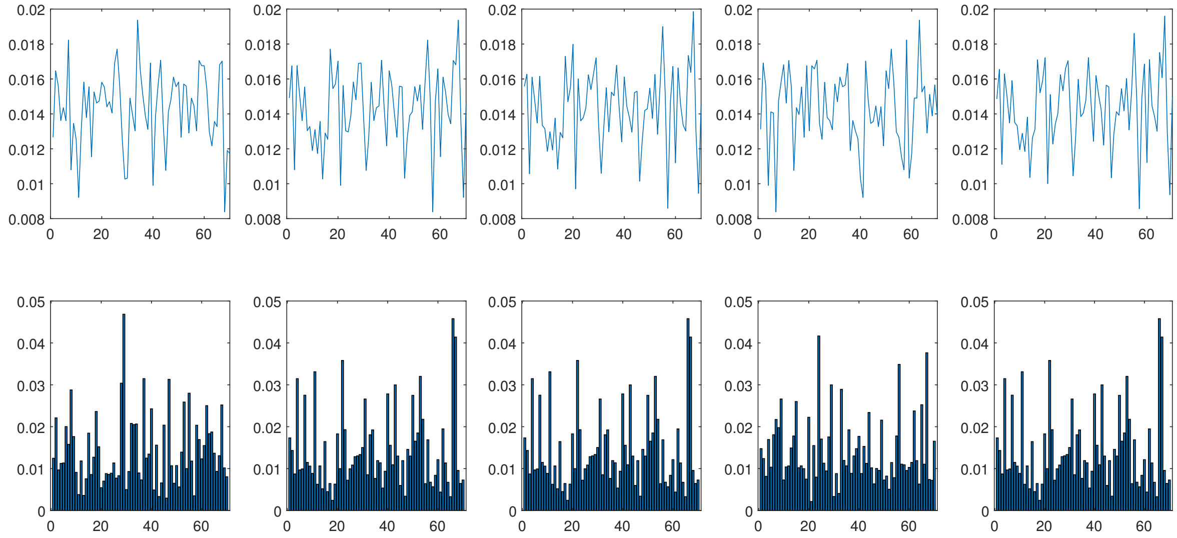

Fig. 2 Processing results from increasing documents size and the number of topics

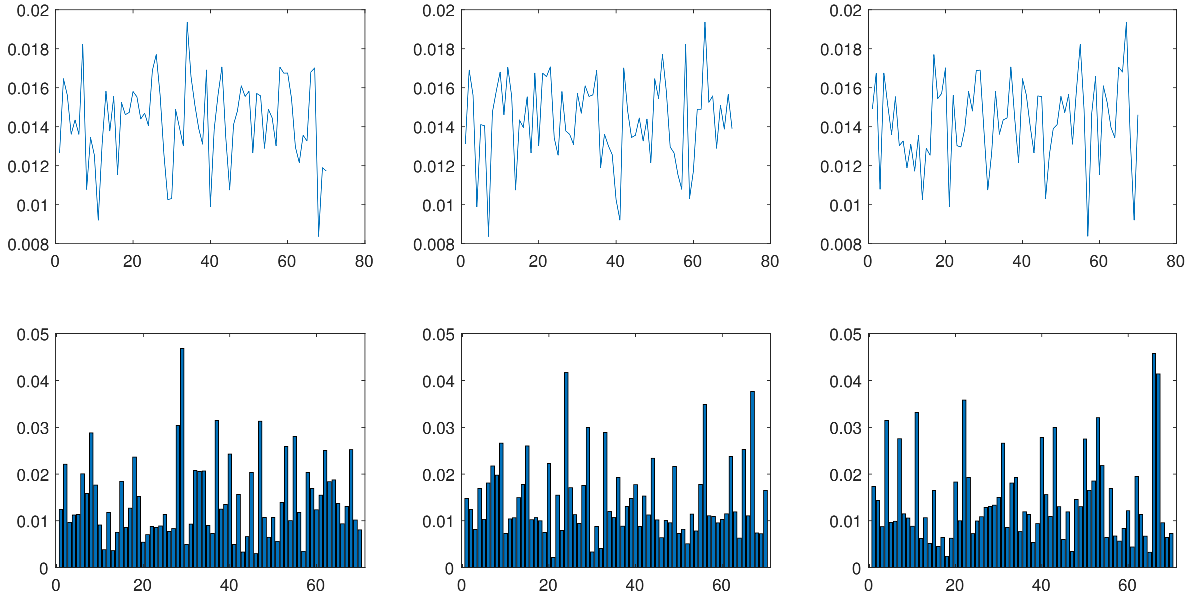

Fig. 3 Three classes from text corpus documents with associated topic structure

Table 4 BoW information for the text document modeling

BoW	Characteristics
Total Counts	53087 × 8805
Vocabulary	[1 × 8805 string]
Total Number of Words	8805
Total Number of Documents	53087

4.3 Natural scene categorization dataset

4.3.1 Preprocessing

In this experiment, we are performing image classification using our proposed hybrid framework. We also used the well-known natural scenes dataset that has 15 categories as shown in [102]. It is a challenging dataset. Here is the list of the classes along with their size: (Suburb,

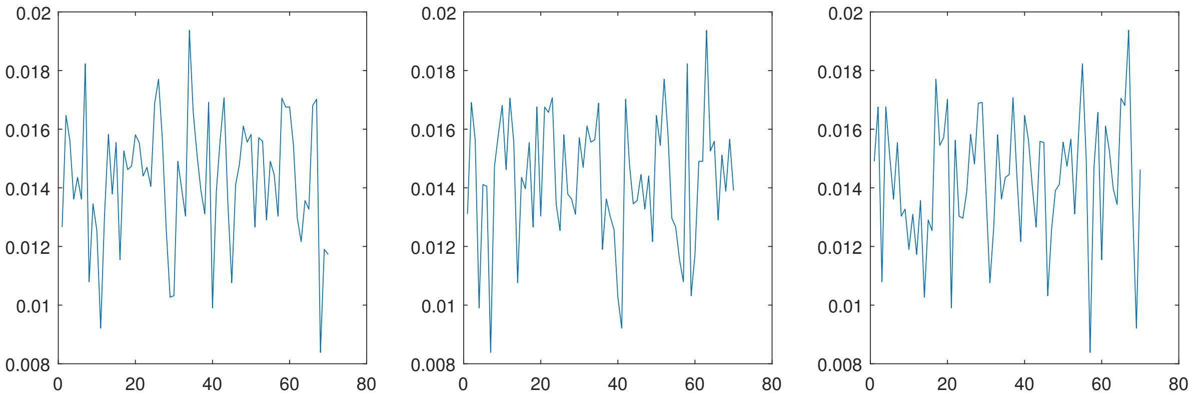

Fig. 4 Multinomial distributions from 3 text documents of different classes

Table 5 Hybrid models performances for the text document dataset

%	BK	KLDK	RK	JSK	ELK
topic model I	61.45	62.16	62.27	63.49	67.51
LDA	45.56	48.67	49.25	50.67	57.89
CVB-LDA	46.12	49.87	57.43	54.89	55.57
CVB-LGDA	50.78	51.65	52.10	53.09	57.16
LGDA	48.36	48.98	49.67	50.18	56.54
topic model II	63.32	63.67	65.74	66.19	68.53

241), (Living room, 289), (Coast, 360), (Forest, 328), (Highway, 260), (Mountain, 374), (Street, 292), (Office, 215), (Store, 315), (Bedroom, 216), (Inside city, 308), (Tall buidling, 356), (Open country, 410), (Kitchen, 210), and (Industrial, 311). The corpus as illustrated in Fig. 5 has in total 4485 images. The dataset is also a collection that contains different categories (for instance, mountain and highway) as well as similar categories (for instance, the 4 indoor categories such as office, living room , kitchen, and bedroom from [32]) to fully characterize the concept of interclass and intraclass variation problems.

From each category, the dataset is split in two groups: the testing set carries 100 samples while the training set gets the remaining. This is similar to our previous work with this data in [5,4] where we used the BoW method to transform the SIFT (scale invariant feature transform) descriptors (from image patches) into codebook or vocabulary after clustering and quantization process [32,6,5,4]. The training set (count data) obtained is then used to build our generative topic models with asymmetric priors. Following the steps in the text classification problem in subsection 4.2, we characetrize each document distribution using subsection 3.3. These documents are then used by our SVM which performs with probabilistic kernels. It is noteworthy that based on our previous work [5,4], the optimal number of topics and vocabulary size are reached at $K = 90$ and $V = 1000$, respectively for the implementation of our generative topic models.This is because of the ability of the GD and BL in pruning irrelevant topics and vocabulary size. We therefore obtained a model selection with very reduced number of topics and vocabulary size.

4.3.2 Results

We showed earlier the low performance of the hybrids with the expected likelihood kernel (ELK): 58.43% for topic Model I, 56.38% with topic Model II. This probabilistic linear kernel was not able to carry enough discriminative information or features that could enhance performances in this image categorization problem. In general, in this experiment, nonlinear probabilistic kernels used in this hybrid generative-discriminative setting have been observed to outperform the ELK. The two topic models in our proposed approach combined with nonlinear probabilistic kernels-based SVM show robustness of our methods with a result around 85% in accuracy from *topic model II*. These two hybrids in our scheme seem to equally perform well with nonlinear probabilistic kernels especially the JSK. They both outperform their competitors such as LDA, CVB-LDA, CVB-LGDA, and LGDA. These results show that nonlinear probabilistic kernels are robust and efficient in image classification than in text categorization. In this experiment, nonlinear kernels are able to characterize the intrinsic properties in images than linear probabilistic kernels represented by the ELK. This justifies the poor performance in the ELK for its inability to adapt to changes in view and illumination in images for instance since such phenomena induce nonlinearity in the dataset resulting in changes in document distributions. This instability in the distributions has a negative impact on linear probabilistic kernel function (ELK).

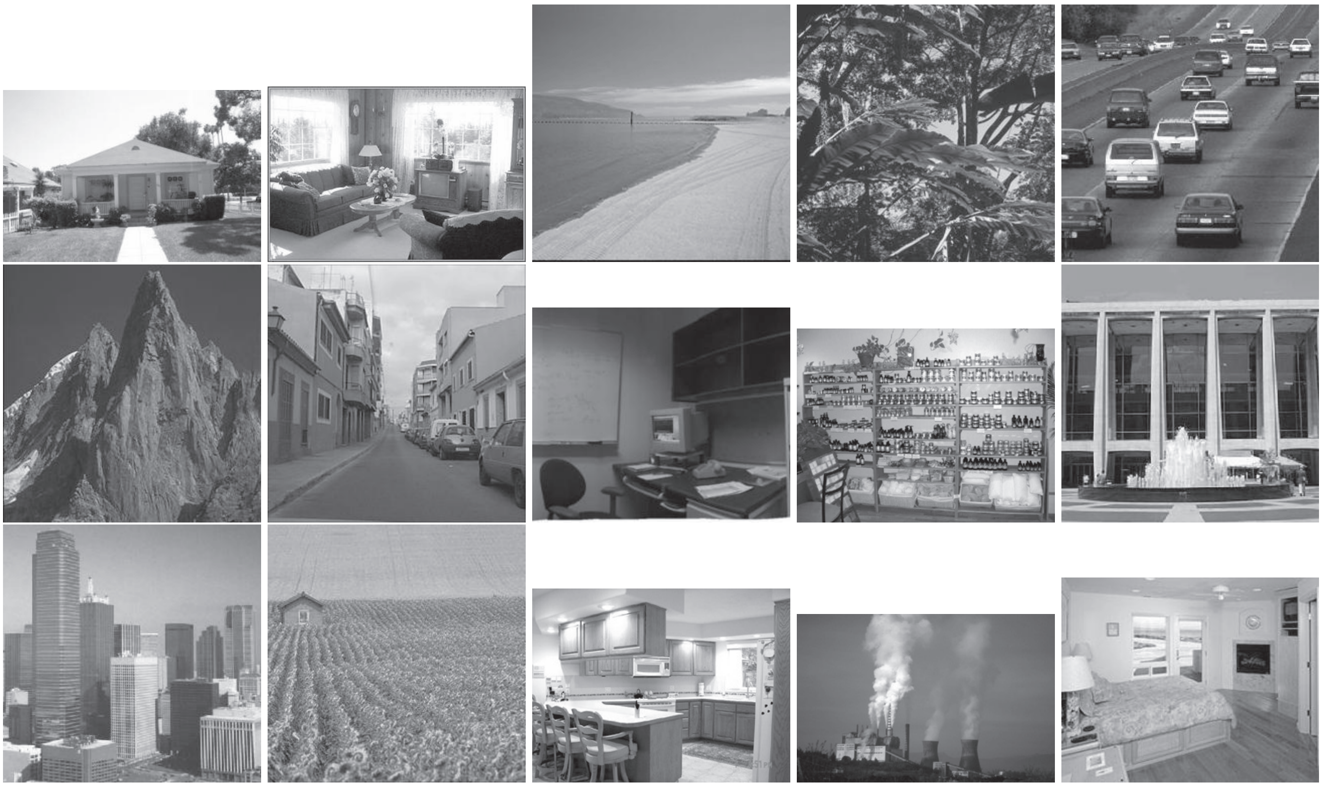

Fig. 5 Examples from the natural scenes image dataset (15 categories).

In addition, the proposed topic models (implemented in this paper) performances have been compared to a Gaussian or RBF kernel-based SVM classifier which operates in the original feature space (75.35% with *topic Model I* and 76.65% with *topic Model II*). The SVM with RBF kernel using orginal feature instead topic distribution provided an accuracy of 68.27%. These topic models outperform the RBF-based classifier. The performance of our method could also be explained by the robustness in the generative topic models for their ability to characterize effectively the documents as probability distributions with a better parameterization. For instance, a random selection of 5 documents has been made whitin the natural scene category dataset. As shown in Figs. 6 and 7, and similar to the scenario presented in our text document classification, the first row, in each figure, illustrates the convergence process while the second row exhibits the word distribution in the documents. The last row provides the topic structure in each document. Under our proposed approach, we can see that the documents are different according to their classes. In Fig. 6 for instance, on the second row, documents 1, 2, and 4 have similar topics and similar distributions over topics. Still on the second row, same observations could be made about documents 3 and 5. These 5 documents ultimately belong to 2 classes from their distribution characteristics. This robustness in approximating effectively the generative topic model facilitates the task for the probabilistic kernel to perform accurately as it measures similarity between distributions within the discriminative framework. As we start increasing the size of the dataset, the number of topics, and the size of the vocabulary during training, we notice improvement in the results with our proposed hybrids. The final optimal number of topics and size of the vocabulary are obtained at $K = 90$ and $V = 600$, respectively. This constitutes the characteristics of the generative approach that we use to construct our discriminative classifier.

Table 6 Hybrid models performances for the natural scenes dataset

%	BK	KLDK	RK	JSK	ELK
topic Model I	78.31	79.18	82.17	82.32	70.65
LDA	59.34	65.54	68.67	69.43	55.41
CVB-LDA	65.38	70.3	70.86	71.57	57.85
CVB-LGDA	70.51	69.96	75.71	80.53	68.34
LGDA	68.45	70.43	75.35	77.64	63.50
topic Model II	78.54	78.98	80.78	85.47	74.67

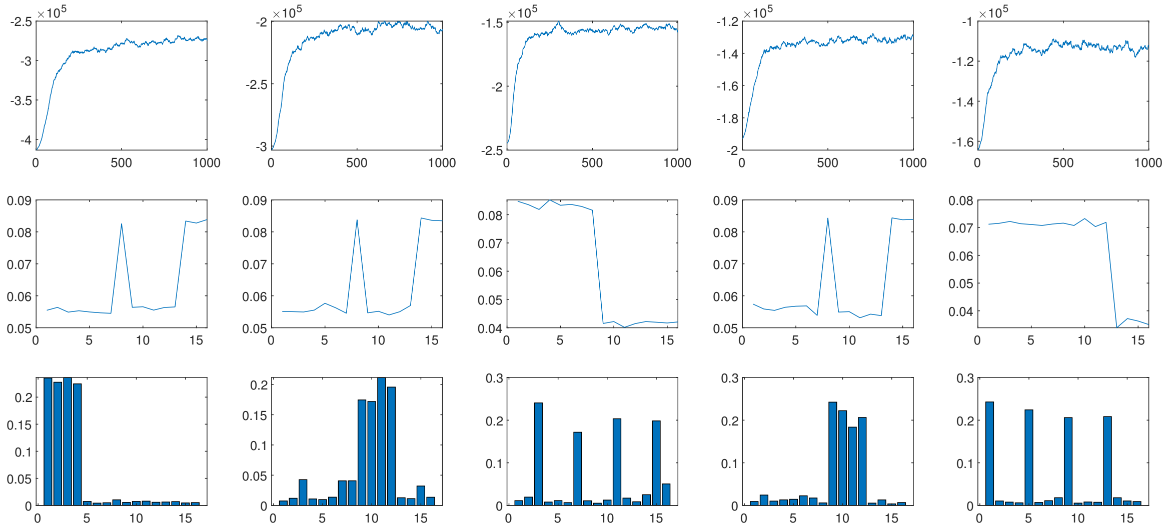

Fig. 6 Five image documents in natural category scene dataset

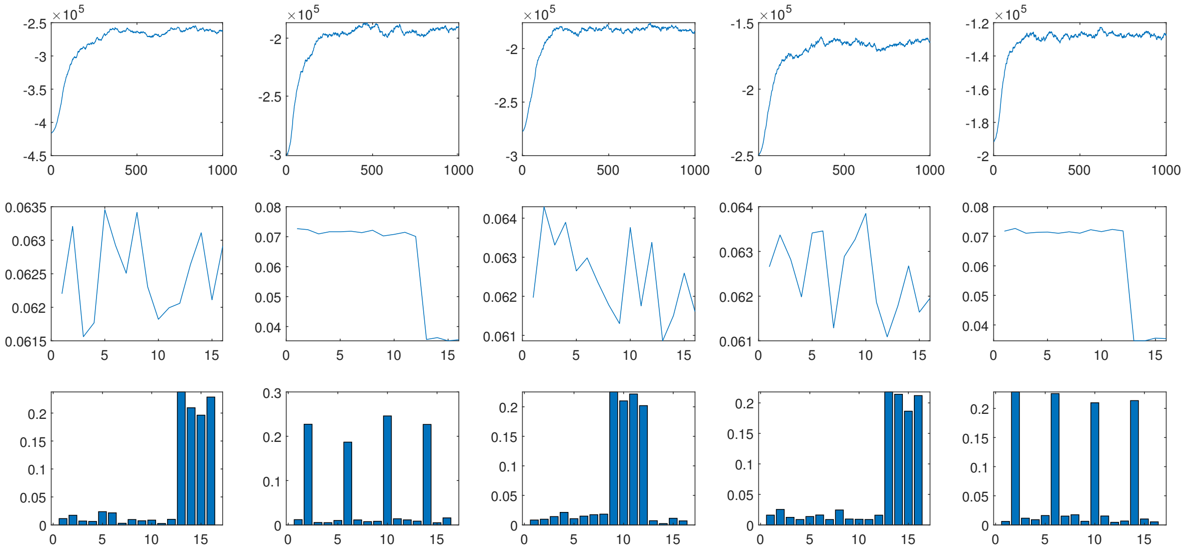

Fig. 7 Analysis of 5 image documents in natural scene category dataset

4.4 COREL dataset

For this second experiment of image classification using our proposed method, we selected the COREL database as illutrated in Fig. 8 from the Corel Photo Gallery [103] for our image classification framework. Over thousands images, the collection contains animals, airplanes, cars, plants, landscape and textures, artistic objects, vehicles, and people. The database has in fact been summarized into 80 categories in total containing 8000 images (100 images per class). Each image in the collection has approximately a size of 325 × 255, in JPEG format. Initially, for feature extraction method, we decided to follow the method implemented in [46] to collect the low frequency features provided by the DCT (Discrete Cosine Transform) from the patches obtained by the sliding window process over the images using MATLAB. These low frequencies in DCT are specialized in capturing relevant characteristics in images. As the generative topic model in our implementation was struggling to be successful with this feature extraction scheme, we decided to use SIFT features similar to the work in [32, 6,5,4] and the one in the previous section in this paper about natural scene categorization. In this work, we used all the 80 categories. The SIFT method and BoW architecture are described in [32,6,5,4] for image representation in feature space.

Once the generative topic models are implemented, we use probabilistic kernels to carry the document topic features to the SVM for classification. The technique ultimately requires the representation of each document in the distribution space to facilitate the work for the probabilistic kernel machine. Then afterwards, we compare the performance of our proposed approach to its competitors in topic modeling. We also maintain an optimal number of topics at $K = 70$ for a vocabulary size of $V = 600$ for the implementation of the generative topic models.

The implementation of our method has shown the performance of the hybrids with nonlinear probabilistic kernels compared to linear probabilistic kernels such as ELK (expected likelihood kernel). From the results (in terms of accuracy) obtained, we can observe that these hybrids performances with ELK were less improved compared to the case of nonlinear kernels such as BK, KLDK, RK, and JSK. This is translated into a low accuracy value for the ELK. The hybrids provided by our proposed generative approach, (*topic model I* and *topic model II*), with the probabilistic kernel-based SVM have demonstrated higher results. The combination *topic model I* and SVM showed an accuracy of 79.83% with the JSK. In overall, these two topic models perform equally within the discriminative setting especially with the JSK.

Similar to the natural scene document modeling case in the previous section, in this COREL dataset also, we randomly selected 5 documents (Figs. 9 and 10). Our proposed topic models were able to show the efficiency of the representation of documents as distributions. Through these distributions characteristics, the documents were able to exhibit their differences. Here, each of these documents (by observing the second row) belongs to a different class as illustrated in Figs. 9 and 10. Our generative models implemented have shown better performance when compared to an RBF-based SVM classifier in the original feature space (*topic model I* with an accuracy of 72.70% and *topic model II* with 70.40%). Implementing the SVM in the original space provided 65.34% as classification accuracy. By using topics, we were able to provide a lower dimemsional space that allows a better compression of the data. The low dimensional space is used to represent the documents.

Table 7 Hybrid models performances for COREL dataset

%	BK	KLDK	RK	JSK	ELK
topic Model I	75.51	76.39	77.98	79.83	67.42
LDA	57.65	60.56	67.43	68.36	55.39
CVB-LDA	60.45	63.78	68.56	69.43	57.54
CVB-LGDA	63.42	65.45	70.12	71.48	58.29
LGDA	62.10	64.33	68.27	70.25	58.87
topic Model II	74.10	74.87	77.21	78.75	70.38

5 Conclusion

In this paper, we demonstrated the effectiveness of documents or data representation (generative features) from the proposed topic generative framework coupled with the implementation of powerful probabilistic kernels-based SVM classifiers that provided good performance in classification. The use of asymmetric GD and BL conjugate priors simultaneously (within the same generative process) in our topic modeling framework led to two models: the CVB-GD-BL-based topic model (*topic model I*) and the CVB-BL-GD-based topic model (*topic model II*). This ultimately characterizes the generative-discriminative setting in our proposed approach. The discretization of the continuous visual feature space due to clustering and quantization schemes for the formation of the visual codebook led to the reformulation of probabilistic kernels from continuous space to discrete space as we deal with empirical (discrete) distributions. Using some challenging datasets in machine learning and computer vision, we are able to extract intrinsic characteristics from text and image documents for the implementation of our hybrid models. Topic representation is an effective summarization method to allow topic models to work in finite dimensional spaces (low dimensional spaces). This automatically presents the advantage of solving memory space (storage) issues in databases. In other words, the space complexity is refined and improved within our proposed framework. The implementation of generative models in the fully collapsed space of latent variables provided a framework (sampling) that allows the computation of probabilistic kernels through empirical likelihood scheme. This setting facilitates the representation and parameterization of our documents (texts and images) as distributions for the kernel machine. This representation has been beneficial for the modeling of our hybrids as documents now have ability to carry effectively local information from generative topic models into discriminative classifiers that operate with distributions. Distributions are always seen as accurate and compact representations of the data since they can efficiently hold some useful properties such as semantics within the observed

Fig. 8 Corel dataset (15 out of 80 categories)

data. This reality is demonstrated in our experiment as we successfully show that despite the performance of standard kernels-based SVMs in the original feature space, probabilistic kernels-based SVMs provide the best performance and results especially when combined with robust topic models. These characteristics illustrate the effectiveness of our hybrid models and their performance within a wide variety of datasets showing therefore the ability for our proposed framework to generalize. The fully collapsed representation was also key to the success of our generative approach by connecting a hybrid inference (the collapsed variational Bayes, seen as one of the state-of-the-art inference techniques in topic modeling with its flexibility to combine both the performance of the variational Bayes and the collapsed Gibbs sampler) to hybrid model (generative-discriminative). The hybrid techniques using CVB-LGDA and CVB-LDA in this generative-discriminative approach have shown better performances compared to the LDA-based hybrids in uncollapsed space. As generalized Dirichlet and Beta-Liouville distributions are more flexible than the Dirichlet, using these priors in topic modeling presents some advantages in the generative-discriminative setting. This ultimately justifies the good performance in our proposed approach as we implement our topic models with these two different priors (asymmetric) used simultaneously within the same generative process. Compared to previous hybrid models, our proposed approaches mostly outperform them in our datasets. As a result, the edge is given to our current proposed

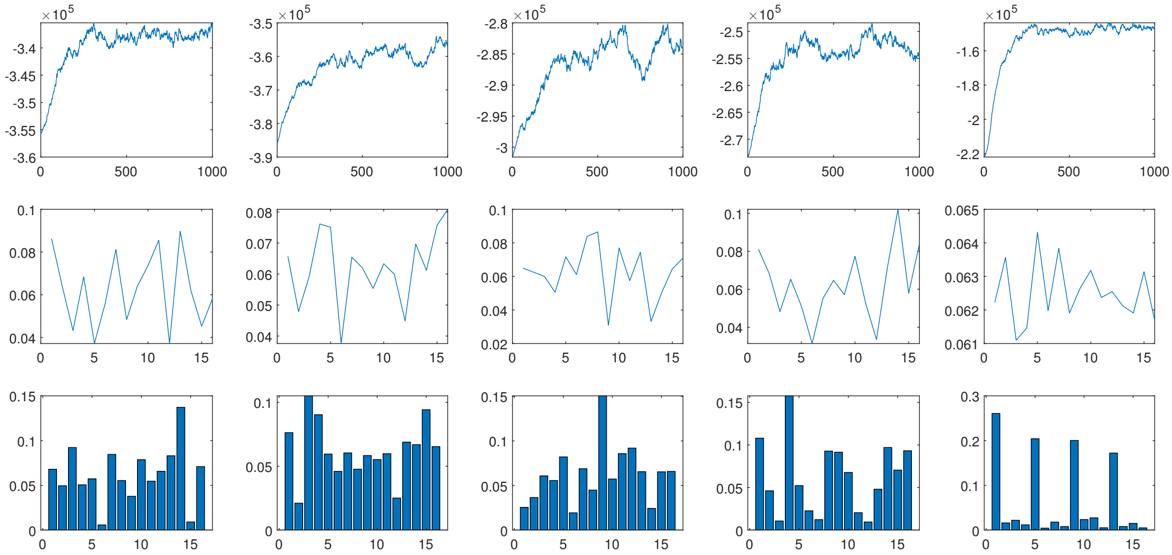

Fig. 9 Analysis of image documents in Corel dataset

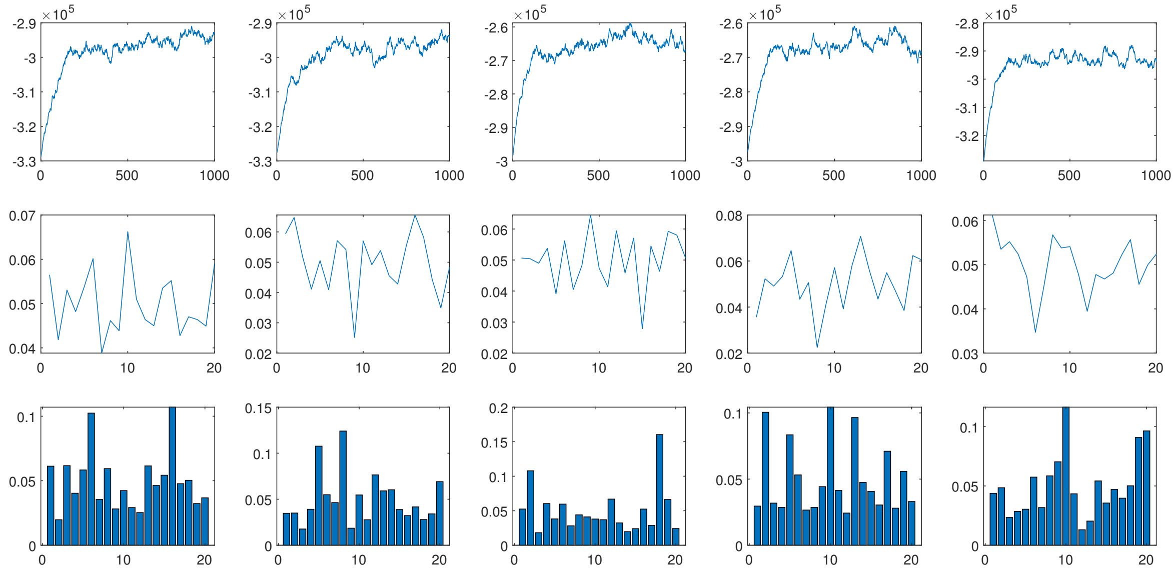

Fig. 10 Characteristics of image documents in Corel dataset

methods. With the right probabilistic kernel, the hybrid methods from topic Models *I* and *II* could also perform almost similarly with the majority of our datasets in a sense that they both provide mostly, robust and coherent generative topic features to the SVM as shown in the performance results compared to their competitors. However, within our proposed methods, the hybrid, *topic model II/SVM* provides a better performance in terms of time complexity in comparison to the hybrid, *topic model I/SVM*. This is mainly due to the intrinsic characteristics of the (asymmetric) Beta-Liouville conjugate prior for the document parameter's modeling besides robustness and flexibility. To its advantage, the distribution (BL) has generally few parameters compared to the GD. As a result, inferences were observed to be faster with the hybrid *topic model II/SVM* as it effectively characterizes or models the document parameter with (asymmetric) BL while also providing robust generative features to the kernel machine. This is in contrast to the hybrid *topic model I/SVM* which samples the document parameter from (asymmetric) GD, and it is observed to be slower in estimations despite its robust performance.

The relationship between our topic generative features and kernel formulations for SVM also demonstrate that our nonlinear probabilistic kernels implemented performed well with images than linear probabilistic kernels such as ELK. Images often provide features that are too complex to be linearly separated. Changes in view and illumination for instance could have impacts on image feature characteristics and therefore on the distributions. Nonlinear probabilistic kernels have ability to adapt to these changes better than linear kernels. On the other hand, text documents classification tends to be well characterized with linear probabilistic kernels. Our models were able exhibit these characteristics through our datasets showing therefore the robustness of the framework. This explains the importance of knowledge about the data as it can influence the choice of the kernel functions in the discriminative framework. Therefore, the strong performance of the JSK (Jensen-Shannon kernel) on our proposed topic models could be explained by the capability of this nonlinear probabilistic kernel in handling and characterizing effectively generative features represented as empirical distributions such as the ones implemented in our topic models. We witnessed, during implementation that the models require many parameter and hyperparameters to initialized. The complexity of the models has been increased. Initialization affect the results. Importantly, our proposed approach remains an alternative to nonparametric models in finite dimensional space (with finite mixtures) for classification. However, as topic models in finite dimensional space always struggle in providing very efficient and accurate model selection criteria, a future work could be about investigating on the possibility to implement a nonparametric model due the high complexity in our datasets. We could also emphasize on inference based on hyperparameter estimation to reduce problems related to poor initializations.

References

1. A. D. Holub, M. Welling, P. Perona, Combining generative models and fisher kernels for object recognition, in: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, Vol. 1, IEEE, 2005, pp. 136–143.
2. A. Y. Ng, M. I. Jordan, On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes, in: Advances in neural information processing systems, 2002, pp. 841–848.
3. R. Nallapati, Discriminative models for information retrieval, in: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2004, pp. 64–71.
4. K. E. Ihou, N. Bouguila, A new latent generalized dirichlet allocation model for image classification, in: Image Processing Theory, Tools and Applications (IPTA), 2017 Seventh International Conference on, IEEE, 2017, pp. 1–6.
5. K. E. Ihou, N. Bouguila, Variational-based latent generalized dirichlet allocation model in the collapsed space and applications, Neurocomputing 332 (2019) 372–395.
6. N. Bouguila, Hybrid generative/discriminative approaches for proportional data modeling and classification, IEEE Transactions on Knowledge and Data Engineering 24 (12) (2012) 2184–2202.
7. N. Bouguila, D. Ziou, A dirichlet process mixture of generalized dirichlet distributions for proportional data modeling, IEEE Transactions on Neural Networks 21 (1) (2010) 107–122.
8. N. Bouguila, Count data modeling and classification using finite mixtures of distributions, IEEE Trans. Neural Networks 22 (2) (2011) 186–198.
9. S. Ullman, M. Vidal-Naquet, E. Sali, Visual features of intermediate complexity and their use in classification, Nature neuroscience 5 (7) (2002) 682.
10. M. Weber, M. Welling, P. Perona, Towards automatic discovery of object categories, in: cvpr, 2000, p. 39.
11. R. Fergus, P. Perona, A. Zisserman, et al., Object class recognition by unsupervised scale-invariant learning, in: CVPR (2), 2003, pp. 264–271.
12. B. Leibe, B. Schiele, Scale-invariant object categorization using a scale-adaptive mean-shift search, in: Joint Pattern Recognition Symposium, Springer, 2004, pp. 145–153.
13. H. Schneiderman, Learning a restricted bayesian network for object detection, CVPR (2) 4 (2004) 639–646.
14. A. S. Bakhtiari, N. Bouguila, A variational bayes model for count data learning and classification, Engineering Applications of Artificial Intelligence 35 (2014) 176–186.
15. A. S. Bakhtiari, N. Bouguila, Online learning for two novel latent topic models, in: Information and Communication Technology: Second IFIP TC 5/8 International Conference, ICT-EurAsia 2014, Bali, Indonesia, April 14-17, 2014, Proceedings, Vol. 8407, Springer, 2014, p. 286.
16. L. Fei-Fei, Learning generative visual models from few training examples, in: Workshop on Generative-Model Based Vision, IEEE Proc. CVPR, 2004.
17. C. M. Bishop, Pattern recognition and machine learning, Springer Science+ Business Media, 2006.
18. C. Yeh, Y. H. Tsai, Y. F. Wang, Generative-discriminative variational model for visual recognition, CoRR abs/1706.02295. arXiv:1706.02295.
19. W. Roth, R. Peharz, S. Tschiatschek, F. Pernkopf, Hybrid generative-discriminative training of gaussian mixture models, Pattern recognition letters 112 (2018) 131–137.
20. W. Zheng, Y. Liu, H. Lu, H. Tang, Discriminative topic sparse representation for text categorization, in: 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Vol. 1, IEEE, 2017, pp. 454–457.
21. T. Jaakkola, D. Haussler, Exploiting generative models in discriminative classifiers, in: Advances in neural information processing systems, 1999, pp. 487–493.
22. T. Jebara, R. Kondor, A. Howard, Probability product kernels, Journal of Machine Learning Research 5 (Jul) (2004) 819–844.

23. N. Vasconcelos, P. Ho, P. Moreno, The kullback-leibler kernel as a framework for discriminant and localized representations for visual recognition, in: European Conference on Computer Vision, Springer, 2004, pp. 430–441.
24. K. Tsuda, M. Kawanabe, G. Rätsch, S. Sonnenburg, K.-R. Müller, A new discriminative kernel from probabilistic models, in: Advances in Neural Information Processing Systems, 2002, pp. 977–984.
25. K. R. Prasad, M. Mohammed, R. Noorullah, Visual topic models for healthcare data clustering, Evolutionary Intelligence (2019) 1–17.
26. L. Xia, D. Luo, C. Zhang, Z. Wu, A survey of topic models in text classification, in: 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), IEEE, 2019, pp. 244–250.
27. H. J. Steinhauer, T. Helldin, G. Mathiason, A. Karlsson, Topic modeling for anomaly detection in telecommunication networks, Journal of Ambient Intelligence and Humanized Computing (2019) 1–12.
28. L. Laib, M. S. Allili, S. Ait-Aoudia, A probabilistic topic model for event-based image classification and multilabel annotation, Signal Processing: Image Communication 76 (2019) 283–294.
29. F. Yao, Y. Wang, Tracking urban geo-topics based on dynamic topic model, Computers, Environment and Urban Systems (2019) 101419.
30. R. Venkatesaramani, D. Downey, B. Malin, Y. Vorobeychik, A semantic cover approach for topic modeling, in: Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 92–102.
31. D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, Journal of machine Learning research 3 (Jan) (2003) 993–1022.
32. L. Fei-Fei, P. Perona, A bayesian hierarchical model for learning natural scene categories, in: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on, Vol. 2, IEEE, 2005, pp. 524–531.
33. Y. Yang, J. Jia, S. Zhang, B. Wu, Q. Chen, J. Li, C. Xing, J. Tang, How do your friends on social media disclose your emotions?, in: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, AAAI Press, 2014, pp. 306–312.
34. L. Yang, M. Qiu, S. Gottipati, F. Zhu, J. Jiang, H. Sun, Z. Chen, Cqarank: jointly model topics and expertise in community question answering, in: Proceedings of the 22nd ACM international conference on Information & Knowledge Management, ACM, 2013, pp. 99–108.
35. B. Leng, J. Zeng, M. Yao, Z. Xiong, 3d object retrieval with multitopic model combining relevance feedback and lda model, IEEE Transactions on Image Processing 24 (1) (2015) 94–105.
36. K. L. Caballero, J. Barajas, R. Akella, The generalized dirichlet distribution in enhanced topic detection, in: Proceedings of the 21st ACM international conference on Information and knowledge management, ACM, 2012, pp. 773–782.
37. J. Foulds, L. Boyles, C. DuBois, P. Smyth, M. Welling, Stochastic collapsed variational bayesian inference for latent dirichlet allocation, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2013, pp. 446–454.
38. B. Ghorbani, H. Javadi, A. Montanari, An instability in variational inference for topic models, in: International Conference on Machine Learning, 2019, pp. 2221–2231.
39. A. Y. Zhang, H. H. Zhou, Theoretical and computational guarantees of mean field variational inference for community detection, arXiv preprint arXiv:1710.11268.
40. A. S. Bakhtiari, N. Bouguila, A latent beta-liouville allocation model, Expert Systems with Applications 45 (2016) 260–272.
41. K. E. Ihou, N. Bouguila, Stochastic topic models for large scale and nonstationary data, Engineering Applications of Artificial Intelligence 88 (2020) 103364.
42. Y. W. Teh, D. Newman, M. Welling, A collapsed variational bayesian inference algorithm for latent dirichlet allocation, in: Advances in neural information processing systems, 2007, pp. 1353–1360.
43. P. Bhagat, P. Choudhary, Image annotation: Then and now, Image and Vision Computing 80 (2018) 1–23.
44. D. Tian, Z. Shi, A two-stage hybrid probabilistic topic model for refining image annotation, International Journal of Machine Learning and Cybernetics (2019) 1–15.
45. W. Fan, N. Bouguila, Learning finite beta-liouville mixture models via variational bayes for proportional data clustering., in: IJCAI, 2013, pp. 1323–1329.
46. P. J. Moreno, P. P. Ho, N. Vasconcelos, A kullback-leibler divergence based kernel for svm classification in multimedia applications, in: Advances in neural information processing systems, 2004, pp. 1385–1392.
47. D. M. Blei, M. I. Jordan, et al., Variational inference for dirichlet process mixtures, Bayesian analysis 1 (1) (2006) 121–144.
48. W. Fan, N. Bouguila, Online data clustering using variational learning of a hierarchical dirichlet process mixture of dirichlet distributions, in: International Conference on Database Systems for Advanced Applications, Springer, 2014, pp. 18–32.
49. H. Zhao, L. Du, W. Buntine, G. Liu, Metalda: a topic model that efficiently incorporates meta information, in: 2017 IEEE International Conference on Data Mining (ICDM), IEEE, 2017, pp. 635–644.
50. P. Kherwa, P. Bansal, Topic modeling: A comprehensive review, ICST Transactions on Scalable Information Systems (2018) 159623.
51. W. Li, A. McCallum, Pachinko allocation: Dag-structured mixture models of topic correlations, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 577–584.
52. L. Liu, H. Huang, Y. Gao, Y. Zhang, X. Wei, Neural variational correlated topic modeling, in: The World Wide Web Conference, ACM, 2019, pp. 1142–1152.
53. G. Xun, Y. Li, W. X. Zhao, J. Gao, A. Zhang, A correlated topic model using word embeddings., in: IJCAI, 2017, pp. 4207–4213.
54. D. Blei, J. Lafferty, Correlated topic models, Advances in neural information processing systems 18 (2006) 147.
55. I. Korshunova, H. Xiong, M. Fedoryszak, L. Theis, Discriminative topic modeling with logistic lda, in: Advances in Neural Information Processing Systems 32, Curran Associates, Inc., 2019, pp. 6767–6777.
56. J. D. Mcauliffe, D. M. Blei, Supervised topic models, in: Advances in neural information processing systems, 2008, pp. 121–128.
57. D. Ramage, D. Hall, R. Nallapati, C. D. Manning, Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1, Association for Computational Linguistics, 2009, pp. 248–256.
58. S. Lacoste-Julien, F. Sha, M. I. Jordan, Disclda: Discriminative learning for dimensionality reduction and

classification, in: Advances in neural information processing systems, 2009, pp. 897–904.

59. A. B. Dieng, F. J. R. Ruiz, D. M. Blei, The dynamic embedded topic model, CoRR abs/1907.05545. arXiv:1907.05545.
60. R. Chi, B. Wu, L. Wang, Expert identification based on dynamic lda topic model, in: 2018 IEEE Third International Conference on Data Science in Cyberspace (DSC), IEEE, 2018, pp. 881–888.
61. D. M. Blei, J. D. Lafferty, Dynamic topic models, in: Proceedings of the 23rd international conference on Machine learning, ACM, 2006, pp. 113–120.
62. J. Chen, J. Zhu, J. Lu, S. Liu, Scalable training of hierarchical topic models, Proceedings of the VLDB Endowment 11 (7) (2018) 826–839.
63. A. Banerjee, I. S. Dhillon, J. Ghosh, S. Sra, Clustering on the unit hypersphere using von mises-fisher distributions, Journal of Machine Learning Research 6 (Sep) (2005) 1345–1382.
64. Y. Li, C. Liu, M. Zhao, R. Li, H. Xiao, K. Wang, J. Zhang, Multi-topic tracking model for dynamic social network, Physica A: Statistical Mechanics and its Applications 454 (2016) 51–65.
65. I. Espinoza, M. Mendoza, P. Ortega, D. Rivera, F. Weiss, Viscovery: Trend tracking in opinion forums based on dynamic topic models, CoRR abs/1805.00457. arXiv:1805.00457.
66. Y. He, C. Lin, W. Gao, K.-F. Wong, Dynamic joint sentiment-topic model, ACM Transactions on Intelligent Systems and Technology (TIST) 5 (1) (2013) 6.
67. J. Fenglei, G. Cuiyun, et al., An online topic modeling framework with topics automatically labeled, in: Proceedings of the 2019 Workshop on Widening NLP, 2019, pp. 73–76.
68. C. Gao, J. Zeng, M. R. Lyu, I. King, Online app review analysis for identifying emerging issues, in: 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE), IEEE, 2018, pp. 48–58.
69. X. Bui, T. Vu, K. Than, Stochastic bounds for inference in topic models, in: International Conference on Advances in Information and Communication Technology, Springer, 2016, pp. 582–592.
70. L. AlSumait, D. Barbará, C. Domeniconi, On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking, in: 2008 eighth IEEE international conference on data mining, IEEE, 2008, pp. 3–12.
71. S. Padó, M. Lapata, Dependency-based construction of semantic space models, Computational Linguistics 33 (2) (2007) 161–199.
72. D. Valdez, A. C. Pickett, P. Goodson, Topic modeling: Latent semantic analysis for the social sciences, Social Science Quarterly 99 (5) (2018) 1665–1679.
73. J. Chang, D. Blei, Relational topic models for document networks, in: Artificial Intelligence and Statistics, 2009, pp. 81–88.
74. D. M. Blei, K. Franks, M. I. Jordan, I. S. Mian, Statistical modeling of biomedical corpora: mining the caenorhabditis genetic center bibliography for genes related to life span, Bmc Bioinformatics 7 (1) (2006) 250.
75. S. Xiong, K. Wang, D. Ji, B. Wang, A short text sentiment-topic model for product reviews, Neurocomputing 297 (2018) 94–102.
76. M. Hajjem, C. Latiri, Combining ir and lda topic modeling for filtering microblogs, Procedia Computer Science 112 (2017) 761–770.
77. M. Fritz, B. Schiele, Decomposition, discovery and detection of visual categories using topic models, in: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008, pp. 1–8.
78. J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, W. T. Freeman, Discovering objects and their location in images, in: Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, Vol. 1, IEEE, 2005, pp. 370–377.
79. R. Fergus, L. Fei-Fei, P. Perona, A. Zisserman, Learning object categories from google's image search.
80. N. Bouguila, Clustering of count data using generalized dirichlet multinomial distributions, IEEE Transactions on Knowledge and Data Engineering 20 (4) (2008) 462–474.
81. N. Bouguila, D. Ziou, J. Vaillancourt, Unsupervised learning of a finite mixture model based on the dirichlet distribution and its application, IEEE Transactions on Image Processing 13 (11) (2004) 1533–1543.
82. T. Hofmann, Probabilistic latent semantic indexing, in: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 1999, pp. 50–57.
83. L. Wu, L. Shen, Z. Li, A kernel method based on topic model for very high spatial resolution (vhsr) remote sensing image classification, ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences XLI-B7 (2016) 399–403.
84. M. Lienou, H. Maitre, M. Datcu, Semantic annotation of satellite images using latent dirichlet allocation, IEEE Geoscience and Remote Sensing Letters 7 (1) (2009) 28–32.
85. Y. Teh, M. Jordan, M. Beal, D. Blei, Hierarchical dirichlet processes, Machine Learning (2006) 1–30.
86. K. Rematas, M. Fritz, T. Tuytelaars, Kernel density topic models: Visual topics without visual words, in: NIPS workshops, Modern Nonparametric Methods in Machine Learning, 2012.
87. V. Nguyen, D. Phung, S. Venkatesh, Topic model kernel classification with probabilistically reduced features, Journal of Data Science 13 (2) (2015) 323–340.
88. P. Hennig, D. Stern, R. Herbrich, T. Graepel, Kernel topic models, in: Artificial Intelligence and Statistics, 2012, pp. 511–519.
89. K. Muandet, K. Fukumizu, F. Dinuzzo, B. Schölkopf, Learning from distributions via support measure machines, in: Advances in neural information processing systems, 2012, pp. 10–18.
90. Y. Yoshikawa, T. Iwata, H. Sawada, Latent support measure machines for bag-of-words data classification, in: Advances in Neural Information Processing Systems, 2014, pp. 1961–1969.
91. T. Bdiri, N. Bouguila, Bayesian learning of inverted dirichlet mixtures for svm kernels generation, Neural Computing and Applications 23 (5) (2013) 1443–1458.
92. K. Than, T. Doan, Guaranteed inference in topic models, arXiv preprint arXiv:1512.03308.
93. H. M. Wallach, D. Mimno, A. McCallum, Rethinking lda: why priors matter, in: Proceedings of the 22nd International Conference on Neural Information Processing Systems, Curran Associates Inc., 2009, pp. 1973–1981.
94. H. M. Wallach, I. Murray, R. Salakhutdinov, D. Mimno, Evaluation methods for topic models, in: Proceedings of the 26th annual international conference on machine learning, ACM, 2009, pp. 1105–1112.

95. A. B. Chan, N. Vasconcelos, P. J. Moreno, A family of probabilistic kernels based on information divergence, Univ. California, San Diego, CA, Tech. Rep. SVCL-TR-2004-1.
96. J. Lin, Divergence measures based on the shannon entropy, IEEE Transactions on Information theory 37 (1) (1991) 145–151.
97. T. Jebara, R. Kondor, Bhattacharyya and expected likelihood kernels, in: Learning theory and kernel machines, Springer, 2003, pp. 57–71.
98. R. Kondor, T. Jebara, A kernel between sets of vectors, in: Proceedings of the 20th International Conference on Machine Learning (ICML-03), 2003, pp. 361–368.
99. J. Zeng, Z.-Q. Liu, X.-Q. Cao, Fast online em for big topic modeling, IEEE Transactions on Knowledge and Data Engineering 28 (3) (2015) 675–688.
100. A. Asuncion, M. Welling, P. Smyth, Y. W. Teh, On smoothing and inference for topic models, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, AUAI Press, 2009, pp. 27–34.
101. X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, in: Advances in neural information processing systems, 2015, pp. 649–657.
102. S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, in: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), Vol. 2, 2006, pp. 2169–2178.
103. J. Z. Wang, J. Li, G. Wiederhold, Simplicity: Semantics-sensitive integrated matching for picture libraries, IEEE Transactions on Pattern Analysis & Machine Intelligence (9) (2001) 947–963.