Latent Beta-Liouville Probabilistic Modeling for Bursty Topic Discovery in Textual Data

Shadan Ghadimi and **Hafsa Ennajari** and **Nizar Bouguila**
CIISE, Concordia University, Montreal, Canada
Emails: s_ghadim@encs.concordia.ca, h_ennaja@encs.concordia.ca, nizar.bouguila@concordia.ca

Abstract

Topic modeling has become a fundamental technique for uncovering latent thematic structures within large collections of textual data. However, conventional models often struggle to capture the burstiness of topics. This characteristic, where the occurrence of a word increases its likelihood of subsequent appearances in a document, is fundamental in natural language processing. To address this gap, we introduce a novel topic modeling framework, integrating Beta-Liouville and Dirichlet Compound Multinomial distributions. Our approach, named Beta-Liouville Dirichlet Compound Multinomial Latent Dirichlet Allocation (BLDCMLDA), is designed to specifically model word burstiness and support a wide range of adaptable topic proportion patterns. Through experiments on diverse benchmark text datasets, the BLDCMLDA model has demonstrated superior performance over conventional models. Our promising results in terms of perplexity and coherence scores demonstrate the effectiveness of BLDCMLDA in capturing the nuances of word usage dynamics in natural language.

Introduction

In the modern era, a vast amount of data is generated across various fields. when properly handled, this data is a valuable source of information (Bouguila 2007). Topic modeling has emerged as a crucial tool for efficiently processing large text datasets. They are adept at uncovering key themes across numerous documents (Bakhtiari and Bouguila 2014a) (Bakhtiari and Bouguila 2014b). Models like Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) identify word clusters, or topics, that frequently appear together, offering a deeper understanding of document content beyond just single words. This method enables a more profound semantic interpretation by focusing on the overarching topics within documents.

The concept of "Burstiness" in language, initially identified by Church, Gale, and Katz (Doyle and Elkan 2009) (Madsen, Kauchak, and Elkan 2005), is inherent in document analysis and topic modeling. It describes the tendency of a rare word to reappear multiple times in a document once it occurs. Beyond text, burstiness is also observed in fields like finance and computer vision (Blei and Lafferty 2007). It is important to distinguish between word burstiness (i.e., the recurrence of specific words in a document) and topic burstiness (i.e., the repetition of topics within a document corpus), as both types play a vital role in analyzing documents and their structure in topic modeling.

Traditional topic models (Blei, Ng, and Jordan 2003) (Das, Zaheer, and Dyer 2015), such as those based on Dirichlet distribution, use basic statistical methods to model word distributions across topics. However, they often struggle to accurately identify new topics, leading to vague or ambiguous interpretations. This issue is mainly due to the inflexibility of their statistical foundations, which are not suited to the dynamic nature of topic trends. As a result, these models are less effective in representing topic burstiness, often producing less clear topics.

In this paper, we introduce the Beta-Liouville Dirichlet Compound Multinomial Latent Dirichlet Allocation (BLDCMLDA) model. This novel topic modeling approach integrates the Beta-Liouville distribution (Bouguila 2012a) (Fan and Bouguila 2015) to overcome the limitations of the Dirichlet distribution priors by allowing for greater flexibility in covariance structure, crucial for capturing the nuances of word burstiness. Our model enhances the adaptability in modeling topic proportions, paving the way for more accurate and coherent topic modeling.

Our contributions to this paper are as follows:

- We propose the BLDCMLDA model, an innovative approach to topic modeling that effectively addresses both word and topic burstiness.
- We demonstrate the superiority of the Beta-Liouville priors in capturing the complex dynamics of topic burstiness, leading to more accurate topic modeling.
- Through extensive experiments on various text datasets, we show that the BLDCMLDA model achieves better semantic coherence and lower perplexity scores compared to traditional models.
- We present comprehensive analyses indicating that BLDCMLDA outperforms existing models in predicting text samples across different topic settings.

The paper is structured as follows: Section reviews relevant literature. Section details the BLDCMLDA model.

Section discusses our experimental results, and Section concludes the paper and explores future research directions.

Background and Related Works

Topic modeling (Blei 2012; Vayansky and Kumar 2020), a technique rooted in generative probabilistic modeling (Liu et al. 2016), is used for identifying hidden topics within textual documents. This method relies on a probabilistic relationship between observable variables and latent parameters for unsupervised analysis of large datasets. It began with Latent Semantic Indexing/Analysis (LSI/LSA) (Landauer and Dumais 2008) which used singular value decomposition for more effective data compression, while Probabilistic LSI/LSA (pLSI/pLSA) by Hofman (Hofmann 2017) introduced a probabilistic approach, enhancing data reduction but with potential overfitting issues. Modern topic models largely employ Bayesian modeling to discover latent patterns in text data (Blei, Ng, and Jordan 2003).

In natural language processing, topic burstiness refers to the sudden and notable increase in certain terms or topics within a text corpus. Analyzing topic burstiness helps in tracking information flow, spotting trends, and understanding topic evolution over time in large datasets. Identifying these bursts and their timing can shed light on public discourse, reveal key events, highlight emerging trends, and enhance our grasp of temporal dynamics. This analysis is particularly useful in areas like social media analytics, news monitoring, and historical document examination, where detecting topical bursts offers critical context for timely and informed decision-making (Kleinberg 2002).

Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003) is a method used to uncover hidden patterns within text corpora and it assumes that each document comprises a mixture of topics. Document generation involves first selecting topics and their corresponding word distributions, and secondly, choosing words from these distributions to form a varied document. The LDA model relies on two key Dirichlet hyperparameters, α and β, which shape its generative process. α influences the topic mixture in documents (θ), with higher values promoting topic similarity and lower ones enhancing diversity. Conversely, β affects the word distribution across topics (ϕ), where a higher β encourages topic similarity, and a lower β leads to more diverse word patterns. Despite not accounting for word burstiness, LDA adequately represents documents through topic distributions (θ), useful for document classification, similarity analysis, and other textual analysis applications(Das, Zaheer, and Dyer 2015).

Dirichlet Compound Multinomial

The Dirichlet Compound Multinomial (DCM) model is designed for text analysis, emphasizing the capture of word burstiness in documents. Differing from Latent Dirichlet Allocation and its related models, DCM focuses on word distribution within each document rather than on topics.

DCM creates each document by choosing a specific multinomial distribution from a common Dirichlet distribution, resulting in documents with words from a distribution that represents a part of a larger topic. This differs from LDA's approach to associating subtopics with documents.

DCM's ability to vary between multinomial and Dirichlet parameters helps it adjust for burstiness, lessening the impact of repeated words as Dirichlet parameters shift. While DCM excels in representing a primary topic with various subtopics, it may not perform as well with documents containing multiple distinct topics. Nevertheless, DCM offers insightful perspectives on word distribution and frequency in documents where word burstiness is prevalent (Huang et al. 2020).

Beta-Liouville

The Beta-Liouville (BL) distribution, part of the Liouville family of distributions, offers a flexible framework for count data modeling. It is defined in a multidimensional setting with positive parameters and is characterized by a generative density function. This distribution stands out for its ability to model a wider range of covariance structures, both positive and negative, unlike the more restrictive Dirichlet distribution. Its flexibility makes it well-suited for applications that involve counting data with complex underlying patterns, such as text categorization or image classification. The Beta-Liouville distribution's unique properties allow it to effectively capture the nuances in data, offering a more accurate and versatile approach to modeling count data compared to traditional methods (Bouguila 2011) (Luo et al. 2023).

Proposed Model

In this section, we will present the mathematical structure and essential aspects of the BLDCMLDA model, including a detailed explanation of the generative process and the method for learning the model parameters.

Model Definition

The proposed BLDCMLDA model has a solid probabilistic foundation, integrating flexible priors including Beta-Liouville and Dirichlet Compound Multinomial. This combination creates a versatile approach for modeling topic burstiness specific to individual documents. The structure of the BLDCMLDA model is depicted in Figure 1.

The Beta-Liouville Dirichlet Compound Multinomial Latent Dirichlet Allocation model combines the Beta-Liouville distribution and Dirichlet Compound Multinomial distribution to enhance the precision and adaptability of representing topic proportions within a document. The Beta-Liouville distribution encompasses the Dirichlet distribution as a particular instance within its framework.

The generative process of BLDCMLDA is outlined in Algorithm 1. The probabilistic assumptions of the proposed model latent variables are described as follows:

$$\theta \sim \text{Beta-Liouville Distribution}(\vec{\delta})$$
$$z \sim \text{Multinomial}(\theta)$$
$$\phi \sim \text{Dirichlet}(\beta)$$

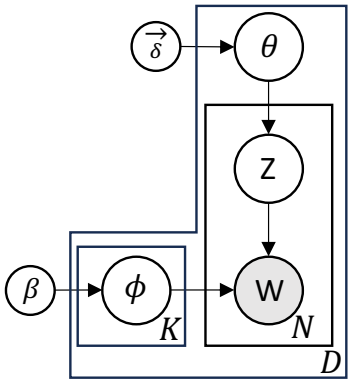


Figure 1: Graphical Model of BLDCMLDA.

Parameter Inference

The BLDCMLDA model with its multiple hidden parameters, necessitates the computation of a posterior distribution that is analytically complex. To address this, we utilize Gibbs sampling (Griffiths and Steyvers 2004), a Markov chain Monte Carlo method that iteratively samples from the conditional distributions of latent variables, aiding in approximating the posterior distribution and estimating model parameters more efficiently. Our model includes unobservable variables: $\vec{\delta}$, β, ϕ, θ, and z. These are split into per-document or per-word parameters (ϕ, θ, z) and hyperparameters ($\vec{\delta}$ and β). During the training phase with document sets, the goal is to find the optimal values for these variables. This involves alternately optimizing the topic parameters (ϕ, θ, z) while keeping the hyperparameters ($\vec{\delta}$ and β) fixed, and then optimizing the hyperparameters based on the refined topic parameters. In situations where hyperparameters are constant, collapsed Gibbs sampling determines the distribution of z in documents, enabling straightforward calculation of ϕ and θ. Additionally, Monte Carlo expectation maximization is used to find the values of $\vec{\delta}$ and β that maximize the likelihood of the training documents, based on the z samples.

Algorithm 1 BLDCMLDA Generative Model

for `document` $d \in \{1, 2, ..., D\}$ **do**
  Draw topic distribution $\theta_d \sim BL(\vec{\delta})$
  for `topic` $k \in \{1, 2, ..., K\}$ **do**
    Draw Topic-Word distribution $\phi_{kd} \sim Dir(\beta_k)$
  end for
  for word w_{dn} in document d, $n \in \{1, ..., n_d\}$ **do**
    Draw topic $z_{dn} \sim \theta_d$
    Draw word $w_{dn} \sim \phi_{z_{dn},d}$
  end for
end for

Following Heinrich et al. (Heinrich 2009), we developed a Gibbs sampling method for efficiently estimating the hidden parameters of the BLDCMLDA model. Initially, we break down the complete likelihood of the model in the following manner:

$$p(w, z|\vec{\delta}, \beta..) = p(w|z, \beta)p(z|\vec{\delta}) \tag{1}$$

The initial probability represents the mean across all potential distributions of ϕ.

$$\begin{aligned} p(w|z, \beta..) &= \int_{\phi} p(z|\phi)p(\phi|\beta)d\phi \\ &= \int_{\phi} p(\phi|\beta) \prod_{d} \prod_{n=1}^{N_d} \phi_{w_{dn} z_{dn}} d\phi \\ &= \int_{\phi} p(\phi|\beta) \prod_{dkt} (\phi_{tkd})^{n_{tkd}} d\phi \end{aligned} \tag{2}$$

Expressing $p(\phi|\beta)$ as a Dirichlet distribution can be written as:

$$\begin{aligned} p(w|z, \beta..) &= \int_{\phi} \left[\prod_{d,k} \frac{1}{B(\beta_{.k})} \prod_{t} (\phi_{tkd})^{\beta_{tk}-1}] \times [\prod_{d,k,t} (\phi_{tkd})^{n_{tkd}} \right] d\phi \\ &= \prod_{d,k} \int_{\phi} \prod_{t} (\phi_{tkd})^{|\beta tk-1+n_{tkd}} d\phi \\ &= \prod_{d,k} \frac{B(n_{.kd} + \beta_{.k})}{B(\beta_{.k})} \end{aligned} \tag{3}$$

In the above equation, $B(\cdot)$ denotes the multivariate Beta function. This function is applied in combination with the count of occurrences of word t associated with topic k in document d, which is indicated by n_{tkd}.

The Beta-Liouville distribution defined in a K-dimensional simplex is characterized by the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$, subject to the constraint $\sum_{k=1}^{K} \theta_k = 1$. Complemented by a hyperparameter vector $\vec{\boldsymbol{\delta}} = (\alpha_1, \alpha_2, \ldots, \alpha_K, \alpha, \gamma)$, it offers more precise control over the distribution shape and scale. The probability density function is formulated as(Bouguila 2012b) (Fan and Bouguila 2013):

$$\begin{aligned} p(\boldsymbol{\theta}|\vec{\boldsymbol{\delta}}) = & \frac{\Gamma\left(\sum_{k=1}^{K-1} \alpha_k\right) \Gamma(\alpha + \gamma)}{\Gamma(\alpha)\Gamma(\gamma)} \prod_{k=1}^{K-1} \frac{\theta_k^{\alpha_k - 1}}{\Gamma(\alpha_k)} \\ & \times \left(\sum_{k=1}^{K-1} \theta_k \right)^{\alpha - \sum_{k=1}^{K-1} \alpha_k} \left(1 - \sum_{k=1}^{K-1} \theta_k \right)^{\gamma - 1} \end{aligned} \tag{4}$$

To infer z, we establish the Gibbs sampling function in the following manner:

$$p(z_i|z^{-i}, w, \vec{\delta}, \beta) = \frac{p(w|z, \beta)p(z|\vec{\delta})}{p(w|z^{-i}, \beta)p(z^{-i}|\vec{\delta})} \tag{5}$$

Hyperparameter EM.

Once the topic parameters are established, the next step involves optimizing the hyperparameters (δ, β) using a Monte Carlo expectation-maximization (EM) technique. This method entails an iterative process of adjusting δ and β values to maximize the likelihood of the training documents.

Earlier studies often used fixed, uniform priors for topic mixtures θ and vocabulary distributions ϕ, with constant parameter values. However, Wallach et al. (Wallach, Mimno,

and McCallum 2009) suggested that asymmetric Dirichlet priors for topic probabilities improve model fitting. In BLDCMLDA, we adopt a novel approach by estimating the BL distribution parameters δ to reveal topic correlations and the parameters for word distributions in topics (β). Directly maximizing the likelihood $p(w|\delta, \beta)$ for data w and hyperparameters δ and β is computationally challenging. We address this by augmenting the likelihood to $p(w, z|\delta, \beta)$ and applying the Monte Carlo Expectation Maximization (MCEM) technique. This involves the Gibbs sampling step for estimating topic assignments (E-step) and optimizing $p(w, z|\delta, \beta)$ (M-step), detailed in Algorithm 2. For β, we maximize the joint distribution based on the expected topic assignments $z = E(w|z, \beta)$ estimated through Gibbs sampling.

Accordingly, we derive the optimal function as follows:

$$\begin{aligned}\beta_{.k}^{new} = \arg\max_{\beta} \sum_{d,t} (\log\Gamma(n_{tkd} + \beta_{tk}) - \log\Gamma(\beta_{tk})) \\ + \sum_{d} [\log\Gamma(\sum_{t} \beta_{tk}) - \log\Gamma(\sum_{t} n_{tkd} + \beta_{tk})]\end{aligned} \tag{6}$$

In this work, we adopt Minka Newton-based methodology (Minka 2012) for fitting the Dirichlet Compound Multinomial distribution. This process involves adjusting the distribution based on K observed vectors, each of a V-dimensional space.

Similarly, the parameters of the Beta-Liouville distribution (Bakhtiari and Bouguila 2016) (Ihou and Bouguila 2020) are determined by maximizing the joint probability distribution:

$$\begin{aligned}\delta^{new} = \arg\max_{\delta} \frac{\Gamma(\sum_{k=1}^{K} a_k)\Gamma(\alpha+\gamma)}{\Gamma(\alpha)\Gamma(\gamma)} \\ \times \int \prod_{k=1}^{K} \frac{\theta_k^{m_k+\alpha_k-1}}{\Gamma(\alpha_k)} (\sum_{k=1}^{K} \theta_k)^{\alpha - \sum_{k=1}^{K} \alpha_k} \\ \times (1 - \sum_{k=1}^{K} \theta_k)^{\gamma-1} d\theta\end{aligned}$$

After estimating the optimal parameters $\delta^{new} = \{\alpha_1^{new}, \alpha_2^{new}, ..., \alpha_k^{new}, \alpha^{new}, \gamma^{new}\}$ through Algorithm 2 and considering the word-topic observations (w, z), we can compute the predictive distribution for a given document d, denoted as $\hat{\theta}_d$.

$$\begin{aligned}\hat{\theta}_{d,k} = \\ \frac{\prod_{i=1}^{K-1}(\alpha + \sum_{k=1}^{K-1} m_{k,\not d} + i - 1)(\gamma + m_{K,\not d})}{\prod_{i=1}^{K}(\sum_{k=1}^{K}(\alpha_k + m_{k,\not d}) + i - 1) \prod_{i=1}^{N_d}(n_{k,\not d} + V\beta + i - 1)} \\ \frac{\prod_{k=1}^{K}(\alpha_k + m_{k,\not d}) \prod_{w=1}^{V}(n_{k,\not d}^{(w)} + \beta}{\prod_{i=1}^{K}(\alpha + \sum_{k=1}^{K-1} m_{k,\not d} + \gamma + m_{K,\not d} + i - 1)}\end{aligned} \tag{7}$$

for the topics $k = 1...K$ and the documents $d = 1...D$.

The probability of words given topics, $\hat{\phi}_k$, can be calculated using the following predictive distribution.

$$\hat{\phi}_{tkd} = \frac{\bar{n}_{w_i z_i d_i} + \beta^*_{w_i z_i} - 1}{\sum_t \bar{n}_t z_i d_i + \beta^*_{t z_i} - 1} \tag{8}$$

It is important to recognize that the likelihood of specific topics in a document, denoted by $\hat{\theta}_d$, varies based on the document itself. Conversely, the likelihood of words within a given topic, indicated by $\hat{\phi}_k$, remains constant. Therefore, when estimating the topic distribution of a new, unseen document, we must account for the document unique topic probabilities while maintaining consistent probabilities for words about their topics.

Algorithm 2 Monte Carlo EM

1: Initialize the parameters δ, β and z
2: **repeat**
3: Run Gibbs Sampling
4: Choose a specific topic assignment for each word using the Gibbs sampling equation
5: Choose δ and β that maximize complete Likelihood $p(w, z|\delta, \beta)$
6: **until** convergence δ, β
7: Choose topic assignment z^* with highest probability
8: Set $\delta^* = \delta, \beta^* = \beta$ **return** δ^*, β^*, z^*

Experimental Results

We conducted comprehensive experiments to assess the BLDCMLDA model's effectiveness in identifying coherent topics and its predictive accuracy. This involved using three public datasets: NIPS, Movie Review, and 20 Newsgroup. The model was compared with GDCMLDA and DCMLDA, which also account for burstiness, offering a relevant benchmark. A notable advantage of Beta-Liouville distributions, as used in our model, is their fewer parameters compared to generalized Dirichlet distributions.

Datasets

Our BLDCMLDA model was tested on three datasets, each with its own set of unique features and challenges. The following is a brief overview of these datasets:

- The NIPS dataset consists of 1740 documents, mostly comprised of papers presented at the NeurIPS conference (formerly known as NIPS), which focuses on Neural Information Processing Systems. These documents cover a period from the first conference in 1987 up to the 2016 conference.
- The Movie Review dataset is commonly used in natural language processing and sentiment analysis studies and comprises textual film reviews. It includes a balanced collection of 2000 reviews, with an equal split of 1000 negative and 1000 positive reviews.
- The 20 Newsgroups dataset comprises around 20,000 documents from newsgroups, evenly distributed across 20 distinct topics. Some newsgroups share close relations, such as comp.sys.ibm.pc.hardware and comp.sys.mac.hardware, while others, like misc.forsale and soc.religion.christian, are markedly different.

Table 1: Examples of topics learned by BLDCMLDA, GDCMLDA, and DCMLDA on the Movie Review dataset.

Horror	Family	Hollywood	Relationships	Fairy Tale	Sci-Fi
BLDCMLDA					
Killed	life	movie	love	story	space
horror	young	scene	wife	disney	planet
killer	character	audience	man	faith	mars
genocide	man	director	children	magic	earth
characters	mother	John	friend	lord	space
scream	woman	role	girl	princess	planet
bad	love	plot	husband	action	alien
GDCMLDA					
horror	life	hollywood	relationship	tale	planet
dead	home	action	love	magic	earth
murder	mother	role	girl	princess	space
kill	father	star	friend	disney	star
wild	love	movie	live	faith	sci-fi
prison	son	story	good	legend	world
movie	woman	work	time	lord	fiction
DCMLDA					
movie	good	watch	great	story	space
horror	character	movie	life	life	earth
killer	young	paul	time	children	alien
scream	lot	director	role	man	movie
characters	written	friend	movie	love	special
sequel	familiar	time	love	disney	planet
time	year	dead	wife	tale	crew

We selected these datasets for their variety and complexity, allowing a robust evaluation of BLDCMLDA. Our preprocessing involved converting all texts to lowercase, tokenizing sentences, and eliminating stop words, punctuation, and words that appear fewer than five times in the corpus.

Topic Coherence

It is crucial to generate topics that are semantically relevant and clear, as they offer deeper insights into the underlying structure of the dataset.

Table 1 displays six topics with associated keywords from the BLDCMLDA, GDCMLDA, and DCMLDA models. Based on their keywords, BLDCMLDA makes it easier to understand these topics. However, GDCMLDA and DCMLDA have unrelated words which make it harder to understand the topics clearly, as they add complexity and obscure the main theme.

Our model effectively pinpointed distinct themes, like the 'Horror' topic with specific words ('genocide', 'scream', 'horror'), contrasting with the DCMLDA model's more general terms ('time'). Similarly, for the 'Relationships' topic, our model captured precise terms ('love', 'wife', 'man', 'children'), reflecting family and emotional aspects, unlike DCMLDA's broader, less-focused approach. This highlights our model's superior ability in identifying and differentiating thematic content in movie reviews, proving its effectiveness across various genres and subjects.

Table 2: Mean coherence scores of the DCMLDA, GDCMLDA, and BLDCMLDA methods.

Dataset	DCMLDA	GDCMLDA	BLDCMLDA
NIPS	0.15	0.34	0.38
20NewsGroups	0.199	0.272	0.37
MovieReview	0.065	0.091	0.104

Our BLDCMLDA method's interpretability was compared to GDCMLDA and DCMLDA using the topic coherence measure (Newman et al. 2010)(Nikolenko, Koltcov, and Koltsova 2017). This metric evaluates how closely the top words in each topic are semantically related, reflecting topic quality. The higher the coherence score, the greater the relevance and connection between the top words. Based on the top 10 words identified by each model, the overall coherence of each model was determined by averaging these scores, facilitating standardized comparisons of topic quality.

The results in Table 2 illustrate BLDCMLDA's superiority over GDCMLDA and DCMLDA in mean coherence score, demonstrating its enhanced capability to generate more meaningful and semantically coherent topics by better understanding semantic word links in the datasets.

The results highlight the substantial promise of the BLDCMLDA method in topic modeling applications. The improved topic coherence indicates that our method can yield more understandable and meaningful topics, a crucial benefit for various use cases.

Perplexity

We also assessed all methods using the perplexity measure, a standard metric in evaluating probabilistic topic models. Perplexity evaluates how well a model predicts a sample, with lower scores indicating better prediction capability. It is calculated inversely to the log-likelihood of the test data. A lower perplexity score suggests a more accurate model in sample prediction.

Our experiments involved training models with varying numbers of topics, specifically 10, 20, 30, 40, and 50. For each configuration, we computed the perplexity score across all topics to assess the performance of the model.

The results of our experiments are displayed in figure 2, showcasing the perplexity scores for DCMLDA, GDCMLDA, and BLDCMLDA across various topic numbers and datasets. In the NIPS dataset, BLDCMLDA consistently shows superiority over GDCMLDA and DCMLDA by achieving lower perplexity scores, indicating a more accurate data representation. This trend is also evident in the 20 NewsGroups and Movie Review datasets, where BLD-

CMLDA surpasses the other methods in predicting test document words, as reflected by its lower perplexity scores.

In Figure 3, the learning curves of BLDCMLDA based on the NIPS, 20 NewsGroups, and Movie Review datasets are shown through perplexity scores. The graph illustrates a consistent decrease in perplexity across all datasets, indicating improved learning with each iteration.

The results demonstrate the efficacy and dependability of the BLDCMLDA method, evident in its consistently high performance across various datasets. Notably, the reduced perplexity scores achieved by BLDCMLDA underscore its capability to effectively discern latent structures and patterns in complex text data.

Topic Diversity

Topic Diversity is an important metric for assessing the quality of the inferred topics. It measures the degree to which information in the topics does not overlap, with a higher score signifying a wider range of topics and, consequently, a more comprehensive semantic coverage of the dataset.

The evaluation of Topic Diversity for the top 10 words in each topic was conducted for BLDCMLDA, GDCMLDA, and DCMLDA methods using these steps:

1. Identify the top 10 words for every topic.
2. Create a unique word set by uniting the top 10 words from all topics.
3. Calculate Topic Diversity as the ratio of the count of unique words to the total word count.

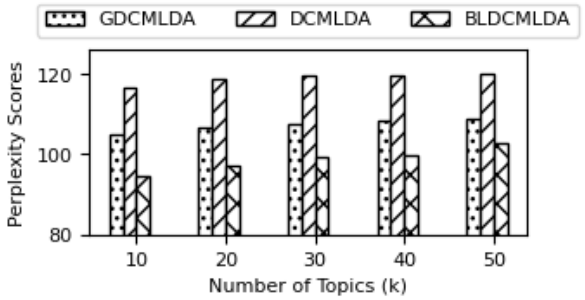
The experiment's findings, detailed in Table 3, show that the BLDCMLDA method outperformed both the GDCMLDA and the standard DCMLDA method in terms of Topic Diversity. This indicates BLDCMLDA's greater effectiveness in representing textual data diversity, as it generated topics encompassing a broader set of unique words, thereby offering a more extensive coverage of the semantic space.

Table 3: Topic Diversity scores of DCMLDA, GDCMLDA, and BLDCMLDA.

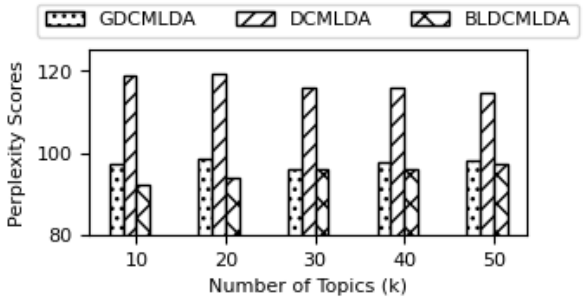
Dataset	DCMLDA	GDCMLDA	BLDCMLDA
NIPS	0.52	0.63	0.67
20NewsGroups	0.72	0.78	0.80
MovieReview	0.38	0.52	0.54

Conclusion

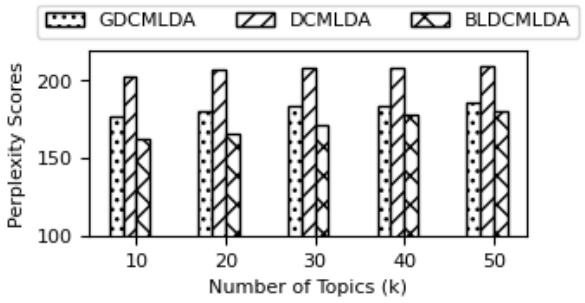
In this work, we present the Beta-Liouville Dirichlet Compound Multinomial Latent Dirichlet Allocation (BLDCMLDA) model. This new approach effectively addresses topic burstiness in text data, demonstrating significant improvements in coherence, perplexity, and topic diversity, thus capturing complex data structures and patterns more accurately. Future developments aim to evolve BLDCMLDA into a non-parametric model using the hierarchical Dirichlet process (Teh et al. 2006) (Fan, Yang, and Bouguila 2022). This will allow for automatic determination of the optimal number of topics, facilitating greater adaptability to diverse datasets without manual topic configuration.



(a) NIPS dataset



(b) 20 NewsGroups dataset



(c) Movie Review dataset

Figure 2: Perplexity scores for the DCMLDA, GDCMLDA, and BLDCMLDA models across varying topic counts on the three datasets.

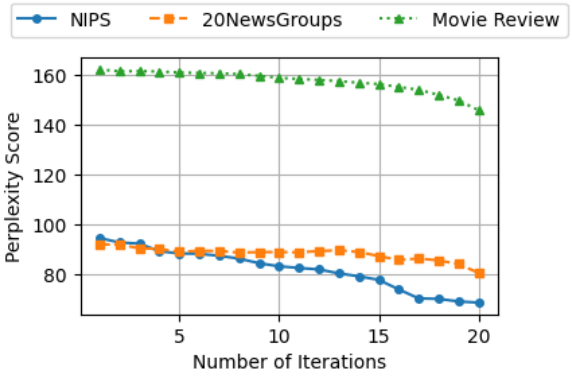


Figure 3: Perplexity Score Trends Over 20 Iterations for Three Different Datasets.

References

Bakhtiari, A. S., and Bouguila, N. 2014a. Online learning for two novel latent topic models. In Linawati; Mahendra, M. S.; Neuhold, E. J.; Tjoa, A. M.; and You, I., eds., *Information and Communication Technology - Second IFIP TC5/8 International Conference, ICT-EurAsia 2014, Bali, Indonesia, April 14-17, 2014. Proceedings*, volume 8407 of *Lecture Notes in Computer Science*, 286–295. Springer.

Bakhtiari, A. S., and Bouguila, N. 2014b. A variational bayes model for count data learning and classification. *Eng. Appl. Artif. Intell.* 35:176–186.

Bakhtiari, A. S., and Bouguila, N. 2016. A latent beta-liouville allocation model. *Expert Systems with Applications* 45:260–272.

Blei, D. M., and Lafferty, J. D. 2007. A correlated topic model of science. *The Annals of Applied Statistics* 1:17–35.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3(Jan):993–1022.

Blei, D. M. 2012. Probabilistic topic models. *Communications of the ACM* 55(4):77–84.

Bouguila, N. 2007. Spatial color image databases summarization. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, volume 1, I–953–I–956.

Bouguila, N. 2011. Count data modeling and classification using finite mixtures of distributions. *IEEE Transactions on Neural Networks* 22(2):186–198.

Bouguila, N. 2012a. Hybrid generative/discriminative approaches for proportional data modeling and classification. *IEEE Transactions on Knowledge and Data Engineering* 24(12):2184–2202.

Bouguila, N. 2012b. Infinite liouville mixture models with application to text and texture categorization. *Pattern Recognit. Lett.* 33(2):103–110.

Das, R.; Zaheer, M.; and Dyer, C. 2015. Gaussian lda for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 795–804.

Doyle, G., and Elkan, C. 2009. Accounting for burstiness in topic models. In Danyluk, A. P.; Bottou, L.; and Littman, M. L., eds., *Proceedings of the 26th Annual International Conference on Machine Learning, ICML 2009, Montreal, Quebec, Canada, June 14-18, 2009*, volume 382 of *ACM International Conference Proceeding Series*, 281–288. ACM.

Fan, W., and Bouguila, N. 2013. Learning finite beta-liouville mixture models via variational bayes for proportional data clustering. In Rossi, F., ed., *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, 1323–1329. IJCAI/AAAI.

Fan, W., and Bouguila, N. 2015. Expectation propagation learning of a dirichlet process mixture of beta-liouville distributions for proportional data clustering. *Eng. Appl. Artif. Intell.* 43:1–14.

Fan, W.; Yang, L.; and Bouguila, N. 2022. Unsupervised grouped axial data modeling via hierarchical bayesian nonparametric models with watson distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44(12):9654–9668.

Griffiths, T. L., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National academy of Sciences* 101(suppl_1):5228–5235.

Heinrich, G. 2009. Parameter estimation for text analysis.

Hofmann, T. 2017. Probabilistic latent semantic indexing. *SIGIR Forum* 51(2):211–218.

Huang, R.; Xu, W.; Qin, Y.; and Chen, Y. 2020. Hierarchical dirichlet multinomial allocation model for multi-source document clustering. *IEEE Access* 8:109917–109927.

Ihou, K. E., and Bouguila, N. 2020. Stochastic topic models for large scale and nonstationary data. *Engineering Applications of Artificial Intelligence* 88:103364.

Kleinberg, J. 2002. Bursty and hierarchical structure in streams. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 7.

Landauer, T. K., and Dumais, S. T. 2008. Latent semantic analysis. *Scholarpedia* 3:4356.

Liu, L.; Tang, L.; Dong, W.; Yao, S.; and Zhou, W. 2016. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* 5.

Luo, Z.; Amayri, M.; Fan, W.; and Bouguila, N. 2023. Cross-collection latent beta-liouville allocation model training with privacy protection and applications. *Appl. Intell.* 53(14):17824–17848.

Madsen, R. E.; Kauchak, D.; and Elkan, C. 2005. Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, 545–552. New York, NY, USA: Association for Computing Machinery.

Minka, T. P. 2012. Estimating a dirichlet distribution.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 100–108.

Nikolenko, S. I.; Koltcov, S.; and Koltsova, O. 2017. Topic modelling for qualitative studies. *Journal of Information Science* 43(1):88–102.

Teh, Y. W.; Jordan, M. I.; Beal, M. J.; and Blei, D. M. 2006. Hierarchical dirichlet processes. *Journal of the American Statistical Association* 101(476):1566–1581.

Vayansky, I., and Kumar, S. A. 2020. A review of topic modeling methods. *Information Systems* 94:101582.

Wallach, H.; Mimno, D.; and McCallum, A. 2009. Rethinking lda: Why priors matter. In Bengio, Y.; Schuurmans, D.; Lafferty, J.; Williams, C.; and Culotta, A., eds., *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.