Adaptive Priors in Probabilistic Topic Models for Bursty Discovery in Textual Data

Shadan Ghadimigheshlaghi

A Thesis

in

The Department

of

Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Quality Systems Engineering) at

Concordia University

Montréal, Québec, Canada

March 2024

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Shadan Ghadimigheshlaghi**

Entitled: **Adaptive Priors in Probabilistic Topic Models for Bursty Discovery in Textual Data**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science (Quality Systems Engineering)

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the Final Examining Committee:

_____ Chair
Dr. Walter Lucia

_____ External Examiner
Dr. Mohsen Ghafouri

_____ Examiner
Dr. Walter Lucia

_____ Supervisor
Dr. Nizar Bouguila

Approved by _____
Dr. Chun Wang, Chair
Department of Concordia Institute for Information Systems Engineering

_____ 2024

_____
Dr. Mourad Debbabi, Dean
Faculty of Engineering and Computer Science

Abstract

Adaptive Priors in Probabilistic Topic Models for Bursty Discovery in Textual Data

Shadan Ghadimigheshlaghi

In the field of natural language processing, topic modeling plays an important role in detecting latent topics in large amounts of text. Models that use traditional methods of representation, however, often fail to capture the 'burstiness' characteristic of natural language - the tendency for previously occurring words to recur within the same document. In order to address this limitation, we introduce two innovative topic modeling frameworks: the Generalized Dirichlet Compound Multinomial Latent Dirichlet Allocation (GDCMLDA) and the Beta-Liouville Dirichlet Compound Multinomial Latent Dirichlet Allocation (BLDCMLDA). Using Dirichlet Compound Multinomial distribution together with Generalized Dirichlet and Beta-Liouville distributions, both frameworks integrate advanced distribution methods. By integrating these concepts, it is possible to model the burstiness phenomenon while maintaining a variety of topic proportion patterns that can be varied and flexible. As a result of our comprehensive evaluations across multiple benchmark text datasets, we conclude that GDCMLDA and BLDCMLDA are superior to existing models. The evidence for this is found in the improved performance metrics, including the scores for perplexity and coherence. Our results confirm that the proposed models are able to capture the complexities of word usage dynamics, thus contributing to a significant advancement in topic modeling.

Acknowledgments

Having completed my Master's degree at Concordia University, I am deeply grateful to many people for their support and contributions. Their invaluable assistance was essential to my achievements.

To begin with, I would like to thank my supervisor, Dr. Nizar Bouguila, for giving me the opportunity to pursue my Master degree at Concordia University and to work on such an exciting topic.

Also, I would like to thank Dr. Hafsa Ennajari for her invaluable advice, support, and patience during my research.

Thanks to Zetane Systems, where I completed a four-month internship through Mitacs Accelerate Explore, I had the opportunity to bridge academia and industry and contribute to cutting-edge research.

I would like to thank all my lab mates at Concordia University's eXplainable Artificial Intelligence (XAI) Lab for our good times and discussions.

I would like to conclude by expressing my heartfelt gratitude to my parents and my sister for their unconditional support throughout my life. Their love and belief in my ambitions have been a constant source of encouragement for me. My achievements would not have been possible without their belief in my potential.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem statement

In the current digital landscape, the large volume of textual data generated on various platforms emphasizes the importance of advanced topic modeling techniques. These techniques play a crucial role in breaking down large amounts of text into understandable themes, making it easier to retrieve, organize, and interpret information Bdiri, Bouguila, and Ziou (2014); Bouguila and Elguebaly (2012); Bouguila and ElGuebaly (2008b); Yang, Fan, and Bouguila (2022); Zamzami and Bouguila (2019a, 2019b, 2022). While foundational methods like Latent Dirichlet Allocation (LDA) have established the basis for identifying themes within extensive text collections, they assume a consistent spread of topics and words throughout the documents. This assumption does not align with the varied and dynamic nature of textual data found in real-world scenarios.

A significant challenge in this area is the phenomenon known as ”topic burstiness” characterized by the concentration of certain topics or words in specific documents rather than being evenly distributed across the entire text corpus. This phenomenon highlights the complex use of language and emphasis on themes that standard models, which rely on simple and rigid statistical distributions, cannot accurately capture. The failure to account for topic burstiness not only affects the clarity of identified topics but also impacts the application of these models in tasks requiring a deep understanding of text, such as analyzing sentiments, recommending content, and classifying information.

Additionally, traditional topic modeling algorithms, which require a predetermined number of

topics, limit their adaptability. This inflexibility is a significant issue when the ideal thematic detail is not known in advance or varies across different data sets. In order to uncover latent thematic structures in text corpora, new topic modeling approaches need to be developed that naturally address the burstiness phenomenon.

In order to overcome these challenges, new topic modeling algorithms must be developed that overcome the limitations of current methods. Statistical distributions should be more adaptable to better represent the complex dynamics of word and topic distributions; mechanisms should be implemented to automatically determine the optimal number of topics that should also be included in such approaches. By implementing these advancements, we aim to enhance our ability to identify underlying thematic patterns in large-scale textual data, leading to deeper insights and more effective decision-making based on data.

This thesis introduces the use of flexible priors as a solution to the burstiness challenge. Initially, we explore the generalized Dirichlet distribution, and subsequently, we integrate the Beta-Liouville distribution. This approach contrasts with traditional topic models that depend on the less flexible Dirichlet distribution. By adopting flexible priors, this research contributes to more precise topic burstiness modeling, significantly improving traditional topic modeling techniques.

1.2 Theoretical background and related works

1.2.1 Fundamentals

Topic modeling is an essential technique in natural language processing and information retrieval, enabling the discovery of hidden thematic structures in large collections of text data. This section describes the foundational concepts and methodologies that have significantly shaped the field of topic modeling.

Latent Semantic Analysis (LSA), introduced by Deerwester et al. Deerwester, Dumais, Furnas, Landauer, and Harshman (1990), is a dimensionality reduction technique that utilizes singular value decomposition (SVD) on the term-document matrix to identify a subspace capturing the most significant patterns in the data. This approach not only facilitates significant data compression but also captures underlying semantic relationships, such as synonymy and polysemy, among words.

Hofmann Hofmann (1999) proposed Probabilistic Latent Semantic Indexing (pLSI) as an advancement over LSA, modeling each word in a document as a sample from a mixture model, where the mixture components correspond to 'topics'. This model allows for a document to be associated with multiple topics, with each word being generated from a topic-specific multinomial distribution. However, pLSI lacks a probabilistic model at the document level, leading to scalability issues and challenges in generalizing to new documents not seen during training.

Latent Dirichlet Allocation (LDA), developed by Blei et al. Blei, Ng, and Jordan (2003), is a generative probabilistic model that extends beyond the capabilities of pLSI by introducing a hierarchical Bayesian framework. In LDA, documents are considered mixtures of topics, where each topic is characterized by a distribution over words. The model is defined by two primary processes: (1) selecting a mixture of topics for each document, and (2) generating each word in the document from one of these topics. LDA's formulation allows for efficient inference and learning, addressing the limitations observed in pLSI by providing a coherent generative process for unseen documents and offering a more scalable solution for large datasets.

The transition from LSA to pLSI, and subsequently to LDA, marks a significant evolution in the approach to topic modeling, with each successive model offering a more comprehensive understanding of document structures and topic distributions. A fundamental assumption behind these models is the 'bag-of-words' approach, treating documents as collections of words regardless of order—assuming exchangeability. De Finetti's theorem Fritz, Gonda, and Perrone (2021) validates this assumption by representing exchangeable collections of random variables as mixtures over latent structures, providing a robust probabilistic foundation for these models.

1.2.2 Literature Review

Latent Dirichlet Allocation (LDA): Latent Dirichlet Allocation Blei et al. (2003) is a technique for determining latent patterns within textual corpora. LDA assumes that each document is composed of various topics. The creation of each document is always a two-step process. Authors first decide which topics to include in a document while assigning each topic to a particular word distribution. Second, authors select words from a determined topic distribution for inclusion in their document, resulting in a diverse document. Although traditional LDA can identify the topic

burstiness to some extent, it often fails to account for the word burstiness within a document. This can result in an excessive level of burstiness at the topic level. The LDA model relies on two essential Dirichlet hyperparameters, α and β, which are crucial in shaping the generative process. α determines the topic proportions in documents, denoted by θ. Increasing α results in more similar topics, while decreasing it leads to greater topic diversity. On the other hand, β influences the word distributions in topics represented by ϕ. A higher value of β leads to more similar topics, while a lower value results in more varied word tendencies. Although the LDA model does not consider word burstiness, it remains a powerful tool for representing documents through shared topics. This modeling approach is particularly valuable for estimating multinomial distributions denoted as θ, which indicate the likelihood of topics occurring in each document. These estimated distributions are beneficial for document classification and similarity analysis and various other applications in the domain of textual analysis Das, Zaheer, and Dyer (2015).

Dirichlet Compound Multinomial (DCM): The Dirichlet Compound Multinomial (DCM) model offers a unique approach to text analysis, specifically focusing on capturing word burstiness within documents. The DCM is different from other models, such as LDA, in that it does not explicitly deal with topics but focuses on the distribution of words within individual documents.

The DCM model generates each document by selecting a document-specific multinomial distribution from a shared Dirichlet distribution. This approach results in each document comprised of words from a single distribution representing a subtopic within a broader topic. The key difference between DCM and LDA lies in their approach to mapping subtopics to documents.

DCM's ability to differentiate between multinomial and Dirichlet parameters allows it to adapt to burstiness by reducing the significance of repeated words as Dirichlet parameters change. However, DCM is primarily designed to represent one top-level topic with various aspects and may struggle when dealing with documents containing words from multiple topics. Despite its limitations, DCM provides valuable insights into the distribution and frequency of words within documents, making it a useful tool in analyzing text data where burstiness is a significant characteristic Huang, Xu, Qin, and Chen (2020).

Dirichlet Compound Multinomial Latent Dirichlet Allocation (DCMLDA): The Dirichlet Compound Multinomial Latent Dirichlet Allocation (DCMLDA) model in Doyle and Elkan (2009)

combines the strengths of DCM and LDA models to handle multiple topics within a document while considering word burstiness.

In the traditional LDA model, a single Dirichlet distribution (β) generates a multinomial distribution (ϕ_k) for each topic (k), which is then applied uniformly across all documents. In DCMLDA, however, instead of a single multinomial distribution, a unique multinomial word distribution ϕ_{kd} is generated for each combination of topic (k) and document (d). Specifically, for each document (d), the value of ϕ_{kd} is drawn from a Dirichlet distribution with parameters β_k, where each topic (k) has its own non-uniform β_k vector. This unique feature considers differences in the likelihood of words within the same topic across various documents, effectively capturing the concept of burstiness within topics.

The focus has shifted from a single set of multinomial topics to multiple sets of multinomial subtopics. LDA focuses on the ϕ, a V × K matrix representing word probabilities of given topics. However, in DCMLDA, ϕ becomes a three-dimensional array V × K × D, which accounts for each word's probability in each document's topic. This variation indicates that ϕ cannot be used for dimensionality reduction in data analysis. Instead, the attention shifts to the β array, a two-dimensional array containing Dirichlet parameters associated with words based on topics. Unlike LDA, the β values in DCMLDA do not have to sum up to one.

The DCMLDA model introduces more flexibility in the β values, effectively adding K extra degrees of freedom to account for word-level burstiness within each topic. The analysis of β values is similar to that of ϕ values in LDA, where higher β values indicate a greater probability of a word being associated with a specific topic. Therefore, β values are useful in identifying the most frequently occurring words within each topic. The DCMLDA model is a significant advancement in analyzing text. It efficiently manages multiple topics in a document while considering word burstiness. With unique β values for each topic and document, DCMLDA captures the complexity of textual data, outperforming traditional models, particularly when burstiness is significant.

Beta-Liouville Distribution:

The Beta-Liouville (BL) distribution, introduced as an extension within the Liouville family of distributions, provides a flexible framework for count data modeling in a multidimensional setting Fan and Bouguila (2013c); Hu, Fan, Du, and Bouguila (2019). Characterized by its generative

density function and a set of positive parameters $\alpha_1, \ldots, \alpha_D$, along with a generating density function parameterized by ξ, this distribution is distinguished for its ability to handle a wider range of covariance structures, unlike the more restrictive Dirichlet distribution. Its versatility and unique properties, including its conjugacy to the multinomial distribution and general covariance structure, make it particularly well-suited for applications involving complex count data, such as text categorization or image classification. The Beta-Liouville distribution's flexibility and comprehensive approach allow for more accurate and nuanced modeling of count data, addressing limitations found in traditional Dirichlet-based models and capturing the nuances in data more effectively Bouguila (2011); Bouguila and ElGuebaly (2008a); Luo, Amayri, Fan, and Bouguila (2023).

1.3 Contributions

This thesis has several contributions that can be listed as follows:

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

  This research has been submitted to the 37th International Florida Artificial Intelligence Research Society (FLAIRS) Conference.

- **Adaptive Priors for Burstiness Analysis in Topic Modeling with Generalized Dirichlet Distributions**

  This research has been accepted in The 9th IEEE International Conference on Data Science and Systems (DSS-2023).

1.4 Thesis Overview

- In chapter 1, we introduce some theoretical background of topic modeling.
- In chapter 2, we present the GDCMLDA model, showcasing its novel approach to modeling burstiness with Generalized Dirichlet distributions and its performance benefits.
- In chapter 3, we introduce the BLDCMLDA model, emphasizing its utilization of Beta-Liouville distributions for enhanced topic discovery and burstiness modeling.

- In chapter 4, we summarize our main findings, contributions, and future works.

Chapter 2

Adaptive Priors for Burstiness Analysis in Topic Modeling with Generalized Dirichlet Distributions

2.1 Introduction

In today's world, large amounts of data are produced from diverse domains. When effectively processed, this data can be a rich source of helpful information. Machine learning models have emerged as essential tools for effectively processing this Bakhtiari and Bouguila (2014a, 2014b). Specifically, topic modeling has proven to be an efficient method when dealing with large amounts of text data, allowing for the extraction of key themes hidden within a large number of documents Blei (2012); Vayansky and Kumar (2020).

Topic Models such as Latent Dirichlet Allocation (LDA) Blei et al. (2003) identify clusters of words (i.e., 'topics') that often occur together in the same context. This approach goes beyond merely considering individual words and instead represents documents based on the underlying topics they cover, allowing for deeper semantic understanding.

One significant linguistic phenomenon that can considerably impact document analysis and topic modeling is "Burstiness". First introduced by Church and Gale and Katz Madsen, Kauchak,

and Elkan (2005), burstiness refers to the pattern where a rare word, once it occurs in a document, is likely to appear again multiple times in the same document. Essentially, this means that words in a document appear in bursts. This phenomenon is not limited to textual data. Evidences of burstiness have also been observed in other fields such as finance, gene expression, and computer vision Blei and Lafferty (2007). This wider occurrence of burstiness highlights its importance in data analysis across multiple domains. Expanding on the concept of burstiness, it is essential to differentiate between two distinct types: word burstiness and topic burstiness. Word burstiness, as mentioned, describes the likelihood of a rare word reoccurring in a single document. Topic burstiness, on the other hand, refers to the recurrence of certain topics within a corpus of documents. This comprehensive analysis of burstiness - considering both word and topic burstiness - allows for a more in-depth analysis of documents and their structures, making this phenomenon an essential factor to consider in topic modeling.

Conventional topic models Blei et al. (2003); Das et al. (2015) use simple statistical approaches to capture words distribution across topics. However, these methods frequently fall short in effectively identifying emerging topics, often resulting in vague and ambiguous interpretations rather than clear and comprehensible ones. This limitation is largely attributed to the rigid nature of the underlying statistical models, such as the Dirichlet distribution, which lacks the flexibility needed to accommodate the dynamic shifts in topic trends. Consequently, these models struggle to adequately represent the burstiness of topics and often generate less interpretable topics.

In order to address these issues, we propose a novel approach: the Generalized Dirichlet Compound Multinomial Latent Dirichlet Allocation (GDCMLDA) model. In this model, documents are modeled more effectively by accurately tackling the challenge of topic burstiness. We integrated flexible Generalized Dirichlet (GD) distributions as priors, allowing us to address the dynamic nature of topic burstiness better. The primary advantage of this model is its inherent flexibility, thanks to the variant nature of the GD distribution, which allows for a more comprehensive and precise modeling of topics. Another unique strength of this model is its ability to process large data volumes without compromising on accuracy.

To conclude, we can summarize our contribution as follows:

- We introduce a novel topic modeling approach, the Generalized Dirichlet Compound Multinomial Latent Dirichlet Allocation, specifically designed to address the challenge of topic burstiness in text data.
- We demonstrate the effectiveness of modeling topic proportions with the GD priors and highlight its notable impact on capturing the intricate dynamics of topic burstiness.
- Results on real-world datasets demonstrate that the GDCMLDA model consistently produced topics with higher semantic coherence, indicating that our model generates more interpretable and contextually meaningful topics.
- We demonstrate that the GDCMLDA model exhibits lower perplexity scores across various topic settings, underscoring its ability to predict text samples more accurately.

The rest of this chapter is organized into two sections. In section 2, we introduce the proposed probabilistic model in detail and, in section 3, we present our experimental results for various text analysis tasks.

2.2 Proposed Model

In this research, we developed the GDCMLDA model by combining the strengths of three different approaches: GD, DCM, and LDA. In this section, we will discuss the mathematical formulation and key components of the GDCMLDA model, including the generative model as well as the parameters learning procedure Bouguila and Ziou (2005); Fan, Sallay, and Bouguila (2017); Ihou and Bouguila (2017).

2.2.1 Model definition

Our proposed GDCMLDA model combines GD, DCM, and LDA to provide a flexible way to model topic burstiness while still keeping them document-specific. The GDCMLDA graphical model is shown in Figure 3.1.

The Generalized Dirichlet Compound Multinomial Latent Dirichlet Allocation model integrates the Generalized Dirichlet distribution to enhance the precision and adaptability of representing topic

proportions within a document. This distribution exhibits greater flexibility compared to the Dirichlet distribution, as it allows independent sampling of each entry in the random vector of topic proportions from Beta distributions.

In GDCMLDA, we assume that for each topic k and document d, a new multinomial word distribution ϕ_{kd} is sampled. The β_k vector varies for each topic k and is not uniform. Each document d has a corresponding ϕ_{kd} that is randomly generated based on a Dirichlet distribution with parameter β_k. This ensures that the occurrences of each topic are connected across all documents in the corpus and accounts for burstiness.

GDCMLDA shifts the modeling focus from a single collection of multinomial topics to multiple collections of multinomial subtopics. Given a corpus of documents, the vocabulary size is denoted as V, the number of topics as K, and the number of documents as D. LDA represents the topic using ϕ, which is a $V \times K$ matrix of word probabilities. In GDCMLDA, the variable ϕ has a three-dimensional structure $V \times K \times D$, representing each word's probability in each document's topic.

The generative model for GDCMLDA is described in Algorithm 3.2.2. Our model's unobserved variables are distributed as follows:

$$\theta \sim GeneralizedDirichlet(a, b)$$

$$z \sim Multinomial(\theta)$$

$$\phi \sim Dirichlet(\beta)$$

Algorithm 1 GDCMLDA Generative Model

for document $d \in \{1, 2, ..., D\}$ **do**
    Draw topic distribution $\theta_d \sim GD(a, b)$
    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

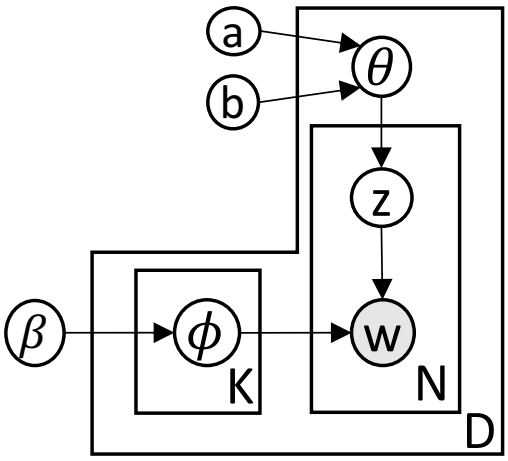

Figure 2.1: Graphical Model of GDCMLDA. Parameters a and b are of size K, where K is the number of topics. Vector β has size $K \times V$, where V is the vocabulary size.

2.2.2 Parameter Inference

The GDCMLDA model involves many hidden parameters, leading to the computation of the posterior distribution, that is not analytically tractable. In such situations, approximation techniques are employed, and one widely adopted method is Gibbs sampling Griffiths and Steyvers (2004). This method is a Markov chain Monte Carlo (MCMC) approach that iteratively samples from the conditional distributions of the latent variables, allowing us to approximate the posterior distribution and efficiently estimate model parameters.

Specifically, our model contains six unobserved variables: a, b, β, ϕ, θ, and z, categorized into per-document or per-word parameters ϕ, θ, and z) and hyperparameters (a, b, and β). When provided with a training document set, the learning process objective is to determine the optimal values of these variables. This involves an iterative optimization of the topic parameters (i.e., ϕ, θ, and z) keeping the hyperparameters (i.e., a, b, and β) fixed and vice versa, optimizing hyper-parameters using the optimized values of the topic parameters. When hyperparameter values are fixed, collapsed Gibbs sampling is employed to determine the z distribution within the documents, and subsequently, ϕ and θ can be simply computed based on the Gibbs sampling main equation. Furthermore, given z samples, values for a, b, and β that maximize the likelihood of the training documents are determined using the Monte Carlo expectation-maximization technique.

Gibbs Sampling.

Following Heinrich et al. Heinrich (2009), we develop an efficient Gibbs sampling technique for GDCMLDA hidden parameters estimation. Firstly, we factorize the model's complete likelihood as follows:

$$p(w, z|(a, b), \beta) = p(w|z, \beta)p(z|(a, b)) \tag{1}$$

The first probability is an average over all possible ϕ distributions.

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

Expanding $p(\phi|\beta)$ as a Dirichlet distribution yields:

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

Above, the multivariate Beta function $B(\cdot)$ is used, along with the frequency of word t assigned to topic k in document d, denoted by n_{tkd}.

A Generalized Dirichlet prior distribution is assumed for the topic mixtures in each document to determine the probability of topics proportions, whereas the topic assignments distribution is modeled using a Multinomial distribution. Therefore, we can conclude that:

$$\begin{aligned} p(z|(a, b)) &= \int_{\theta} p(z|\theta)p(\theta|(a, b))d\theta \\ &= \prod_{j=1}^{D} \prod_{k=1}^{K-1} \frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)} \prod_{k=1}^{K-1} \frac{\Gamma(a_k^j) + \Gamma(b_k^j)}{\Gamma(a_k^j + b_k^j)} \end{aligned} \tag{4}$$

where: $a_k^j = a_k + n_{.kd}, b_k^j = b_k + n_{.d,k+1} + ... + n_{.dK}$.

The topic assignments z are unobserved. We use n_{tkd} to count word t in topic k and document d in the complete corpus $w_{-i} \cup w_i$, and use n'_{tkd} to count in the limited corpus w_{-i}.

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

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

while the Dirichlet part will lead to:

$$\frac{p(w|z, \beta)}{p(w|z^{-i}, \beta)} \propto \frac{n_{w_i z_i d_i} + \beta_{w_i z_i} - 1}{\sum_t n_t z_i d_i + \beta_{t z_i} - 1} \tag{6}$$

and the Generalized Dirichlet part:

$$\frac{p(z|a, b)}{p(z^{-i}|a, b)} \propto$$

$$\begin{cases} \frac{a_k + n'_{.kd}}{a_k + b_k + \sum_{l=1}^{K} n'_{.ld}} & k = 1 \\ \frac{a_k + n'_{.kd}}{a_k + b_k + \sum_{l=1}^{K} n'_{.ld}} \prod_{m=1}^{k-1} \frac{b_m + \sum_{l=m+1}^{K} n'_{.ld}}{a_m + b_m + \sum_{l=m}^{K} n'_{.ld}} & 1 < k < K \\ \prod_{m=1}^{k-1} \frac{b_m + \sum_{l=m+1}^{K} n'_{.ld}}{a_m + b_m + \sum_{l=m}^{K} n'_{.ld}} & k = K \end{cases} \tag{7}$$

Hyperparameter EM.

After determining the topic parameters, the next step is to optimize the hyperparameters (a, b, and β). This is achieved by using a Monte Carlo expectation-maximization (EM) approach. The process involves iteratively optimizing the values of a, b, and β until the likelihood of the training documents is maximized.

Prior studies have typically assumed fixed, uniform priors for topic mixtures θ and vocabulary distributions for topics ϕ, represented by constant values for parameters a, b, and β Griffiths and Steyvers (2004). However, Wallach et al. Wallach, Mimno, and McCallum (2009) have shown that

utilizing asymmetric Dirichlet priors for topic probabilities leads to better model fitting. In the context of GDCMLDA, we take a different approach by estimating the GD distribution parameters a and b to uncover topic correlations. We also estimate the parameters for the prior distribution of words given topics (i.e., β). Ideally, we would directly maximize the likelihood $p(w|a, b, \beta)$ for observed data w and hyperparameters a, b, and β. Unfortunately, this distribution is not computationally tractable for this model. To address this challenge, we augment the likelihood to $p(w, z|a, b, \beta)$ and employ Monte Carlo Expectation Maximization (MCEM) Wei and Tanner (1990). Given hyperparameters a, b, and β, we employ Gibbs sampling to estimate the posterior distribution of topic assignments for each word (E-step) as discussed above. Subsequently, given the expected topic assignments and words, we optimize $p(w, z|a, b, \beta)$ (M-step). Algorithm 3.2.2 provides a detailed description of these iterative processes. For fitting β, we maximize the joint distribution conditioned on the expected topic assignments $z = E(z|a, b, \beta)$, which are estimated through Gibbs sampling. We have the optimal function as follows:

$$\begin{aligned}\beta_{.k}^{new} = \arg\max_{\beta} \sum_{d,t} (log\Gamma(n_{tkd} + \beta_{tk}) - log\Gamma(\beta_{tk})) \\ + \sum_{d} [log\Gamma(\sum_{t} \beta_{tk}) - log\Gamma(\sum_{t} n_{tkd} + \beta_{tk})]\end{aligned} \tag{8}$$

We follow the Newton-based approach proposed by Minka Minka (2012). Here, we have a DCM distribution to fit from K observed vectors of dimension V.

Similarly, we estimate the parameters of the GD distribution Caballero Barajas, Barajas, and Akella (2012) by maximizing the joint distribution:

$$\begin{aligned}a^{new}, b^{new} = \arg\max_{a,b} \prod_{j=1}^{D} \prod_{k=1}^{K-1} \frac{\Gamma(a_k + b_k)}{\Gamma(a_k)\Gamma(b_k)} \\ \times \prod_{k=1}^{K-1} \frac{\Gamma(a_k^j) + \Gamma(b_k^j)}{\Gamma(a_k^j + b_k^j)}\end{aligned} \tag{9}$$

where: $a_k^j = a_k + \bar{N}_{j,k}$, $b_k^j = b_k + \bar{N}_{j,k+1} + ... + \bar{N}_{j,k}$.

After obtaining the optimal parameters (a^{new}, b^{new}) from Algorithm 3.2.2 and the word-topic

observations (w, z), the predictive distribution for document j, $\hat{\theta}_j$ can be estimated.

$$\hat{\theta}_{d,k} = \begin{cases} \frac{a_k^* + \bar{n}_{.kd}}{a_k^* + b_k^* + \sum_{l=1}^{K} \bar{n}_{.ld}} & k = 1 \\ \frac{a_k^* + \bar{n}_{.kd}}{a_k^* + b_k^* + \sum_{l=1}^{K} \bar{n}_{.ld}} \prod_{m=1}^{k-1} \frac{b_m^* + \sum_{l=m+1}^{K} \bar{n}_{.ld}}{a_m^* + b_m^* + \sum_{l=m}^{K} \bar{n}_{.ld}} & 1 < k < K \\ \prod_{m=1}^{k-1} \frac{b_m^* + \sum_{l=m+1}^{K} \bar{n}_{.ld}}{a_m^* + b_m^* + \sum_{l=m}^{K} \bar{n}_{.ld}} & k = K \end{cases} \tag{10}$$

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

The probability of words given topics, $\hat{\phi}_k$, is estimated through the following predictive distribution Griffiths and Steyvers (2004).

$$\hat{\phi}_{tkd} = \frac{\bar{n}_{w_i z_i d_i} + \beta^*_{w_i z_i} - 1}{\sum_t \bar{n}_t z_i d_i + \beta^*_{t z_i} - 1} \tag{11}$$

It is important to note that the probability of certain topics, represented by $\hat{\theta}_j$, depends on the specific document in question. On the other hand, the probability of words, given their corresponding topic, is represented by $\hat{\phi}_k$. This means that when predicting the distribution of topics for a new, unseen document, we need to consider that document's unique probabilities while keeping the probabilities of words and their topics constant.

Algorithm 2 Monte Carlo EM

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

2.3 Experimental Results

To validate the effectiveness of our proposed model, we conducted extensive experiments on a variety of evaluation tasks. Our primary objective is to assess the GDCMLDA model's ability to

discover both interpretable and coherent topics while also evaluating its performance in predicting held-out documents. Based on three publicly available datasets, the NIPS, Movie Review, and the 20 Newsgroup dataset, we validated the performance of our model. We compare GDCMLDA with the DCMLDA model because the DCMLDA model also accounts for burstiness, making it a relevant baseline of comparison to assess the effectiveness of GDCMLDA in addressing this aspect.

2.3.1 Datasets

We evaluated our proposed GDCMLDA model on three different datasets, each presenting its unique of characteristics and challenges. These datasets are outlined as follows:

- NIPS: This dataset comprises 1740 documents, primarily papers from the Conference and Workshop on Neural Information Processing Systems (NeurIPS, formerly NIPS). The papers span from the inaugural 1987 conference to the 2016 conference.
- Movie Review: This dataset is a collection of text-based movie reviews often employed in natural language processing and sentiment analysis research. This dataset contains a balanced set of 2000 reviews, including 1000 negative samples and 1000 positive samples.
- 20 Newsgroups: This dataset is a collection of approximately 20,000 newsgroup documents, partitioned nearly evenly across 20 different newsgroups, each corresponding to a different topic. Some of the newsgroups are very closely related to each other (e.g., comp.sys.ibm.pc. hardware and comp.sys.mac.hardware), while others are highly unrelated (e.g., misc.forsale and soc.religion.christian).

We chose these datasets due to their diversity and complexity, which collectively provide a comprehensive test of the GDCMLDA model's capabilities. To preprocess all datasets, we first convert all words into lowercase, tokenize sentences, and then remove stop words, punctuation, and rare words appearing less than five times in the corpus.

Table 2.1: Examples of topics learned by GDCMLDA and DCMLDA on the Movie Review dataset.

Horror	Family	Hollywood	Relationships	Fairy Tale	Sci-Fi
			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

2.3.2 Topic Coherence

In topic modeling, the interpretability of the generated topics is of utmost importance. A model that generates semantically coherent and meaningful topics is more valuable as it provides more insightful information about the underlying structure of the dataset.

In Table 3.1, we present six randomly selected topics along with top words that show the strongest positive associations with each of them. These topics have been derived from both GDCMLDA and DCMLDA models, and we tried to align the topics discovered by both models to ensure consistency with their respective topic labels. From the results, we can see that for GDCMLDA, it is straightforward to conclude their respective topics by examining their top respective words. In contrast, with DCMLDA, the presence of numerous intruder words complicates the interpretation of each topic, introducing nuances that obscure the overall topic and make it challenging to grasp the intended meaning.

For instance, The "Horror" topic inferred by our model contains consistent words such as "horror" "dead" "murder" and "kill", whereas DCMLDA some generic words such as "time". Also, GDCMLDA successfully characterized the "Family" topic by terms like "love," "mother," "father," and "son," capturing the familial and emotional aspects typical of reviews about family-oriented movies The DCMLDA model on the other hand, struggled to convey the essence of the "Family" topic. It exhibited a broader range of terms, including some unrelated ones like "lot" and "written" which diluted the specific familial context seen in GDCMLDA's representation of the topic. These examples demonstrate the model's ability to distinguish diverse thematic content within the movie review dataset, highlighting its effectiveness in topic identification across various genres and subjects.

To quantitatively assess the interpretability of our proposed GDCMLDA method compared to the DCMLDA method, we use the topic coherence measure Newman, Lau, Grieser, and Baldwin (2010); Nikolenko, Koltcov, and Koltsova (2017). Topic coherence provides a quantitative measure of the semantic interpretability of the generated topics, which is a crucial aspect of the quality of topic models. It is a measure of how semantically similar the top words are. A higher coherence score indicates that the top words of the topic are more meaningful and related to each other. We

calculated the coherence score for each generated topic based on the top 10 words for both the GDCMLDA and DCMLDA methods. The overall coherence score for a model is then calculated as the mean of the coherence scores of all the generated topics. This yields a unified metric for assessing and comparing the quality of topics generated by various methods.

The results are presented in Table 3.2. The GDCMLDA method achieved a higher mean coherence score than the DCMLDA method, producing more semantically coherent topics. This suggests that the GDCMLDA method is more effective in capturing the semantic relationships between words in the dataset, leading to more meaningful topic generation.

Table 2.2: Mean coherence scores of the DCMLDA and GDCMLDA methods.

Dataset	DCMLDA	GDCMLDA
NIPS	0.15	0.34
20NewsGroups	0.199	0.272
Movie Review	0.065	0.091

These findings also demonstrate the considerable potential of the GDCMLDA method for downstream topic modeling tasks. The increased topic coherence suggests that our approach can provide more interpretable and semantically meaningful topics, which is a significant advantage in many applications. Further experiments will investigate the performance of the GDCMLDA method on more diverse and challenging datasets.

2.3.3 Perplexity

In addition to topic coherence, we also evaluated the performance of the proposed GDCMLDA method and the DCMLDA method using the perplexity measure Jelinek, Mercer, Bahl, and Baker (1977). Perplexity is a widely used metric for evaluating probabilistic models such as topic models. It measures how well a probability distribution predicts a sample and is inversely proportional to the log-likelihood of the test data. Topic modeling aims to minimize the perplexity since a lower perplexity score indicates that the model is better at predicting the sample. Formally, perplexity can be defined as:

$$Perplexity(D_{test}) = \exp\left(\frac{\sum_{d=1}^{D} \log p(\mathbf{w}_d)}{\sum_{d=1}^{D} N_d}\right)$$

where:

- D_{test} is the number of documents in the test dataset.
- w_d represents the words in document d.
- N_d is the number of words in document d.
- $p(w_d)$is the likelihood of document d given the learned topic model.

We conducted our experiments with different numbers of topics (i.e., 10, 20, 30, 40, and 50). We trained the models for each setting and then calculated the perplexity score over all topics.

The results of our study are presented in figure 3.2, 3.3 and 3.4. These figures show the perplexity scores of the two different methods, DCMLDA and GDCMLDA, as we vary the number of topics across different datasets. Our analysis of the NIPS dataset consistently demonstrates that the GDCMLDA method outperforms DCMLDA in terms of providing a more accurate representation of the data, as indicated by consistently lower perplexity scores. This pattern is further illustrated in our examination of the 20NewsGroups and Movie Review datasets, demonstrating that GDCMLDA is more effective than DCMLDA in predicting the words within the test documents, as reflected in the lower perplexity scores it achieves.

In figure3.5, we plotted the learning curves for the NIPS, 20NewsGroups, and Movie Review datasets using perplexity scores. As depicted, perplexity scores consistently decrease across all three datasets, as it continues to learn from the data over successive iterations.

These findings highlight the robustness and reliability of the GDCMLDA method in achieving superior performance across a variety of datasets. This superior performance, as indicated by lower perplexity scores, signifies the enhanced ability of GDCMLDA to uncover the latent structure and patterns within complex textual data.

2.3.4 Topic Diversity

Topic Diversity is another important metric for evaluating the generated topics quality. This measure aims to quantify the extent of non-overlapping information across the generated topics. A

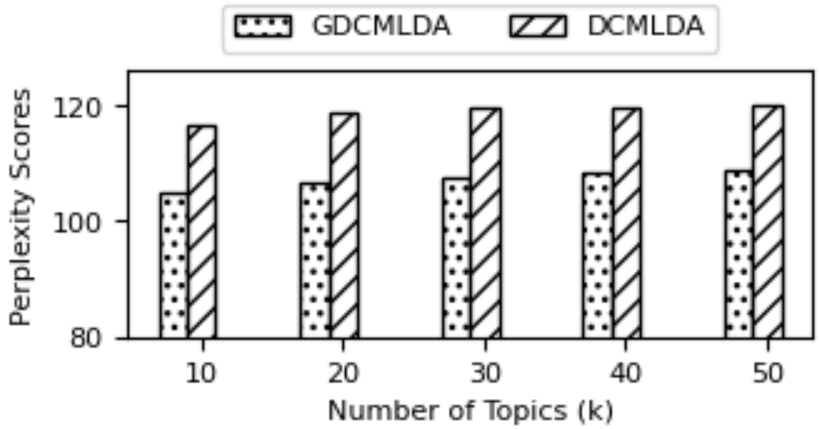

Figure 2.2: Perplexity scores of the DCMLDA and GDCMLDA methods for different numbers of topics for NIPS dataset.

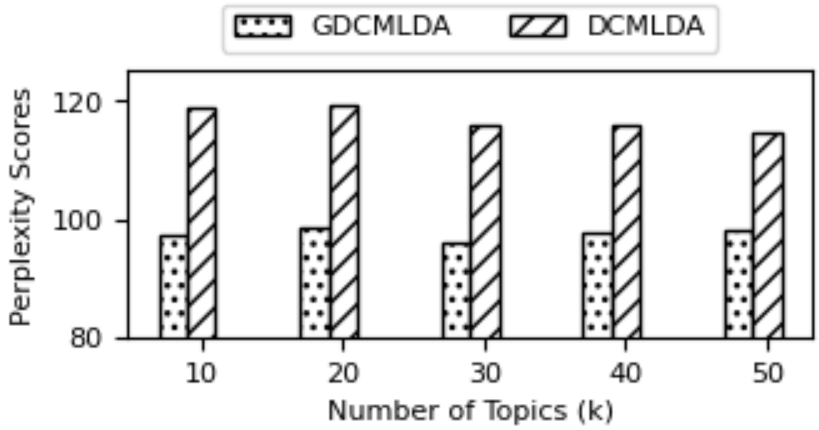

Figure 2.3: Perplexity scores of the DCMLDA and GDCMLDA methods for different numbers of topics for 20NewsGroups dataset.

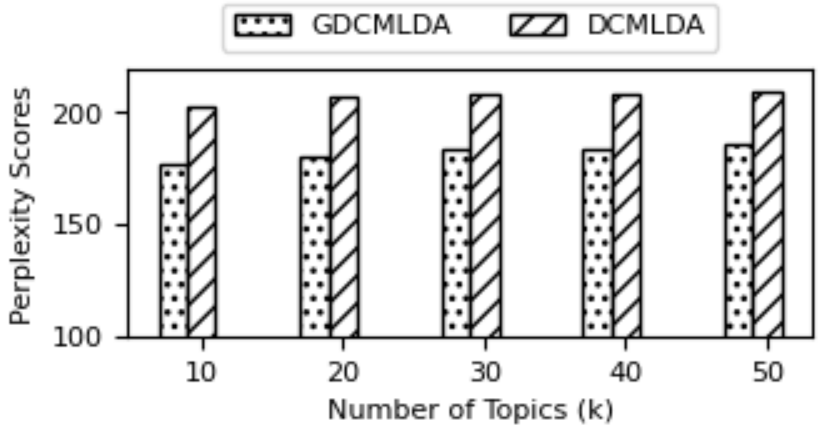

Figure 2.4: Perplexity scores of the DCMLDA and GDCMLDA methods for different numbers of topics for Movie Review dataset.

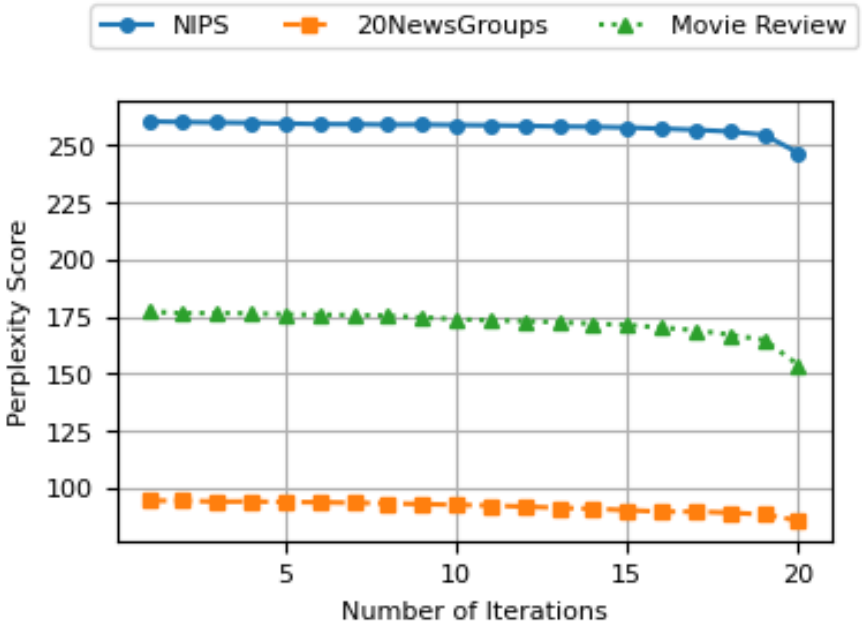

Figure 2.5: Learning Curve for Perplexity Scores Across 20 Iterations for Three Datasets

higher Topic Diversity score indicates a greater variety of topics, suggesting a broader coverage of the dataset semantic space.

We measured the Topic Diversity of the top 10 words for each topic for both the GDCMLDA and DCMLDA methods. This was performed using the following procedure:

(1) For each topic, the top 10 words were identified.

(2) A set of unique words is then created by taking the union of the top 10 words across all topics.

(3) The Topic Diversity is then calculated as the ratio of the number of unique words to the total number of words.

The results of this experiment are presented in Table 3.3. The GDCMLDA method achieved a higher Topic Diversity score than the traditional DCMLDA method, producing topics with a wider range of unique words. These results suggest that the GDCMLDA method is more effective in capturing the diversity of the textual data, leading to a broader coverage of the semantic space.

Table 2.3: Topic Diversity scores of the DCMLDA and GDCMLDA methods.

Dataset	DCMLDA	GDCMLDA
NIPS	0.52	0.63
20NewsGroups	0.72	0.78
Movie Review	0.38	0.52

These findings highlight the potential of the GDCMLDA method for topic modeling tasks. The increased Topic Diversity suggests that our approach can provide a more comprehensive understanding of the corpus, which is a significant advantage for many NLP downstream applications.

Chapter 3

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

3.1 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. 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, 2014b). Models like Latent Dirichlet Allocation Blei et al. (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 Madsen et al. (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 et al. (2003); Das et al. (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 distributionBouguila (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 rest of this chapter is organized into two sections. In section 2, we introduce the proposed model in detail and, in section 3, we discuss our experimental results.

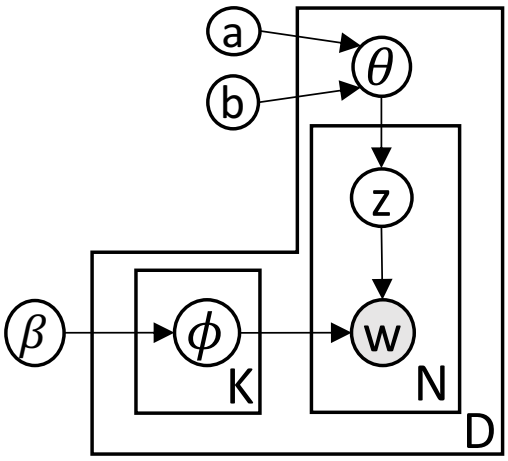

Figure 3.1: Graphical Model of BLDCMLDA.

3.2 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.

3.2.1 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 3.1.

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.

Algorithm 3 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

The generative process of BLDCMLDA is outlined in Algorithm 3.2.2. The probabilistic assumptions of the proposed model latent variables are described as follows:

$$\theta \sim \text{Beta-Liouville Distribution}(\vec{\delta})$$

$$z \sim \text{Multinomial}(\theta)$$

$$\phi \sim \text{Dirichlet}(\beta)$$

3.2.2 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.

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:

$$p(w, z|\vec{\delta}, \beta..) = p(w|z, \beta)p(z|\vec{\delta}) \tag{12}$$

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

$$\begin{aligned} p(w|z, \beta..) &= \int_{\phi} p(z|\phi)p(\phi|\beta)d\phi \\ &= \int_{\phi} p(\phi|\beta) \prod_{d} \prod_{n=1}^{N_d} \phi_{w_{dn} z_{dn} d} d\phi \\ &= \int_{\phi} p(\phi|\beta) \prod_{d,k,t} (\phi_{tkd})^{n_{tkd}} d\phi \end{aligned} \tag{13}$$

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

$$\begin{aligned} p(w|z, \beta..) &= \int_{\phi} \left[\prod_{d,k} \frac{1}{B(\beta_{.k})} \prod_{t} (\phi_{tkd})^{\beta_{tk}-1}] \times [\prod_{d,k,t} (\phi_{tkd})^{n_{tkd}} \right] d\phi \\ &= \prod_{d,k} \int_{\phi} \prod_{t} (\phi_{tkd})^{|\beta tk-1+n_{tkd}} d\phi \\ &= \prod_{d,k} \frac{B(n_{.kd} + \beta_{.k})}{B(\beta_{.k})} \end{aligned} \tag{14}$$

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 asBouguila (2012b); Fan and

Bouguila (2013a):

$$p(\boldsymbol{\theta}|\vec{\boldsymbol{\delta}}) = \frac{\Gamma\left(\sum_{k=1}^{K-1}\alpha_k\right)\Gamma(\alpha+\gamma)}{\Gamma(\alpha)\Gamma(\gamma)}\prod_{k=1}^{K-1}\frac{\theta_k^{\alpha_k-1}}{\Gamma(\alpha_k)} \times \left(\sum_{k=1}^{K-1}\theta_k\right)^{\alpha-\sum_{k=1}^{K-1}\alpha_k}\left(1-\sum_{k=1}^{K-1}\theta_k\right)^{\gamma-1} \tag{15}$$

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

$$p(z_i|z^{-i}, w, \vec{\delta}, \beta) = \frac{p(w|z,\beta)p(z|\vec{\delta})}{p(w|z^{-i},\beta)p(z^{-i}|\vec{\delta})} \tag{16}$$

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 et al. (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 3.2.2. 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:

$$\begin{aligned}\beta_{.k}^{new} = \arg\max_{\beta} \sum_{d,t} (\log\Gamma(n_{tkd} + \beta_{tk}) - \log\Gamma(\beta_{tk})) \\ + \sum_{d} [\log\Gamma(\sum_{t} \beta_{tk}) - \log\Gamma(\sum_{t} n_{tkd} + \beta_{tk})]\end{aligned} \tag{17}$$

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 3.2.2 and considering the word-topic observations (w, z), we can compute the predictive distribution for a given document d, denoted as $\hat{\theta}_d$.

$$\begin{aligned}\hat{\theta}_{d,k} = \\ \frac{\prod_{i=1}^{K-1}(\alpha + \sum_{k=1}^{K-1} m_{k,\not d} + i - 1)(\gamma + m_{K,\not d})}{\prod_{i=1}^{K}(\sum_{k=1}^{K}(\alpha_k + m_{k,\not d}) + i - 1) \prod_{i=1}^{N_d}(n_{k,\not d} + V\beta + i - 1)} \\ \frac{\prod_{k=1}^{K}(\alpha_k + m_{k,\not d}) \prod_{w=1}^{V}(n_{k,\not d}^{(w)} + \beta}{\prod_{i=1}^{K}(\alpha + \sum_{k=1}^{K-1} m_{k,\not d} + \gamma + m_{K,\not d} + i - 1)}\end{aligned} \tag{18}$$

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.

$$\hat{\phi}_{tkd} = \frac{\bar{n}_{w_i z_i d_i} + \beta^*_{w_i z_i} - 1}{\sum_t \bar{n}_t z_i d_i + \beta^*_{t z_i} - 1} \tag{19}$$

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 4 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^*

3.3 Experimental Results

To evaluate the effectiveness of the proposed BLDCMLDA model, we conducted comprehensive experiments across various tasks. Our goal is to assess the proficiency of the model in identifying interpretable and coherent topics and its predictive accuracy for unseen documents. We used three public datasets, namely, NIPS, Movie Review, and 20 Newsgroup. We compare our model against GDCMLDA and DCMLDA models, which also consider burstiness, providing a relevant benchmark for assessing BLDCMLDA's performance. One of the advantages of Beta-Liouville distributions over generalized Dirichlet distributions is that they have fewer parameters than generalized Dirichlets.

3.3.1 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.forsaleandsoc.religion.christian, are markedly different.

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

3.3.2 Topic Coherence

In topic modeling, the clarity and relevance of the topics generated are crucial. A model that produces topics that are semantically clear and relevant is considered more effective, as it offers deeper insights into the dataset's underlying structure.

In Table 3.1, six topics are showcased, each with keywords strongly associated with them, derived from the BLDCMLDA, GDCMLDA, and DCMLDA models. For BLDCMLDA, understanding these topics is more straightforward based on their keywords. However, with GDCMLDA and DCMLDA, the presence of unrelated words makes it more difficult to interpret the topics clearly, as they add complexity and obscure the main theme.

Our model effectively identified specific themes, such as the "Horror" topic with words like "genocide", "scream", and "horror" unlike the DCMLDA model which included more general terms such as "time". Similarly, the BLDCMLDA model accurately captured the "Relationships" topic

Table 3.1: 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

Table 3.2: Mean coherence scores of the DCMLDA, GDCMLDA, and BLDCMLDA methods.

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

with words like ”love”, ”wife”, ”man”, and ”children”, reflecting family and emotional aspects. In contrast, the DCMLDA model's representation of the same topic was less focused, including broader and unrelated terms that diluted the specific context. These instances showcase the superior capabilities of our model in distinguishing varied thematic content in movie reviews, demonstrating its effectiveness in topic identification across different genres and subjects.

We evaluated the interpretability of our BLDCMLDA method against GDCMLDA and DCMLDA using the topic coherence measure Newman et al. (2010); Nikolenko et al. (2017). This measure quantifies how semantically related the top words in a topic are, indicating the quality of the topics. Higher coherence scores mean the top words are more relevant and related to each other. We computed this score for each topic from the top 10 words inferred by BLDCMLDA, GDCMLDA, and DCMLDA. The overall coherence of the model is the average of these scores, providing a standardized way to compare the topic quality across methods.

The findings, detailed in Table 3.2, show that the BLDCMLDA method outperformed GDCMLDA and DCMLDA in terms of mean coherence score. This indicates that BLDCMLDA is more adept at understanding semantic links between words in the three datasets, resulting in the generation of topics that are more meaningful and semantically coherent.

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.

3.3.3 Perplexity

We also assessed the BLDCMLDA, GDCMLDA, and DCMLDA 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 figures 3.2, 3.3, and 3.4, 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 BLDCMLDA surpasses the other methods in predicting test document words, as reflected by its lower perplexity scores.

In Figure 3.5, 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.

3.3.4 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.

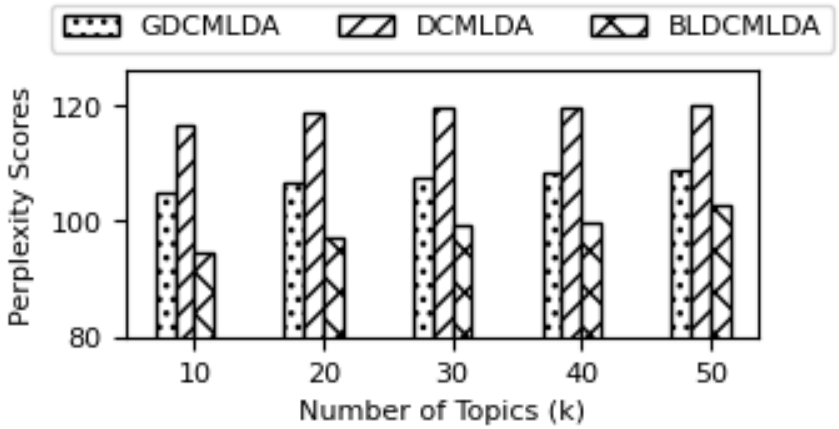

Figure 3.2: Perplexity scores for the DCMLDA, GDCMLDA, and BLDCMLDA models across varying topic counts on the NIPS dataset.

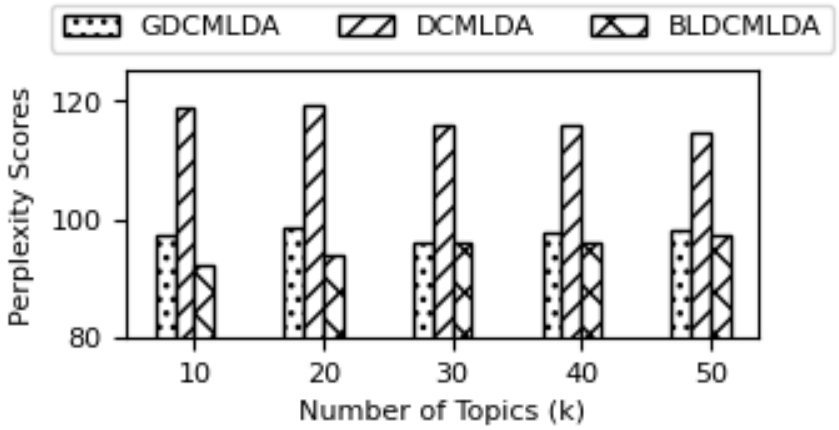

Figure 3.3: Perplexity scores for the DCMLDA, GDCMLDA, and BLDCMLDA models across varying topic counts on the 20 NewsGroups dataset.

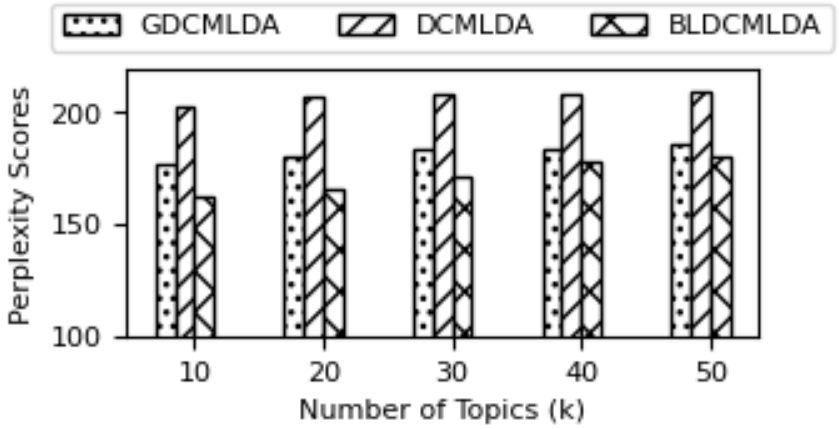

Figure 3.4: Perplexity scores for the DCMLDA, GDCMLDA, and BLDCMLDA models across varying topic counts on the Movie Review dataset.

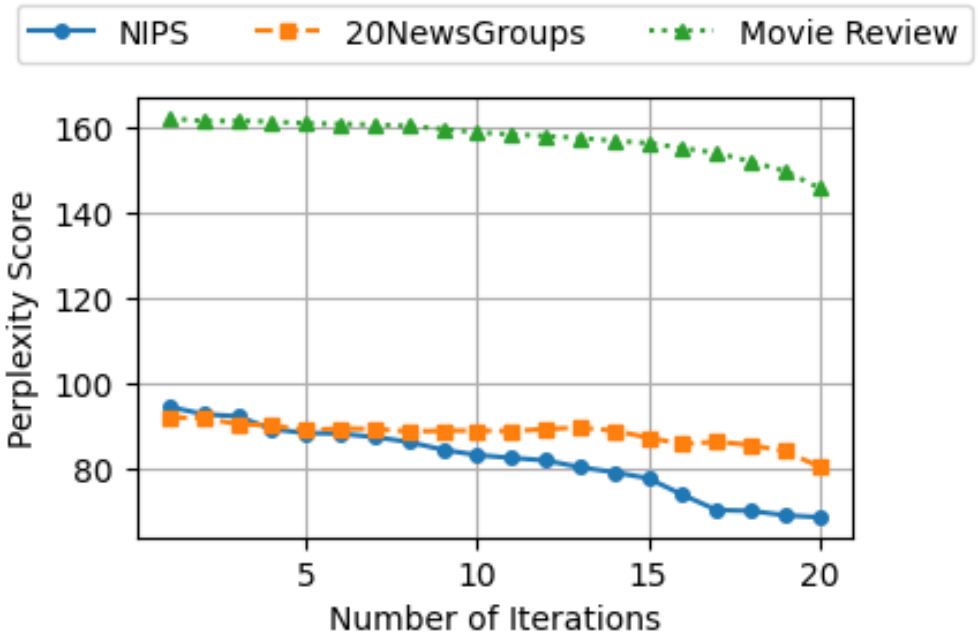

Figure 3.5: Perplexity Score Trends Over 20 Iterations for Three Different Datasets.

(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 3.3, 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 3.3: Topic Diversity scores of DCMLDA, GDCMLDA, and BLDCMLDA.

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

Chapter 4

Conclusion

This thesis has focused on a comprehensive exploration of advanced topic modeling techniques, specifically focusing on addressing the challenge of burstiness in textual data. Through a detailed examination of traditional models and their limitations, particularly concerning their inability to capture the dynamic and bursty nature of topics, this work has introduced novel approaches that significantly improve upon existing methodologies.

The foundational critique of conventional models like Latent Dirichlet Allocation (LDA) highlighted their shortcomings in dealing with the phenomenon of topic burstiness, where specific topics or words are concentrated in certain documents in a non-uniform way. This critique set the stage for the development of new models that are based on flexible priors and distributions, such as the Generalized Dirichlet and Beta-Liouville distributions, to more accurately reflect the complexities of real-world textual data.

The introduction of the Generalized Dirichlet Compound Multinomial Latent Dirichlet Allocation (GDCMLDA) model in Chapter 2 marked a significant advancement in the field of topic modeling. By integrating flexible Generalized Dirichlet distributions as priors, the GDCMLDA model demonstrated an enhanced ability to capture topic burstiness, leading to topics with higher semantic coherence and lower perplexity scores. This model represents a substantial step forward in the creation of more interpretable and accurate topic models.

Further innovation was presented in Chapter 3 with the development of the Beta-Liouville Dirichlet Compound Multinomial Latent Dirichlet Allocation (BLDCMLDA) model. As a result

of this model's integration of the Beta-Liouville distribution with the Dirichlet Compound Multinomial distribution, it provided a robust framework for capturing the nuances of word burstiness and topic distribution. Because of BLDCMLDA's superior semantic coherence and predictive accuracy, it has the potential to revolutionize topic modeling by better representing topic distributions and adapting to language burstiness.

This thesis makes several important contributions. It finds and tackles an important missing piece in traditional methods of analyzing topics within texts. Moreover, it introduces new, tested solutions that enhance the ability to make decisions based on deep analysis of text data. These improvements are applicable in various areas, including scholarly research and different industry sectors.

In summary, this thesis has shown that using more adaptable and detailed methods for analyzing topics can greatly improve the precision and clarity of topics identified from large amounts of text data. The introduced models, GDCMLDA and BLDCMLDA, are key advancements in this area, offering new ways to analyze text data more insightfully and effectively. Future studies should keep improving these models, checking how well they work with different types of data and making them more computationally efficient. This research adds valuable knowledge to the discussion on topic analysis and offers useful tools for addressing the ongoing challenges of analyzing text in the digital era.

The future work could involve enhancing our topic modeling frameworks by transitioning them to a non-parametric approach, specifically through the adoption of the hierarchical Dirichlet process (HDP) Bouguila and Ziou (2008, 2012); Fan and Bouguila (2013b); Fan, Bouguila, Du, and Liu (2019); Fan, Yang, and Bouguila (2022). This shift could allow for the automatic determination of the optimal number of topics directly from the data, therefore eliminating the need for manual parameterization of topic numbers. Additionally, another future work could include online learning, such as online variational learning Bdiri, Bouguila, and Ziou (2016); Epaillard and Bouguila (2019); Fan and Bouguila (2014); Manouchehri, Dalhoumi, Amayri, and Bouguila (2020). This method offers the advantage of fast and efficient updates to the model without requiring the processing of the entire data corpus at each iteration. This enables the model to learn continuously from new data inputs. This approach would not only facilitate the analysis process but also ensure that the models

remain adaptable. This would make them highly effective for ongoing and real-time data analysis challenges.

References

Bakhtiari, A. S., & Bouguila, N. (2014a). Online learning for two novel latent topic models. In Linawati, M. S. Mahendra, E. J. Neuhold, A. M. Tjoa, & I. You (Eds.), *Information and communication technology - second IFIP TC5/8 international conference, ict-eurasia 2014, bali, indonesia, april 14-17, 2014. proceedings* (Vol. 8407, pp. 286–295). Springer.

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

Bakhtiari, A. S., & Bouguila, N. (2016). A latent beta-liouville allocation model. *Expert Systems with Applications*, *45*, 260–272. doi: 10.1016/j.eswa.2015.09.044

Bdiri, T., Bouguila, N., & Ziou, D. (2014). Object clustering and recognition using multi-finite mixtures for semantic classes and hierarchy modeling. *Expert Syst. Appl.*, *41*(4), 1218–1235.

Bdiri, T., Bouguila, N., & Ziou, D. (2016). Variational bayesian inference for infinite generalized inverted dirichlet mixtures with feature selection and its application to clustering. *Appl. Intell.*, *44*(3), 507–525.

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

Blei, D. M., & Lafferty, J. D. (2007). A correlated topic model of science. *The Annals of Applied Statistics*, *1*, 17-35. Retrieved from https://api.semanticscholar.org/CorpusID:8872108

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

Bouguila, N. (2011). Count data modeling and classification using finite mixtures of distributions. *IEEE Transactions on Neural Networks*, *22*(2), 186-198. doi: 10.1109/TNN.2010.2091428

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.

Bouguila, N., & Elguebaly, T. (2012). A fully bayesian model based on reversible jump MCMC and finite beta mixtures for clustering. *Expert Syst. Appl.*, *39*(5), 5946–5959.

Bouguila, N., & ElGuebaly, W. (2008a). A generative model for spatial color image databases categorization. In *2008 ieee international conference on acoustics, speech and signal processing* (p. 821-824).

Bouguila, N., & ElGuebaly, W. (2008b). On discrete data clustering. In T. Washio, E. Suzuki, K. M. Ting, & A. Inokuchi (Eds.), *Advances in knowledge discovery and data mining, 12th pacific-asia conference, PAKDD 2008, osaka, japan, may 20-23, 2008 proceedings* (Vol. 5012, pp. 503–510). Springer.

Bouguila, N., & Ziou, D. (2005). Mml-based approach for high-dimensional unsupervised learning using the generalized dirichlet mixture. In *IEEE conference on computer vision and pattern recognition, CVPR workshops 2005, san diego, ca, usa, 21-23 september, 2005* (p. 53). IEEE Computer Society.

Bouguila, N., & Ziou, D. (2008). A dirichlet process mixture of dirichlet distributions for classification and prediction. In *2008 ieee workshop on machine learning for signal processing* (p. 297-302).

Bouguila, N., & Ziou, D. (2012). A countably infinite mixture model for clustering and feature selection. *Knowl. Inf. Syst.*, *33*(2), 351–370.

Caballero Barajas, K., Barajas, J., & Akella, R. (2012, 10). The generalized dirichlet distribution in enhanced topic detection..

Das, R., Zaheer, M., & 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)* (pp. 795–804).

Deerwester, S., Dumais, S., Furnas, G., Landauer, T., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science 41*, 391-407.

Doyle, G., & Elkan, C. (2009). Accounting for burstiness in topic models. In *Proceedings of the 26th annual international conference on machine learning* (p. 281–288). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1553374.1553410 doi: 10.1145/1553374.1553410

Epaillard, E., & Bouguila, N. (2019). Variational bayesian learning of generalized dirichlet-based hidden markov models applied to unusual events detection. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(4), 1034-1047.

Fan, W., & Bouguila, N. (2013a). Learning finite beta-liouville mixture models via variational bayes for proportional data clustering. In F. Rossi (Ed.), *IJCAI 2013, proceedings of the 23rd international joint conference on artificial intelligence, beijing, china, august 3-9, 2013* (pp. 1323–1329). IJCAI/AAAI.

Fan, W., & Bouguila, N. (2013b). Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference. *IEEE Transactions on Neural Networks and Learning Systems*, *24*(11), 1850-1862.

Fan, W., & Bouguila, N. (2013c). Variational learning of finite beta-liouville mixture models using component splitting. In *The 2013 international joint conference on neural networks (ijcnn)* (p. 1-8).

Fan, W., & Bouguila, N. (2014). Online data clustering using variational learning of a hierarchical dirichlet process mixture of dirichlet distributions. In W. Han, M. Lee, A. Muliantara, N. A. Sanjaya, B. Thalheim, & S. Zhou (Eds.), *Database systems for advanced applications - 19th international conference, DASFAA 2014, international workshops: Bdma, damen, SIM - 3 - , uncrowd; bali, indonesia, april 21-24, 2014, revised selected papers* (Vol. 8505, pp. 18–32). Springer.

Fan, W., & 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., Bouguila, N., Du, J.-X., & Liu, X. (2019). Axially symmetric data clustering through

dirichlet process mixture models of watson distributions. *IEEE Transactions on Neural Networks and Learning Systems*, *30*(6), 1683-1694.

Fan, W., Sallay, H., & Bouguila, N. (2017). Online learning of hierarchical pitman-yor process mixture of generalized dirichlet distributions with feature selection. *IEEE Trans. Neural Networks Learn. Syst.*, *28*(9), 2048–2061.

Fan, W., Yang, L., & 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.

Fritz, T., Gonda, T., & Perrone, P. (2021, November). De finetti's theorem in categorical probability. *Journal of Stochastic Analysis*, *2*(4). Retrieved from http://dx.doi.org/10.31390/josa.2.4.06 doi: 10.31390/josa.2.4.06

Griffiths, T. L., & 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.. Retrieved from https://api.semanticscholar.org/CorpusID:7566772

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the twenty-second annual international sigir conference on research and development in information retrieval.*

Hu, C., Fan, W., Du, J., & Bouguila, N. (2019). A novel statistical approach for clustering positive data based on finite inverted beta-liouville mixture models. *Neurocomputing*, *333*, 110–123.

Huang, R., Xu, W., Qin, Y., & Chen, Y. (2020). Hierarchical dirichlet multinomial allocation model for multi-source document clustering. *IEEE Access*, *8*, 109917-109927. doi: 10.1109/ACCESS.2020.3002107

Ihou, K. E., & Bouguila, N. (2017). A new latent generalized dirichlet allocation model for image classification. In *Seventh international conference on image processing theory, tools and applications, IPTA 2017, montreal, qc, canada, november 28 - december 1, 2017* (pp. 1–6). IEEE.

Ihou, K. E., & Bouguila, N. (2020). Stochastic topic models for large scale and nonstationary data. *Engineering Applications of Artificial Intelligence*, *88*, 103364. doi: 10.1016/j.engappai.2019.103364

Jelinek, F., Mercer, R. L., Bahl, L. R., & Baker, J. K. (1977). Perplexity—a measure of the difficulty of speech recognition tasks. *The Journal of the Acoustical Society of America*, *62*(S1), S63–S63.

Luo, Z., Amayri, M., Fan, W., & 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., & Elkan, C. (2005). Modeling word burstiness using the dirichlet distribution. In *Proceedings of the 22nd international conference on machine learning* (p. 545–552). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1102351.1102420 doi: 10.1145/1102351.1102420

Manouchehri, N., Dalhoumi, O., Amayri, M., & Bouguila, N. (2020). Variational learning of a shifted scaled dirichlet model with component splitting approach. In *2020 third international conference on artificial intelligence for industries (ai4i)* (p. 75-78).

Minka, T. P. (2012). Estimating a dirichlet distribution.. Retrieved from https://api.semanticscholar.org/CorpusID:6959923

Newman, D., Lau, J. H., Grieser, K., & 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* (pp. 100–108).

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

Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.

Wallach, H., Mimno, D., & McCallum, A. (2009). Rethinking lda: Why priors matter. In Y. Bengio, D. Schuurmans, J. Lafferty, C. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (Vol. 22). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2009/file/0d0871f0806eae32d30983b62252da50-Paper.pdf

Wei, G., & Tanner, M. (1990, September). A monte carlo implementation of the em algorithm and

the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, *85*(411), 699–704. doi: 10.1080/01621459.1990.10474930

Yang, L., Fan, W., & Bouguila, N. (2022). Clustering analysis via deep generative models with mixture models. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(1), 340-350.

Zamzami, N., & Bouguila, N. (2019a). Model selection and application to high-dimensional count data clustering - via finite EDCM mixture models. *Appl. Intell.*, *49*(4), 1467–1488.

Zamzami, N., & Bouguila, N. (2019b). A novel scaled dirichlet-based statistical framework for count data modeling: Unsupervised learning and exponential approximation. *Pattern Recognit.*, *95*, 36–47.

Zamzami, N., & Bouguila, N. (2022). Sparse count data clustering using an exponential approximation to generalized dirichlet multinomial distributions. *IEEE Transactions on Neural Networks and Learning Systems*, *33*(1), 89-102.