Novel Probabilistic Frameworks for Author-Level Topic Modeling

Faiza Tahsin

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 (Information Systems Security) at

Concordia University

Montréal, Québec, Canada

February 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Faiza Tahsin**

Entitled: **Novel Probabilistic Frameworks for Author-Level Topic Modeling**

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

Master of Applied Science (Information Systems Security)

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. Farnoosh Naderkhani

______________________________ Examiner
Dr. Walter Lucia

______________________________ Examiner
Dr. Farnoosh Naderkhani

______________________________ Supervisor
Dr. Nizar Bouguila

Approved by ______________________________
Chun Wang, Chair
Concordia Institute for Information Systems Engineering

__________ 2025

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

Abstract

Novel Probabilistic Frameworks for Author-Level Topic Modeling

Faiza Tahsin

The increasing complexity of textual data in modern applications, such as social media and academic literature analysis, needs improved topic modeling techniques that capture sparsity, variability, and nuanced author-topic relationships. Because of their rigorous assumptions and inadequate adaptability in representing various data, traditional models generally fail to address these shortcomings. We present two novel probabilistic models, Author Dirichlet Multinomial Allocation with Generalized Distribution (ADMAGD) and Author Beta-Liouville Multinomial Allocation (ABLiMA) to overcome these drawbacks while strengthening the state of author-specific topic modeling. To depict complex author-topic relationships, ADMAGD incorporates the Generalized Dirichlet distribution. For datasets with uneven or absent topic representations, ABLiMA uses the Beta-Liouville distribution to adjust for topic distribution variability and sparsity. By comparing these models to common datasets like the NIPS and 20 Newsgroups datasets, the research presented here demonstrates how well these models manage sparsity, capture complex theme preferences, and generate coherent subjects. The results show that the models can be applied to many situations. Coherence measure and author-topic relationship visualizations further validate their interpretability and usefulness.

Acknowledgments

It has been an incredible journey completing my Master's degree at Concordia University, and I am immensely thankful to all of the people who have helped me along the way. This accomplishment is the result of their support, encouragement, and faith in my skills.

For giving me the chance to pursue my Master's degree and learn more about this exciting field of study, I would like to sincerely thank my supervisor, Dr. Nizar Bouguila. Throughout this journey, his wise counsel, mentoring, and unwavering support have been precious.

I also want to express my gratitude to Dr. Hafsa Ennajari for all of her help, support, and tolerance throughout my research. Her knowledge and insightful remarks have greatly influenced my work and inspired me to overcome obstacles and advance as a researcher.

Above all, I want to express my sincere gratitude to my family for their constant belief in me, love, and support. Throughout my life, and especially during my Master's studies, their support has been a source of inspiration and strength. Without their constant encouragement, this achievement would not have been possible.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Problem statement

Conventional topic modeling techniques have been crucial in textual data analysis and the identification of latent topics in a variety of areas Bdiri, Bouguila, and Ziou (2014); Yang, Fan, and Bouguila (2022). Regardless of their extensive application and fundamental importance, these models encounter significant challenges when applied to contemporary, complex, and diverse datasets. These drawbacks result from their inflexible presumptions and inability to adapt to the nuances of actual textual data, such as unpredictability, sparsity, and complex author-topic relationships. One of the most significant challenges traditional models encounters is sparsity in topic distributions. Many real-world datasets, such as those derived from social media platforms, online reviews, or other forms of short-form content, exhibit sparse or uneven distributions of topics. In these datasets, certain topics may be entirely absent or weakly represented in specific documents. For instance, a social media post might focus on a niche subject, making it difficult for traditional models to capture patterns due to their assumptions of comprehensive topic representation across documents. This sparsity leads to suboptimal performance, as traditional approaches often fail to identify less prominent but crucial thematic elements. Another critical limitation is the variability in author contributions. In datasets with distinct authorship, such as academic papers or journalistic articles, individual authors often exhibit unique thematic preferences and styles. Capturing these preferences is essential for understanding the underlying structure of the dataset. However, traditional models, which do not explicitly model author-specific dynamics or thematic variations, struggle to account for this variability. As a result, they fail to provide a clear picture of how different authors contribute to or influence the thematic composition of the dataset. In addition to sparsity and variability, traditional models suffer from inadequate flexibility. They operate under assumptions of even topic distributions and fixed relationships between authors and topics. Although these presumptions make computation easier, they are not appropriate for the dynamic, diverse form of contemporary textual data. Real-world datasets frequently show complex relationships between topics that change over time, interact with one another, or fluctuate greatly depending on the environment. Fixed models neglect these dynamic interactions, leading to results that may not align with the data's actual structure. Finally, limited interpretability is the outcome of these defects. Conventional models can occasionally offer imprecise or inconsistent results, which reduces the significance for future applications such as content recommendation, sentiment analysis, and authorship identification. The models' inability to adapt to the specifics of the data leads to a discrepancy between the topics that are generated and the thematic structures that are found in the dataset, which results in this lack of interpretability. These challenges show how urgently new advanced frameworks that can get around the limitations of traditional subject modeling are needed. Effective analysis of contemporary textual groups requires robust and adaptable models that can manage sparsity, account for authorship fluctuation, and capture fluid relationships between authors and topics. Such developments would allow for wider applications in fields ranging from academic research and authorship attribution to social media analysis and content suggestion, in addition to enhancing the quality and applicability of topic modeling results. Next-generation models have the potential to revolutionize the way we study and understand textual data by filling in these gaps and revealing deeper insights.

1.2 Theoretical background and related works

1.2.1 Fundamentals

Topic modeling is a statistical technique designed to uncover latent thematic structures within large collections of textual data. It is predicated on the idea that articles are made up of several subjects, each of which is a probabilistic distribution across a word vocabulary. Topic modeling uncovers hidden themes through studying word co-occurrence patterns both inside and between documents, providing important information about the composition and content of a dataset. The document-topic distribution, which depicts each document as a collection of topics with proportions indicating the dominance of each topic inside the document, is a basic component of topic modeling. The document's thematic composition is made more apparent by this distribution. Similarly, topic-word distributions characterize topics as probabilistic distributions over a vocabulary, with higher-probability words providing semantic cues about the topic's meaning. The generative process underlying topic modeling explains how documents are created. This typically involves sampling a topic for each word in a document based on the document's topic distribution and generating words from the selected topic's word distribution. The introduction of probabilistic models marked a pivotal moment in the evolution of topic modeling. Latent Semantic Analysis (LSA) introduced by Deerwester, Dumais, Furnas, Landauer, and Harshman (1990), which reduced the dimensionality of term-document matrices and revealed hidden semantic connections using singular value decomposition (SVD), was one of the first innovations. But because LSA was based on linear algebra and lacked a probabilistic basis, it was difficult to interpret in terms of probabilities and was vulnerable to noise. Probabilistic Latent Semantic Analysis (PLSA) was created to address the shortcomings of LSA. By assuming that documents were produced using a combination of latent subjects, each represented as a probability distribution over words, this model provided a probabilistic approach. Although PLSA enhanced interpretability, it was not appropriate for larger datasets because of overfitting and a lack of a defined generative process.

Latent Dirichlet Allocation (LDA), introduced by Blei, Ng, and Jordan (2003), is one of the most widely used frameworks in topic modeling. It assumes that document-topic and topic-word distributions follow Dirichlet priors, enabling flexible topic proportions. LDA employs Bayesian inference techniques, such as Gibbs Sampling or Variational Inference, to estimate latent distributions. While LDA has proven effective in capturing thematic structures, its rigid priors and assumption of uniform topic distributions often limit its adaptability to complex datasets. Building on LDA, the Author-Topic Model (ATM) Rosen-Zvi, Griffiths, Steyvers, and Smyth (2004) incorporates authorship information into the generative process. ATM assumes that each author has a unique distribution over topics, which influences the topic composition of the documents they write. This allows ATM to account for author-specific thematic preferences, enhancing its utility for analyzing datasets with distinct authorship. However, ATM retains many of LDA's limitations, particularly in managing sparsity and variability in topic distributions. Traditional models like LDA and ATM rely heavily on fixed priors and assume an even representation of topics across documents. These assumptions reduce their effectiveness in analyzing datasets characterized by sparse topics, uneven thematic distributions, or complex relationships between authors and topics. Such limitations emphasize the need for more advanced and flexible models capable of adapting to the intricate dynamics of modern textual datasets.

1.2.2 Literature review

Topic modeling has long been a foundational technique in natural language processing, offering a probabilistic framework to analyze and interpret textual data. Over the years, researchers have developed several models to address the challenges of extracting meaningful themes from diverse datasets. This section reviews foundational and advanced models, highlighting their contributions, limitations, and relevance to the proposed approaches in this thesis.

Foundational Models

Latent Dirichlet Allocation (LDA) and its extensions, such as the Author-Topic Model (ATM), have been widely used for topic modeling. While LDA assumes that documents are mixtures of topics and topics are distributions over words, ATM incorporates authorship information, modeling each author's thematic preferences. Despite their significance, these models face limitations:

- **Sparsity:** LDA struggles with sparse datasets, where certain topics are absent or weakly represented.
- **Independent Topics:** LDA assumes topic independence, which is unrealistic for datasets with interrelated themes.
- **Authorship Representation:** ATM lacks flexibility in capturing nuanced author-specific variations in writing styles and word choices. To address these limitations, researchers have explored more flexible probabilistic distributions and advanced modeling techniques.

Advanced models with Generalized Dirichlet Distribution have emerged as a powerful alternative for improving topic modeling frameworks. Unlike the Dirichlet distribution, the Generalized Dirichlet allows for richer representations of dependencies between topics, enhancing coherence and interoperability Luo, Amayri, Fan, Ihou, and Bouguila (2024); Ihou and Bouguila (2019).

Correlated Topic Models (CTM) (Blei & Lafferty, 2007) extend LDA by incorporating topic correlations, enabling the modeling of interrelated themes. However, CTM does not address sparsity or author-specific contributions. Zero-Inflated Latent Dirichlet Allocation (zinLDA) Tang and Chen (2019) utilizes the Generalized Dirichlet distribution to handle structural zeros, demonstrating its versatility in applications such as microbiome analysis. Smoothed Generalized Dirichlet Models Najar and Bouguila (2022) improve topic detection in sparse datasets, particularly those with bursty and uneven count data. These models highlight the flexibility of the Generalized Dirichlet distribution, but they primarily focus on content structure rather than author-specific dynamics.

Advanced models with Beta-Liouville Distribution have been introduced to address sparsity and variability in topic distributions, making it particularly effective for datasets like social media and short-form content.

Latent Beta-Liouville Allocation Model (LBLAM) (Bakhtiari & Bouguila, 2016) enhances topic modeling by incorporating Beta-Liouville priors, capturing latent structures in high-dimensional and count data. Amirkhani, Manouchehri, and Bouguila (2021) proposed a Birth-Death MCMC approach for multivariate Beta mixture models in medical applications. Online learning models (Bakhtiari & Bouguila, 2014a) utilize Beta-Liouville distributions to update topic distributions in real-time, catering to dynamic datasets such as social media feeds and news articles. Expectation Propagation models (Fan & Bouguila, 2015) demonstrate the efficiency of Beta-Liouville distributions in document clustering and proportional data modeling, especially in sparse and skewed datasets. The infinite Liouville mixture model has been applied to text and texture categorization Bouguila (2012). These models showcase the potential of the Beta-Liouville distribution in advanced topic modeling but lack integration with author-specific information.

1.3 Contributions

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

- **Author Dirichlet Multinomial Allocation Model with Generalized Distribution (ADMAGD):** This research was accepted at the 11th International Symposium on Networks, Computers and Communications (ISNCC'24) Tahsin, Ennajari, and Bouguila (2024).

- **Author Beta-Liouville Multinomial Allocation Model (ABLiMA):** This research was accepted at the 27th International Conference on Enterprise Information Systems (ICEIS'25) Tahsin, Ennajari, and Bouguila (2025).

1.4 Thesis overview

- In chapter 1, we introduce the fundamental concepts of topic modeling, tracing its evolution from early clustering methods to modern probabilistic approaches.

- In chapter 2, we present the Author Dirichlet Multinomial Allocation with Generalized Distribution (ADMAGD) model. This chapter focuses on how the integration of Generalized Dirichlet distribution enhances the modeling of complex dependencies between authors and topics.

- In chapter 3, we introduce the Author Beta-Liouville Multinomial Allocation (ABLiMA) model, emphasizing its use of the Beta-Liouville distribution to handle sparsity and variability in topic distributions.

- In chapter 4, we summarize the main findings and contributions of this thesis, highlighting how ADMAGD and ABLiMA address the limitations of traditional topic modeling frameworks. We reflect on the practical applications of the proposed models and suggest future research directions, including hybrid modeling approaches, scalability enhancements, and applications in multilingual and dynamic datasets.

Chapter 2

Author Dirichlet Multinomial Allocation Model with Generalized Distribution (ADMAGD)

2.1 Introduction

Topic modeling is a robust technique in natural language processing (NLP) and machine learning that aims to reveal latent topic structures within large textual datasets Bakhtiari and Bouguila (2014b); Blei (2012); Blei et al. (2003). Topic modeling algorithms allow researchers to extract meaningful insights, facilitate document organization, and support a variety of downstream tasks such as document clustering, information retrieval, classification, and recommendation systems by automatically identifying recurring word patterns across documents Ennajari, Bouguila, and Bentahar (2021). Variational learning of finite scaled Dirichlet mixture models has been explored for data clustering Nguyen, Azam, and Bouguila (2019). Zamzami, Alsuroji, Eromonsele, and Bouguila (2020) proposed a proportional data modeling approach using a finite mixture of scaled Dirichlet distributions. In topic modeling, documents are assumed to be composed of distinct topics, each characterized by its distribution of words. Each topic is defined by a probability distribution across the vocabulary of the corpus, indicating the likelihood of each word being associated with that topic.

In this context, the Dirichlet distribution Bouguila and Ziou (2006) is commonly used to model the distribution of topics over a set of documents, where it serves as a prior distribution for the topic proportions of each document. Bouguila and Ziou (2005a) introduced an MML-based approach for estimating and selecting finite Dirichlet mixtures. Bouguila and Ziou (2005c) They also proposed an approach for fitting finite Dirichlet mixtures using ECM and MML. Authorship is a vital attribute in any text. Traditional topic models often struggle to capture the diverse and nuanced aspects of textual data, such as the varying writing styles of different authors, the evolution of topics over time, and the presence of ambiguous or polysemous words. The Latent Dirichlet Allocation (LDA) model does not consider the information about text's authorship. Although the author-topic model attempted to incorporate this attribute, it remains limited, particularly in capturing the complexities of the author-topic relationship. In this chapter, we introduce a novel probabilistic topic model, ADMAGD, designed to address these limitations by capturing complex author-topic relationships effectively. Our model leverages the Generalized Dirichlet distribution to account for the variability in writing styles and topic preferences among different authors Epaillard and Bouguila (2018); Fan, Sallay, and Bouguila (2016); Bouguila and Ziou (2005b); Ihou and Bouguila (2017). This distribution has a more flexible covariance structure, allowing for richer dependencies between topics within a document, which can better capture the nuanced ways authors combine topics in their writing. It also provides more control over the variability in topic proportions across documents from different authors. Consequently, ADMAGD can more accurately reflect the subtle variations and patterns in the data, leading to improved topic coherence and interpretability. Extensive experiments across multiple datasets demonstrate that ADMAGD effectively detects intricate patterns in authors' writing on a wide range of topics. The rest of this chapter is structured as follows: Section 2 presents the proposed ADMAGD model. In Section 3, we provide a detailed explanation of the Gibbs sampling approach, which is utilized to infer the model parameters, experiments and results of our model on the 20-newsgroup and NIPS datasets, respectively.

2.2 Proposed model

Model description

In topic modeling, the Dirichlet distribution is fundamentally used as a prior distribution for the topic proportions in documents and for the word distributions within topics Fan and Bouguila (2012); Bouguila (2007). This distribution is parameterized by a vector of positive reals, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ that determines the cluster of the distribution within k categories. The Dirichlet distribution is a conjugate prior for the multinomial distribution Bouguila and Ziou (2007). This property simplifies the computation of posterior distributions, making the inference process more tractable. The probability Density Function of Dirichlet distribution for a vector $\boldsymbol{x} = (x_1, \ldots, x_K)$, where each x_k = proportion of category k :

$$f(x; \alpha) = \frac{1}{B(\alpha)} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

where $B(\alpha)$ is the multivariate Beta function, defined as:

$$B(\alpha) = \frac{\prod_{k=1}^{K} \Gamma(\alpha_k)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}$$

and α_k are the parameters that shape the distribution.

The Generalized Dirichlet distribution (GD) is defined for a vector of probabilities $\boldsymbol{x} = (x_1, \ldots, x_K)$ and is parameterized by two vectors $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)$.

The probability density function is given by:

$$f(x; \alpha, \beta) = \frac{\Gamma(\sum_{k=1}^{K} (\alpha_k + \beta_k))}{\prod_{k=1}^{K} \Gamma(\alpha_k + \beta_k)} \times \prod_{k=1}^{K} \pi_k^{\alpha_k - 1} \times \prod_{k=1}^{K} \left(1 - \sum_{i=1}^{k} \pi_i\right)^{\beta_k - \alpha_{k+1}}$$

In our proposed ADMAGD model, we assume that both topic and word distributions are drawn from a Generalized Dirichlet distribution. Formally, we are assuming following generative process for ADMAGD:

Table 2.1: Summary of Mathematical Notations

Notation	**Meaning**
ϕ_k	The word distribution for topic k.
a_k, b_k	Parameters of the generalized Dirichlet distribution for the word distribution within topic k.
θ_a	The topic distribution for author a.
α, β	Hyperparameters for the Dirichlet priors for word distributions within topics and topic distributions within authors, respectively.
$z_{d,i}$	The topic assigned to the i-th word in document d.
$W_{d,i}$	The i-th word in document d.

- Topic-level Word distributions

  For each topic k:

  A word distribution ϕ_k is drawn from a generalized Dirichlet distribution with parameters a and b. The word distribution ϕ_k is modeled as:

$$\phi_k \sim \text{Generalized Dirichlet}(a_k, b_k)$$

  The probability density for ϕ_k:

$$p(\phi_k \mid a_k, b_k) \propto \prod_{w=1}^{W} \phi_{kw}^{a_{kw}-1} \left(1 - \sum_{i=1}^{w} \phi_{ki}\right)^{b_{kw}-1}$$

- Author-level Topic distributions

  For each author a:

  A topic distribution θ_a is drawn from a Generalized Dirichlet distribution with parameters α and β. θ_a, modeled as:

$$\theta_a \sim \text{Generalized Dirichlet}(\alpha_a, \beta_a)$$

  The probability density for θ_a is:

$$p(\theta_a \mid \alpha_a, \beta_a) \propto \prod_{k=1}^{K} \theta_{ak}^{\alpha_{ak}-1} \left(1 - \sum_{i=1}^{k} \theta_{ai}\right)^{\beta_{ak}-1}$$

- Document-level Topic selection

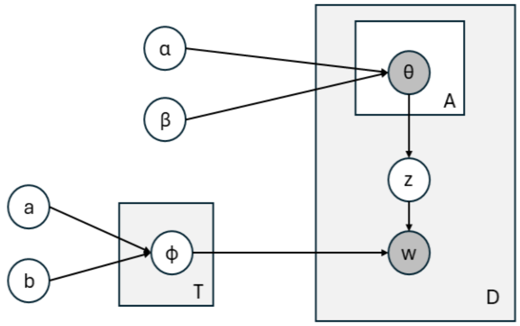

Figure 2.1: Graphical model for ADMAGD

For each document d and author a:

A topic z_d is taken from the distribution over topics θ_a.

$$Z_{d,n} \mid \theta_a \sim \text{Multinomial}(\theta_a)$$

The probability of assigning topic k to a word in this context is given by:

$$p(Z_{d,n} = k \mid \theta_a) = \theta_{ak}$$

- Word generation in Documents

  For each word w in document d:

  A word w is drawn from the word distribution ϕ_{z_d}.

The joint probability distribution of the ADMAGD model:

$$p(\Theta, \Phi, Z, W \mid \alpha, \beta, a, b) = \left(\prod_{a=1}^{A} p(\theta_a \mid \alpha_a, \beta_a)\right) \times \left(\prod_{k=1}^{K} p(\phi_k \mid a_k, b_k)\right) \times \left(\prod_{d=1}^{D} \prod_{n=1}^{N_d} p(Z_{d,n} \mid \theta_a) \times p(W_{d,n} \mid \phi_{Z_{d,n}})\right)$$

Figure 2.1 illustrates the graphical representation of our model.

Model setup

The model has been configured with a predetermined number of topics, which acts as a foundation for the process of topic modelling. Selecting and initializing various parameters and hyperparameters are done to configure the model. A topic word distribution displays the predetermined number of topics.

In our model, we assume that authors are associated with probability distributions over topics, which indicates how likely they are to get involved in discussions about specific topics. Next, the hyperparameters that influence the distributions of topics and words, as well as the specific parameters which are unique to ADMAGD, were initialized.

Corpus During the initialization phase, ADMAGD obtains a corpus that includes a collection of documents. Each author is associated with one or more documents in the corpus.

Mapping and Hyperparameter Two mappings, id to word and author to doc, are key elements of its configuration: id to word is a dictionary that associates different identifiers with words, and author to doc connects authors to their respective documents.

The model is additionally parameterized with hyperparameters, namely α and β which have a significant impact on the distributions of topics and words. Furthermore, the inclusion of a and b as parameters for the generalized Dirichlet distribution distinguishes ADMAGD from conventional models. Boukhers and Staab (2020) claim that hyperparameters are crucial in determining topic model output. In their research, they have shown that by adjusting the values of the hyperparameters, the coherence of the topics can be significantly impacted, also the quality of the model can be improved.

Model fitting

In order to infer the model hidden parameters, we developed a Gibbs Sampling approach. It is a Markov Chain Monte Carlo (MCMC) method that is well-suited for complex probabilistic models and allows for efficient estimation of the posterior distributions of the model parameters Fan and Bouguila (2012); Bouguila and Elguebaly (2012); Elguebaly and Bouguila (2010). This iterative sampling method is vital for understanding the latent thematic structures within the corpus. The algorithm begins by randomly assigning topics to each of the words in each document, following a distribution of probabilities which is uniform across the entire set of topics.

Iterative Sampling Process Repeat for a specified number of iterations or until convergence:

- For each document d: For each word $w_{d,n}$ in document d:
- Remove Current Topic Assignment: Temporarily remove the current topic assignment $z_{d,n}$ of word $w_{d,n}$ and update the count matrices accordingly.
- Update count matrices:

$$\begin{aligned} N_k^{-d,n} &: \text{Count of words assigned to topic } k, \\ N_{k,w}^{-d,n} &: \text{Count of word } w \text{ assigned to topic } k, \\ N_{a,k}^{-d,n} &: \text{Count of topic } k \text{ for author } a \end{aligned}$$

- Compute Conditional Distribution:

$$\begin{aligned} P(z_{d,n} = k \mid z_{-d,n}, W, \Theta, \Phi) \propto (&\theta_{a,k}^{(g)} + \alpha_k - 1) \\ &\times (\phi_{k,w_{d,n}}^{(g)} + \beta_{w_{d,n}} - 1) \end{aligned}$$

  where $\theta_{a,k}^{(g)}$ and $\phi_{k,w_{d,n}}^{(g)}$ are calculated considering the generalized Dirichlet parameters.

- Sample New Topic: Draw a new topic $z_{d,n}$ for word $w_{d,n}$ based on the normalized conditional probabilities.

- Update Count Matrices:

$$
\begin{aligned}
N_k &: \text{Count of words assigned to topic } k, \\
N_{k,w} &: \text{Count of word } w \text{ assigned to topic } k, \\
N_{a,k} &: \text{Count of topic } k \text{ associated with author } a
\end{aligned}
$$

Compute Final Distributions Calculate the final topic distributions θ and word distributions ϕ after the last iteration:

$$
\begin{aligned}
\theta_{a,k} &= \frac{N_{a,k} + \alpha}{\sum_{k'} (N_{a,k'} + \alpha)} \\
\phi_{k,w} &= \frac{N_{k,w} + \beta}{\sum_{w'} (N_{k,w'} + \beta)}
\end{aligned}
$$

The conditional probability of the topic assignment k to word $W_{d,n}$ in document d and author a is given by:

$$
P(Z_{d,n} = k \mid Z_{-d,n}, W, \Theta, \Phi) \propto \theta_{a,k} \times \phi_{k,w}
$$

The first part denotes the probability that a given word $W_{d,n}$ will be assigned to a specific topic k by previous word-topic assignments. The last part computes the sum of all documents d to account for the influence of each author's preference towards topic k.

Convergence This procedure is repeated by multiple iterations for each word in the corpus until the topic assignment converges. Generally, convergence is assessed according to the consistency of the topic distributions across consecutive iterations.

After completing the Gibbs sampling iterations, the final topic assignments are used to compute the posterior distributions of the model parameters, the topic-author distribution (θ) and the word-topic distribution (ϕ). The distributions are obtained by adjusting the count matrices using the corresponding summations and the hyperparameters of the generalized Dirichlet distribution.

2.3 Experiments

2.3.1 Datasets and setup

To evaluate the performance of the Author Dirichlet Multinomial Allocation Model with Generalized Distribution (ADMAGD), we conducted a series of experiments on two widely-used benchmark datasets: 20-newsgroups and NIPS. These datasets were selected due to their varied authorship patterns and rich topic structures.

The Newsgroup dataset comprises an estimated 20,000 newsgroup documents, partitioned across 20 different newsgroups. The data is obtained from an assortment of newsgroups, covering a wide-ranging collection of topics such as technology, athletics, politics, and religion. It is a benchmark for the classification of texts and topic modeling tasks due to its extensive variety Lang (1995). The diversity and association of each document with certain authors in the dataset facilitated evaluating the robustness and flexibility of the ADMAGD model.

The NIPS dataset Kaggle (n.d.) on the other hand, consists of 1740 papers from the Neural Information Processing Systems (NIPS) conferences, with metadata including authorship information. This dataset is particularly suited for exploring complex author-topic relationships due to the high variability in author contributions across different topics.

All datasets were preprocessed to remove noise and less important content and focus on the main text Bird, Klein, and Loper (2009) which includes eliminating headers, footers and quotes from the documents. We also performed Tokenization, stop word removal, and Lemmatization in order to concentrate solely on the primary content. We created a dictionary containing 5315 words by filtering out tokens that come in less than 15 documents or more than 50% of the documents. Subsequently, we represented each document as a TF-IDF vector.

2.3.2 Experimental Results

Table 2.2 demonstrates an example of 6 topics, out of a total of 20, that was obtained by the model for the 20-newsgroup dataset. We derived the topics from a sample that was collected during the 200th iteration of the Gibbs sampling algorithm. The summary of each topic provides the top 10 words which are the most probable outcomes based on the topic, along with their respective probabilities, that are likely to be generated. In Topic 7, the top words (God, Christian Jesus, Bible) are highly likely to occur frequently (prob. 0.0077, 0.0058, 0.0052, 0.0043) in the topic referring to religion.

Table 2.2: Word Probabilities per Topic on 20 NewsGroup.

TOPIC 1		TOPIC 2		TOPIC 5	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
Window	0.0083	Year	0.0034	People	0.0078
Run	0.0070	Space	0.0034	Israel	0.0064
File	0.0068	New	0.0031	Right	0.0058
Problem	0.0067	Research	0.0029	State	0.0055
Work	0.0060	Development	0.0025	Israeli	0.0055
Program	0.0059	Science	0.0025	Country	0.0044
Try	0.0055	Information	0.0023	Jew	0.0043
Look	0.0047	World	0.0023	Arab	0.0043
Help	0.0046	Program	0.0023	Good	0.0042
Write	0.0041	Write	0.0022	Way	0.0042
TOPIC 7		**TOPIC 10**		**TOPIC 14**	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
God	0.0077	Email	0.0047	Game	0.0096
Christian	0.0058	Software	0.0040	Team	0.0090
People	0.0053	Send	0.0037	Year	0.0080
Jesus	0.0052	Computer	0.0037	Good	0.0076
Thing	0.0046	List	0.0035	Player	0.0065
Believe	0.0045	Ftp	0.0035	Play	0.0065
Bible	0.0043	Include	0.0033	Season	0.0050
Question	0.0042	Mail	0.0033	League	0.0049
Way	0.0040	Work	0.0033	Look	0.0048
Good	0.0040	Help	0.0032	Run	0.0041

Table 2.3 displays the two most prominent topics associated with each author. The $Topics$ column shows pairs of numbers that represent the two topics that are most common or frequent in the writings of each author, according to the topic model's analysis. From the table, we can see some renowned authors (e.g., Guy Kuo, Joe Green, Jonathan McDowell, and Brian Manning Delaney) and their interests in the area of topics. It can be seen that Joe Green refers to two topics; the first one is related to religious beliefs, which means this is the topic the author is more likely to write about. For the second topic, an email was found written by Joe Green about the graphic chip, which is why he also referred to the computer graphic topic.

Table 2.4 shows the top words most likely to occur in NIPS dataset. In Topic 5 (Recognition,

Table 2.3: Author-Topic Distribution in 20 NewsGroup.

Author	Topics
guykuo@carson.u.washington.edu (Guy Kuo)	9, 6
twillis@ec.ecn.purdue.edu (Thomas E Willis)	17, 12
jgreen@amber (Joe Green)	7, 20
jcm@head-cfa.harvard.edu (Jonathan McDowell)	15, 20
jcm@head-cfa.harvard.edu (Jonathan McDowell)	2, 16
bmdelane@quads.uchicago.edu (Brian Manning Delaney)	8, 2
bgrubb@dante.nmsu.edu (GRUBB)	6, 9
holmes7000@iscsvax.uni.edu	10, 19
kerr@ux1.cso.uiuc.edu (Stan Kerr)	1, 15
irwin@cmptrc.lonestar.org (Irwin Arnstein)	3, 20
...	...

Table 2.4: Word Probabilities per Topic in NIPS.

TOPIC 2		**TOPIC 5**		**TOPIC 6**	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
Noise	0.0033	Recognition	0.0048	Norm	0.0047
Recover	0.0031	Vision	0.0047	Convex	0.0045
Dimensional	0.0030	Image	0.0042	Descent	0.0040
Iid	0.0025	Cvpr	0.0042	Minimization	0.0035
Entry	0.0025	Object	0.0038	Regularization	0.0034
high	0.0025	Visual	0.0037	Operator	0.0032
Row	0.0025	Convolutional	0.0032	regularize	0.0031
Noisy	0.0025	pixel	0.0031	Continuous	0.0030
Furthermore	0.0024	Extract	0.0029	Converge	0.0030
Signal	0.0024	Classification	0.0028	Write	0.0029

TOPIC 7		**TOPIC 8**		**TOPIC 10**	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
Intelligence	0.0032	Bengio	0.0052	Likelihood	0.0057
Determine	0.0030	Deep	0.0049	Inference	0.0051
Node	0.0028	Architecture	0.0047	Bayesian	0.0051
Graph	0.0027	Layer	0.0045	Posterior	0.0047
Tree	0.0025	Preprint	0.0042	Marginal	0.0040
Search	0.0025	Hinton	0.0040	Latent	0.0039
Share	0.0025	Hidden	0.0034	Markov	0.0039
Artificial	0.0024	Unit	0.0033	Density	0.0038
Associate	0.0024	Kingma	0.0033	Variational	0.0036
Probabilistic	0.0024	Recurrent	0.0033	Family	0.0035

Table 2.5: Author-Topic Distribution in NIPS.

Author	Topics
Sebastian Stober	0, 7, 6
Daniel J. Cameron	9, 8, 7
Jessica A. Grahn	9, 8, 7
Aurel A. Lazar	0, 1, 6
Yevgeniy Slutskiy	9, 8, 7
Chen-Yu Wei	8, 2, 6
Yi-Te Hong	9, 8, 7
Chi-Jen Lu	9, 8, 7
Katherine A. Heller	9, 5, 7
David B. Dunson	1, 9, 5
...	...

Vision, Image, Object), these words have a higher probability rate (0.0048, 0.0047, 0.0042, 0.0038).

Table 2.5 shows the authors (e.g., Sebastian Stober, Jessica A. Grahn, David B. Dunson) and their topics of interest.

We also compared the performances of our model with the Author-Topic model Rosen-Zvi et al. (2004) and the Latent Dirichlet Allocation model Blei et al. (2003).

From Table 2.6, we can see that some words contain zero-weight probability. Also, the topics are less coherent and include some frequent and less informative words.

In Table 2.7, we are using the news agency companies as the authors. shows most authors have strong preferences for certain topics. Because of this, the authorship association exhibits less variability.

The topic distribution in ADMAGD is more balanced than in the ATM model, with distinct author thematic preferences. Authors cover a variety of primary and secondary topics, presenting a more complete picture of their thematic focus.

Coherence Score

In the evaluation of the topic model, the coherence score is often used by considering the frequency of word co-occurrences in documents. The u_mass measure is highly useful for its straightforwardness and direct utilization of document frequencies Mimno, Wallach, Talley, Leenders, and

Table 2.6: Word Probabilities per Topic ATM in 20 NewsGroup.

TOPIC 1		TOPIC 2		TOPIC 4	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
News	0.032	President	0.010	Trump	0.0037
Reuters	0.016	Trump	0.008	State	0.0012
Trump	0.010	Year	0.007	President	0.0011
Business	0.008	New	0.007	Clinton	0.007
World	0.008	House	0.006	Campaign	0.006
Percent	0.007	State	0.006	Vote	0.006
State	0.007	Time	0.005	Republican	0.006
Market	0.007	City	0.005	Party	0.005
President	0.006	Officials	0.005	House	0.005
Company	0.006	Include	0.005	Republicans	0.005

TOPIC 9		TOPIC 12		TOPIC 15	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
Super	0.000	Archiveteam	0.000	Archiveteam	0.000
Like	0.000	Like	0.000	Company	0.000
Peak	0.000	Company	0.000	Article	0.000
New	0.000	People	0.000	Facebook	0.000
Time	0.000	New	0.000	Time	0.000
Play	0.000	Time	0.000	Future	0.000
Facebook	0.000	Write	0.000	Like	0.000
Learn	0.000	Work	0.000	New	0.000
Company	0.000	Year	0.000	Group	0.000
Story	0.000	Article	0.000	Story	0.000

McCallum (2011). It is defined as:

$$\text{Coherence} = \frac{1}{M}\sum_{i=2}^{N}\sum_{j=1}^{i-1}\log\frac{D(w_i, w_j)+1}{D(w_j)}$$

Figure 3.2 displays the UMass coherence score for each of the top words. The UMass score ranges from -14 to 14, indicating a modest coherence score for our top words. It is also noticed that more or less the number of top words alters the coherence score, thus the proper amount of top words should be utilized to retain the quality of topics generated by the model.

Qualitative Analysis

We have manually inspected the topics generated by the model, based on the technique of how humans interpret topic models by Chang, Gerrish, Wang, Boyd-Graber, and Blei (2009), which

Table 2.7: Author-Topic Distribution ATM in 20 NewsGroup.

Author	Topics
Atlantic	1, 4, 18
Breibart	1, 4, 18
Business Insider	1, 2, 4, 18
Buzzfeed News	1, 2, 4, 18
CNN	2, 4, 18
Fox News	1, 2, 4, 18
Los Angeles Times	2, 18
NPR	1, 2, 4, 18
New York Post	2, 4, 18
New York Times	2, 4, 18
...	...

Table 2.8: Word Probabilities per Topic LDA in 20 NewsGroup.

TOPIC 1		**TOPIC 2**		**TOPIC 4**	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
Image	0.017	Gun	0.012	Need	0.009
File	0.011	File	0.011	Use	0.008
Use	0.010	Use	0.011	Gun	0.007
Bike	0.010	Make	0.008	State	0.007
Know	0.006	Know	0.008	Like	0.007
Good	0.006	Like	0.008	Dod	0.006
Like	0.005	Say	0.008	Apr	0.006
Email	0.005	Right	0.007	File	0.006
Jpeg	0.005	Dod	0.006	Say	0.006
Just	0.005	Just	0.006	Make	0.005

TOPIC 6		**TOPIC 8**		**TOPIC 9**	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
Say	0.008	Make	0.0012	Bike	0.0016
Fbi	0.008	Law	0.008	Like	0.0010
Child	0.008	Right	0.008	Just	0.008
Compound	0.007	Good	0.008	Time	0.008
Make	0.007	Time	0.007	Dog	0.007
Batf	0.006	Use	0.007	Good	0.007
Come	0.006	Like	0.006	Right	0.006
Start	0.005	Public	0.006	Make	0.006
Roby	0.005	Country	0.006	Turn	0.005
Day	0.005	Say	0.006	Know	0.005

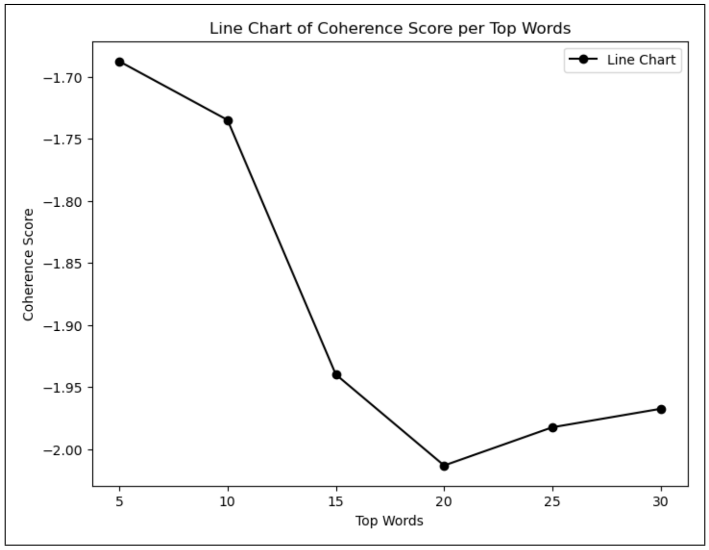

Figure 2.2: Coherence Score per Top Words in 20 NewsGroup.

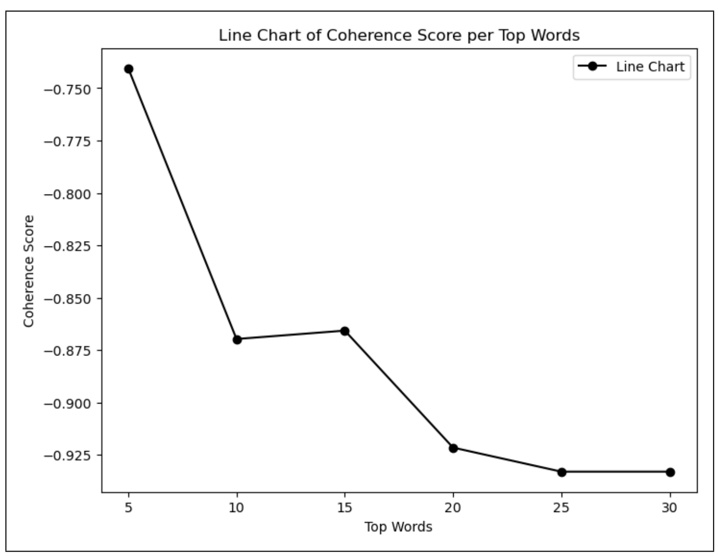

Figure 2.3: Coherence Score per Top Words in NIPS.

involves the authorship attribute analysis and how the model accurately represents the differences in topics among various authors. This analysis can provide important insights into the topic emphasis, the writing style of each author and the evolution over time.

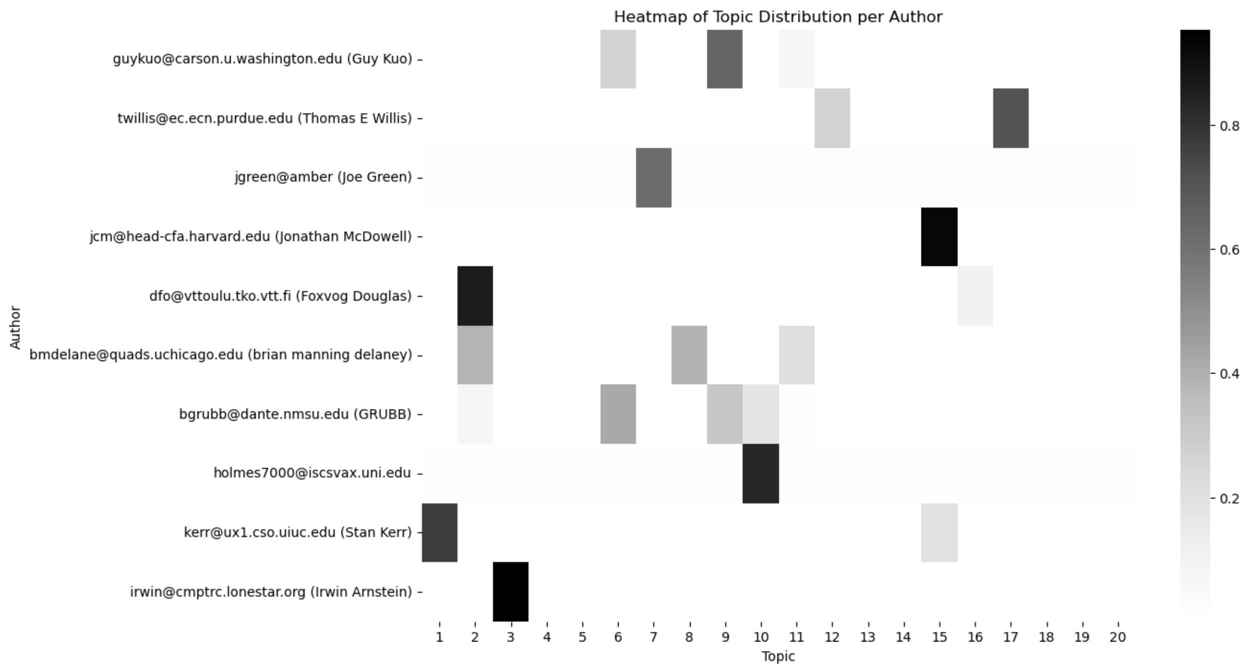

Figure 2.4: Topic Distribution per Author in 20 NewsGroup.

After training the ADMAGD model, we extracted the topic distributions for each author and their probability distribution that reflects their preferences in various topics. Then, we manually inspected and compared the topic distributions across multiple authors to identify differences in their thematic focus. Then, we assessed whether the topics assigned by our model correspond closely with what we perceive in their works. The heatmaps in figures 4 and 5 illustrate which topics are more prominent and how they are distributed across the documents or authors.

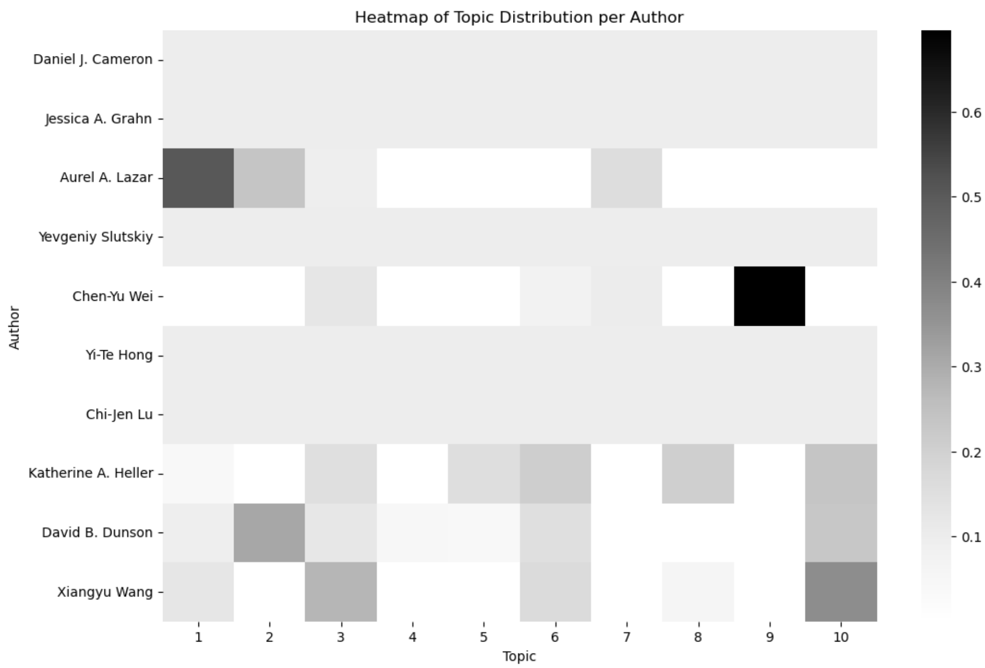

Figure 2.5: Topic Distribution per Author in NIPS.

Chapter 3

Author Beta-Liouville Multinomial Allocation Model (ABLiMA)

3.1 Introduction

The rapidly expanding field of text analytics has made topic modeling a vital technique, enabling the extraction of thematic structures from vast text corpora. Conventional models, such Latent Dirichlet Allocation (LDA) Blei et al. (2003), have improved the understanding of latent topics in texts by claiming that each document comprises a fixed number of topics. Nonetheless, fixed attributes and shortcomings of these models to tackle topic scarcity and the fluctuating relevance of topics across documents provide significant challenges, particularly in the analysis of social media and other forms of dynamic textual data. Recent improvements in probabilistic topic modeling seek to address these limitations by using more flexible distributions that more accurately represent the complex structure of real-world textual data. In this context, we propose the Author Beta-Liouville Multinomial Allocation (ABLiMA) model, which integrates the Beta-Liouville distribution to provide an advanced approach to topic modeling. This model outperforms traditional frameworks by allowing topic proportions to be less than one, hence offering a more precise representation of topic absence and sparsity, a common feature in many current datasets.

In addition to flexibly modeling topic proportions, ABLiMA incorporates the influence of author-specific factors on topic distribution throughout the modeling process. It emphasizes that authors may possess distinct topic perspectives that strongly influence the content. This attribute is essential in contexts where the author's identity impacts the material, such as academic literature, journalistic articles, and especially in social media, where personal expression and individual differences are significant.

The incorporation of the Beta-Liouville distribution in ABLiMA addresses the absence of topics and allows for a more flexible response to varying levels of author engagement with specific topics. This capability is particularly beneficial for datasets with high diversity. It enables the model to competently manage the different distributions of topics across texts, leading to improved precision compared to conventional models.

Our contributions in this chapter are as follows:

- We introduce the ABLiMA model, a novel approach to author-topic modeling that integrates the Beta-Liouville distribution, enabling more flexible and accurate representation of topic distributions.

- We showcase the effectiveness of Beta-Liouville priors in capturing the complex dynamics of thematic structures and author-specific preferences, efficiently addressing challenges related to sparsity and thematic diversity.

- Through comprehensive experiments on the 20 Newsgroups and NIPS datasets, we demonstrate that the ABLiMA model outperforms traditional models like LDA, achieving higher semantic coherence.

- We present thorough analyses showing that ABLiMA surpasses existing models in effectively capturing the thematic focus of authors, particularly in cases with significant topic variability and sparsity.

The structure of the chapter is as follows: Section 2 outlines the ABLiMA model, covering its generative process and mathematical formulation. Section 3 presents the experimental results obtained from various datasets.

3.2 Proposed Model

In this section, we present the proposed Author Beta-Liouville Multinomial Allocation (ABLiMA) model, describing its generative process, parameter inference, and hyperparameter optimization. In order to flexibly represent author-specific topic distributions, we first define the generative process of ABLiMA, which uses the Beta-Liouville distribution. This is followed by a breakdown of the Gibbs sampling method for parameter inference, which makes it feasible to estimate latent variables effectively. Lastly, we discuss the techniques for optimizing hyperparameters to enhance the model's performance.

3.2.1 Model Definition

The Author Beta-Liouville Multinomial Allocation ABLiMA model is an advanced author-topic model that uses the Beta-Liouville distribution for modeling author-specific topic distributions and a Dirichlet distribution for topic-word distributions.

Generative Process

The generative process of the ABLiMA model involves the following steps:

- Author-Level Topic Proportions: For each author $a \in \{1, \ldots, A\}$, we draw the author-level topic proportions from a Beta-Liouville distribution parameterized by vectors $\vec{\alpha}$ and $\vec{\beta}$. This models the variability and sparsity in author-specific thematic focus.

  $$\theta_a \sim \text{Beta-Liouville}(\vec{\alpha}, \vec{\beta})$$

  Here, θ_a is a vector representing the proportion of different topics for author a. The Beta-Liouville distribution provides greater flexibility than the standard Dirichlet distribution by allowing more diverse topic proportion patterns.

- Topic-Word Distribution: For each topic $k \in \{1, \ldots, K\}$, draw a topic-word distribution ϕ_k from a Dirichlet distribution parameterized by β. This distribution ensures that each topic is associated with a distinct distribution over words.

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

Here, ϕ_k represents the probability distribution over words for topic k.

- Document-Level Topic Assignment and Word Generation For each document $d \in \{1, \ldots, D\}$ authored by an author a, and for each word position $n \in \{1, \ldots, N_d\}$:

  - A topic $z_{d,n}$ is drawn for the n-th word from the author's topic distribution θ_a:

    $$z_{d,n} \sim \text{Multinomial}(\theta_a)$$

    This step assigns a topic to each word in a document based on the thematic focus of the document's author.

  - The word $w_{d,n}$ is drawn from the topic-word distribution $\phi_{z_{d,n}}$:

    $$w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$$

    This step generates the word based on the topic assigned in the previous step.

We have outlined the generative process of ABLiMA in the algorithm provided below:

3.2.2 Parameter Inference

To estimate the hidden parameters of the Author Beta-Liouville Multinomial Allocation (ABLiMA) model, we utilize a Gibbs Sampling approach Griffiths and Steyvers (2004), which is a Markov Chain Monte Carlo (MCMC) method that allows efficient inference of the posterior distributions for complex probabilistic models. The latent parameters that need to be inferred in ABLiMA include the author-level topic proportions (θ_a), the topic-word distributions (ϕ_k), and the topic assignments for each word in each document ($z_{d,n}$). Below, we describe how each of these components is inferred iteratively.

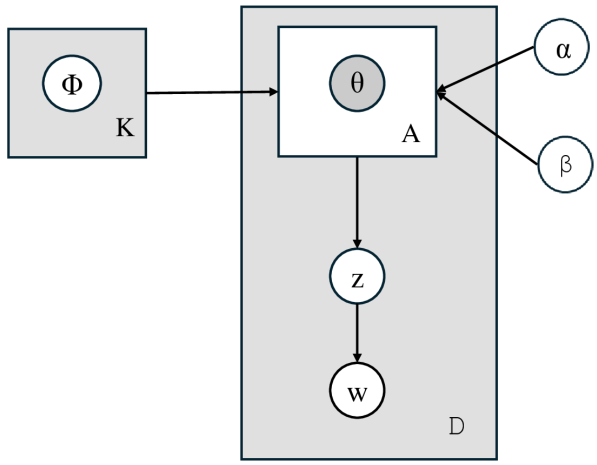

Figure 3.1: Graphical Model of ABLiMA.

Table 3.1: Summary of Mathematical Notations

Notation	**Meaning**
ϕ_k	The word distribution for topic k.
a, b	Parameters of the Beta-Liouville distribution for the word distribution within topic k.
θ_a	The topic distribution for author a.
$\vec{\alpha}, \vec{\beta}$	Hyperparameters for the Beta-Liouville distribution for author-level topic proportions.
$z_{d,n}$	The topic assigned to the n-th word in document d.
$w_{d,n}$	The n-th word in document d.
A	The number of authors in the dataset.
k	The number of topics in the model.
d	The number of documents in the dataset.
N_d	The number of words in document d.

Algorithm 1 Generative Process of the ABLiMA Model

Step 1: Draw Author-Level Topic Proportions
for each author $a \in \{1, \ldots, A\}$ **do**
  Draw author-level topic proportions $\theta_a \sim$ Beta-Liouville$(\vec{\alpha}, \vec{\beta})$
end for
Step 2: Draw Topic-Word Distributions
for each topic $k \in \{1, \ldots, K\}$ **do**
  Draw topic-word distribution $\phi_k \sim$ Dirichlet(β)
end for
Step 3: Generate Words for Documents
for each document $d \in \{1, \ldots, D\}$ authored by author a **do**
  for each word position $n \in \{1, \ldots, N_d\}$ **do**
    Draw topic $z_{d,n} \sim$ Multinomial(θ_a)
    Draw word $w_{d,n} \sim$ Multinomial$(\phi_{z_{d,n}})$
  end for
end for
Output: Generated words for each document.

The Beta-Liouville distribution, defined over a K-dimensional simplex, is characterized by the parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$, subject to the constraint $\sum_{k=1}^{K} \theta_k = 1$. It is complemented by the hyperparameter vector $\boldsymbol{\delta} = (\alpha_1, \alpha_2, \ldots, \alpha_K, \alpha, \gamma)$, providing precise control over the distribution's shape and scale.

The probability density function is given by Fan and Bouguila (2013):

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

where $\Gamma(\cdot)$ represents the Gamma function.

Here is the joint probability density function for ABLiMA:

$$
\begin{aligned}
p(\theta_a, \phi_k, Z, W \mid \vec{\alpha}, \vec{\beta}, a, b) = &\prod_{a=1}^{A} p(\theta_a \mid \vec{\alpha}, \vec{\beta}) \prod_{k=1}^{K} p(\phi_k \mid a, b) \\
&\prod_{d=1}^{D} p(Z_d \mid \theta_a) p(W_d \mid \phi_{Z_d}),
\end{aligned} \tag{2}
$$

The Gibbs Sampling function is given by:

$$
\begin{aligned}
p(z_{d,n} = k \mid z_{-d,n}, w, \vec{\alpha}, \vec{\beta}, a, b) \propto &(\theta_{a,k} + \alpha_k - 1) \\
&\cdot (\phi_{k,w_{d,n}} + b_{w_{d,n}} - 1)
\end{aligned} \tag{3}
$$

To optimize the hyperparameters, we use a Monte Carlo Expectation-Maximization (MCEM) approach. The goal of MCEM is to iteratively refine the hyperparameters in such a way that they maximize the likelihood of the observed data. The MCEM process consists of two main steps: the E-step (Expectation) and the M-step (Maximization). In the E-step, we use Gibbs Sampling to approximate the latent variables. For each word in a document, we draw topic assignments based on the conditional distributions. These topic assignments provide estimates for the hidden topic structure in the corpus. By repeating the Gibbs Sampling procedure for a sufficiently large number of iterations, we approximate the expected value of the latent variables given the current set of hyperparameters. In the M-step, we maximize the expected complete-data likelihood of the training documents with respect to the hyperparameters. Specifically, we find the values of the hyperparameters ($\vec{\alpha}$, $\vec{\beta}$, a, and b) that maximize the joint likelihood of the data and the topic assignments. To optimize the hyperparameters of the Beta-Liouville distribution, we follow a likelihood maximization approach. Specifically, for the author-level topic distribution hyperparameters $\vec{\alpha}$ and $\vec{\beta}$, and the word distribution hyperparameters a and b, we maximize the likelihood of the observed word distributions within each topic.

The objective in the M-step is to maximize the complete-data likelihood:

$$
p(w, z \mid \vec{\alpha}, \vec{\beta}, a, b) = p(w \mid z, a, b)\, p(z \mid \vec{\alpha}, \vec{\beta})
$$

where:

- $p(w \mid z, a, b)$ represents the probability of words given the topic assignments.
- $p(z \mid \vec{\alpha}, \vec{\beta})$ represents the probability of the topic assignments given the author-level topic proportions.

To optimize the hyperparameters, we solve the following optimization problem for $\vec{\alpha}$, $\vec{\beta}$, a, and b:

$$(\vec{\alpha}^*, \vec{\beta}^*, a^*, b^*) = \arg \max_{\vec{\alpha}, \vec{\beta}, a, b} E_{z \sim p(z \mid w, \vec{\alpha}, \vec{\beta}, a, b)} \left[\log p(w, z \mid \vec{\alpha}, \vec{\beta}, a, b) \right]$$

where E represents the expectation over the latent variables z drawn from the conditional distribution $p(z \mid w, \vec{\alpha}, \vec{\beta}, a, b)$.

Algorithm 2 Monte Carlo EM for ABLiMA Hyperparameter Optimization

Require: Training corpus, initial hyperparameters $\vec{\alpha}$, $\vec{\beta}$, and topic assignments Z
Ensure: Optimized hyperparameters $\vec{\alpha}^*$, $\vec{\beta}^*$

```
Initialization: Set initial values for α⃗, β⃗, and topic assignments Z
repeat convergence of α⃗, β⃗
    E-Step: Gibbs Sampling
    Perform Gibbs sampling to update the topic assignments Z
    M-Step: Hyperparameter Maximization
    Maximize the likelihood p(W, Z | α⃗, β⃗) with respect to α⃗ and β⃗
    Update α⃗ and β⃗ based on the expected topic assignments Z
until convergence
Return optimized hyperparameters α⃗*, β⃗*
```

The specific form of the expectation in the E-step is:

$$E_z \left[\sum_{k=1}^{K} \sum_{w=1}^{V} C_{k,w} \log \phi_{k,w} + \sum_{a=1}^{A} \sum_{k=1}^{K} C_{a,k} \log \theta_{a,k} \right],$$

where the counts $C_{k,w}$ and $C_{a,k}$ are approximated using Gibbs Sampling. These terms represent the expected contribution of the current topic and author assignments to the overall likelihood of the observed data, given the current hyperparameters.

3.3 Experimental Results

In this section, we present the results of our proposed Author Beta-Liouville Multinomial Allocation (ABLiMA) model on benchmark datasets, including the 20 Newsgroups and NIPS datasets.

3.3.1 Datasets

- 20 Newsgroups Dataset: This dataset contains documents from 20 different newsgroups, representing a wide variety of topics. It is commonly used for evaluating the performance of topic modeling techniques.

- NIPS Conference Papers Dataset: This dataset includes papers from the Neural Information Processing Systems (NIPS) conference, covering a diverse range of topics in machine learning. It is suited to evaluate how a topic modeling approach can capture author-specific topics.

Table 3.2 shows the word probabilities for selected topics, where the most probable words are displayed for six representative topics. The probability of each word indicates its significance within a particular topic, helping to understand the semantic focus of each topic. For instance, ”Topic 6” is centered around religion-related terms, while ”Topic 7” represents sports, evidenced by terms like ”Game” and ”Team”.

Table 3.3 illustrates the author-topic distributions, showing each author's association with a set of topics that represent the subjects they most frequently address. For example, Irwin Arnstein is primarily associated with topics 3, 15, and 2, suggesting a diverse thematic focus across different subject areas. This table illustrates the connection between authors and the dominant themes in their writing.

The following tables present the results of the topic analysis conducted on the NIPS dataset. Table 3.4 provides word probabilities for different topics, indicating the most representative words for each topic. For instance, Topic 2 primarily relates to nodes, graphs, and groups, suggesting a focus on network structures. Topic 3 contains terms like ”layer” and ”deep,” indicating a focus on deep learning and neural network architecture.

Table 3.5 shows the topic distributions for various authors in the NIPS dataset. For example, Xiangyu Wang is most associated with topics 3, 4, and 6, reflecting a combination of interests that

Table 3.2: ABLiMA-Word Probabilities per Topic on 20 Newsgroups Dataset.

TOPIC 6		TOPIC 7		TOPIC 8	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
God	0.0167	Game	0.0181	Gun	0.0118
Christian	0.0111	Team	0.0152	People	0.0096
Jesus	0.0086	Play	0.0116	Right	0.0093
Bible	0.0080	Player	0.0105	Law	0.0090
Believe	0.0066	Year	0.0105	State	0.0085
Christ	0.0064	Win	0.0082	Government	0.0076
Church	0.0063	Season	0.0080	Weapon	0.0071
Life	0.0055	League	0.0072	Kill	0.0063
People	0.0055	Score	0.0062	Crime	0.0061
Word	0.0052	Fan	0.0060	Case	0.0056
TOPIC 10		**TOPIC 12**		**TOPIC 15**	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
Space	0.0164	Work	0.0102	People	0.0090
Launch	0.0077	Power	0.0094	Israel	0.0075
Earth	0.0073	Good	0.0069	War	0.0063
NASA	0.0071	Signal	0.0067	Israeli	0.0063
Year	0.0068	Design	0.0063	State	0.0062
Orbit	0.0066	Wire	0.0062	Government	0.0061
Data	0.0059	Current	0.0061	Jew	0.0059
Program	0.0055	Radio	0.0061	Attack	0.0053
Project	0.0055	Device	0.0061	Kill	0.0052
Large	0.0054	Low	0.0060	Right	0.0050

could include deep learning, optimization, and related fields. These tables collectively illustrate the thematic preferences of both the topics and the authors, providing insights into their research focus areas.

Table 3.6 shows the word probabilities across several topics for the 20 Newsgroups dataset forATM (Author-Topicc model). In Topic 1, high-probability words such as News, Reuters, and Trump suggest a focus on current events, media, and political figures, with additional emphasis on financial terms like Market and Company. Topic 2 continues with political themes, with words like President, Trump, and House indicating government and public administration discussions.

Table 3.7 displays the distribution of author topics within the 20 Newsgroups dataset. It shows that many prominent news outlets, such as Atlantic, Breitbart, and Fox News, frequently cover Topics 1, 4, and 18, indicating shared themes or areas of focus among these sources. Other publications

Table 3.3: ABLiMA-Author-Topic Distribution on 20 Newsgroups dataset.

Author	Topics
irwin@cmptrc.lonestar.org	3, 15, 2
david@terminus.ericsson.se	5, 8, 15
rodc@fc.hp.com	19, 18, 1
jgreen@amber	11, 19, 8
jllee@acsu.buffalo.edu	0, 1, 5
mathew	15, 8, 5
ab@nova.cc.purdue.edu	10, 1, 15
CPKJP@vm.cc.latech.edu	3, 17, 1
ritley@uimrl7.mrl.uiuc.edu	11, 19, 15
abarden@tybse1.uucp	10, 19, 8

like CNN, New York Post, and New York Times have significant coverage of Topics 2, 4, and 18, reflecting a possible emphasis on political and current events.

Table 3.8 outlines the LDA model word probabilities for several topics in the 20 Newsgroups dataset. In Topic 1, terms such as Image, File, and Jpeg suggest discussions related to digital media and file handling, with frequent references to files and images. Topic 2 features words like Gun, File, and Right, indicating a focus on rights and possibly legal or policy-related content.

3.3.2 Coherence Score

Topic coherence measures the quality of topics generated by a model, reflecting how interpretable and meaningful the topics are to human readers. It quantifies the semantic similarity between the most representative words in a topic, aiming to determine if the words typically occur together in real-world contexts. A high coherence score indicates that the generated topics consist of related words, making them easier to interpret and understand. This metric is crucial for evaluating the effectiveness of topic models, as it ensures the topics extracted are insightful and relevant to the underlying dataset Ennajari et al. (2021). It is defined by:

$$\text{Coherence} = \frac{1}{M} \sum_{i=2}^{N} \sum_{j=1}^{i-1} \log \left(\frac{D(w_i, w_j) + 1}{D(w_j)} \right)$$

Figures 3.2 and 3.3 illustrate the coherence scores of topics derived from the ABLiMA model, as the number of top words used for coherence calculation increases from 5 to 30. The first chart

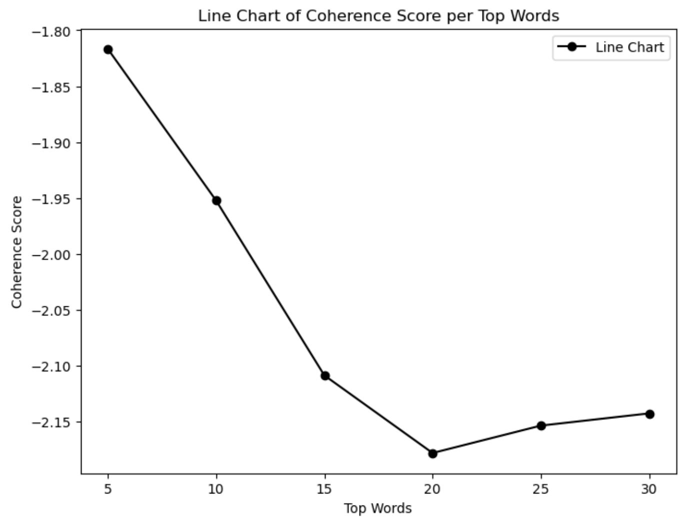

Figure 3.2: Coherence Score of 20 Newsgroups dataset.

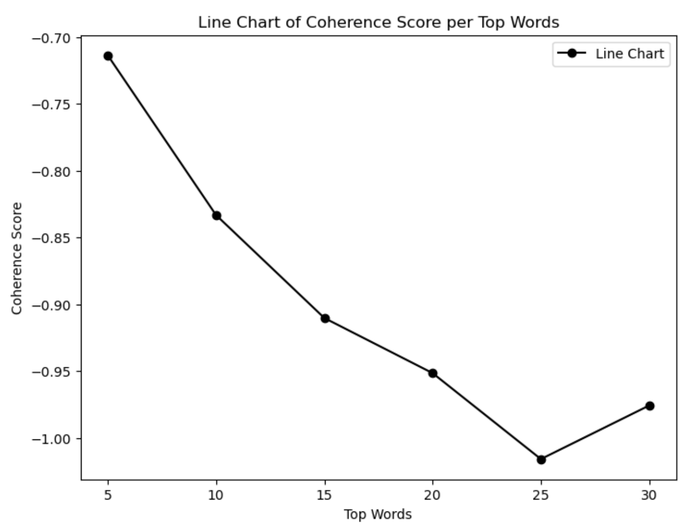

Figure 3.3: Coherence Score of NIPS dataset.

Table 3.4: ABLiMA-Word Probabilities per Topic on NIPS Dataset.

TOPIC 2		TOPIC 3		TOPIC 4	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
Node	0.0043	Layer	0.0057	Bayesian	0.0038
Binary	0.0039	Architecture	0.0055	Posterior	0.0037
Graph	0.0038	Deep	0.0054	Likelihood	0.0036
Assign	0.0038	Bengio	0.0052	Noise	0.0031
Group	0.0036	Hinton	0.0051	Inference	0.0030
Edge	0.0035	Convolutional	0.0043	Variance	0.0030
Capture	0.0033	Sutskever	0.0041	Dynamic	0.0029
Identify	0.0032	Unit	0.0039	Simulation	0.0027
Connect	0.0032	Activation	0.0035	Fit	0.0024
Partition	0.0029	Lecun	0.0034	Equation	0.0024

TOPIC 5		TOPIC 6		TOPIC 8	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
IID	0.0040	Convex	0.0076	CVPR	0.0055
Sense	0.0034	Descent	0.0062	Recognition	0.0053
Family	0.0033	Minimization	0.0057	Visual	0.0053
Finite	0.0033	Norm	0.0049	Vision	0.0048
Uniform	0.0031	Regularization	0.0045	Object	0.0042
Turn	0.0031	Dual	0.0044	Human	0.0039
Literature	0.0029	Convexity	0.0043	Pixel	0.0039
Establish	0.0029	Smooth	0.0040	Pattern	0.0038
Implies	0.0029	Regularize	0.0039	Scene	0.0037
Distance	0.0028	Program	0.0038	Image	0.0037

corresponds to the 20 Newsgroups dataset, while the second chart represents the NIPS dataset. For both datasets, we observe a general trend of decreasing coherence scores as the number of top words grows, indicating diminishing coherence between the additional words. The coherence scores of the ABLiMA model were computed following the methodology described by Mimno et al. (2011), which has been shown to effectively reflect the semantic consistency of topics.

3.3.3 Qualitative Analysis

The qualitative analysis is done by manual inspection. Chang et al. (2009) explored how well humans can interpret the output of topic models.

The heatmaps infigurese 3.4 and 3.5 show the topic distributions for authors in the two datasets: 20 Newsgroups and NIPS. Each row represents an author, while each column corresponds to a topic.

Table 3.5: ABLiMA-Author-Topic Distribution in NIPS dataset.

Author	**Topics**
Xiangyu Wang	3, 4, 6
Fangjian Guo	9, 8, 7
Lars Buesing	3, 0, 2
David Silver	0, 8, 3
Daan Wierstra	9, 8, 7
Nicolas Heess	3, 2, 0
Oriol Vinyals	2, 0, 7
Razvan Pascanu	2, 7, 3
Danilo Jimenez Rezende	3, 2, 0
Theophane Weber	9, 8, 7

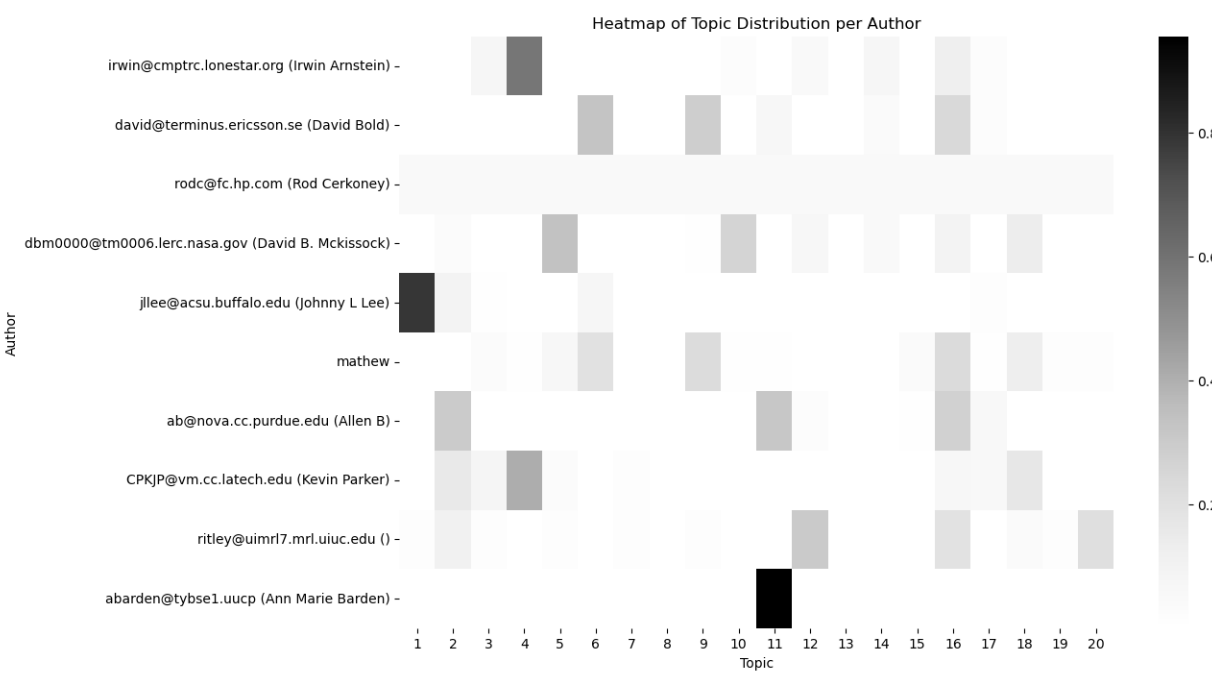

Figure 3.4: Heatmap of 20newsgroup dataset.

Table 3.6: ATM-Word Probabilities per Topic on 20 Newsgroups dataset.

TOPIC 1		TOPIC 2		TOPIC 4	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
News	0.032	President	0.010	Trump	0.0037
Reuters	0.016	Trump	0.008	State	0.0012
Trump	0.010	Year	0.007	President	0.0011
Business	0.008	New	0.007	Clinton	0.007
World	0.008	House	0.006	Campaign	0.006
Percent	0.007	State	0.006	Vote	0.006
State	0.007	Time	0.005	Republican	0.006
Market	0.007	City	0.005	Party	0.005
President	0.006	Officials	0.005	House	0.005
Company	0.006	Include	0.005	Republicans	0.005

TOPIC 9		TOPIC 12		TOPIC 15	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
Super	0.000	Archiveteam	0.000	Archiveteam	0.000
Like	0.000	Like	0.000	Company	0.000
Peak	0.000	Company	0.000	Article	0.000
New	0.000	People	0.000	Facebook	0.000
Time	0.000	New	0.000	Time	0.000
Play	0.000	Time	0.000	Future	0.000
Facebook	0.000	Write	0.000	Like	0.000
Learn	0.000	Work	0.000	New	0.000
Company	0.000	Year	0.000	Group	0.000
Story	0.000	Article	0.000	Story	0.000

The intensity of the color indicates the strength of association between the author and the respective topic. In the 20 Newsgroups dataset, we see some authors strongly aligned with particular topics, as indicated by the darker shades. Similarly, the NIPS dataset heatmap reveals varying topic preferences among the authors, showcasing some strong associations to specific topics, especially by authors such as Oriol Vinyals and Fangjian Guo. These visualizations help understand the thematic focus of different authors in both datasets.

Table 3.7: ATM-Author Topics Distribution on 20 Newsgroups dataset

Author	**Topics**
Atlantic	1, 4, 18
Breibart	1, 4, 18
Business Insider	1, 2, 4, 18
Buzzfeed News	1, 2, 4, 18
CNN	2, 4, 18
Fox News	1, 2, 4, 18
Los Angeles Times	2, 18
NPR	1, 2, 4, 18
New York Post	2, 4, 18
New York Times	2, 4, 18

Table 3.8: LDA- Word Probabilities per Topic on 20 Newsgroups Dataset.

TOPIC 1		**TOPIC 2**		**TOPIC 4**	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
Image	0.017	Gun	0.012	Need	0.009
File	0.011	File	0.011	Use	0.008
Use	0.010	Use	0.011	Gun	0.007
Bike	0.010	Make	0.008	State	0.007
Know	0.006	Know	0.008	Like	0.007
Good	0.006	Like	0.008	Dod	0.006
Like	0.005	Say	0.008	Apr	0.006
Email	0.005	Right	0.007	File	0.006
Jpeg	0.005	Dod	0.006	Say	0.006
Just	0.005	Just	0.006	Make	0.005

TOPIC 6		**TOPIC 8**		**TOPIC 9**	
WORD	**PROB.**	**WORD**	**PROB.**	**WORD**	**PROB.**
Say	0.008	Make	0.0012	Bike	0.0016
Fbi	0.008	Law	0.008	Like	0.0010
Child	0.008	Right	0.008	Just	0.008
Compound	0.007	Good	0.008	Time	0.008
Make	0.007	Time	0.007	Dog	0.007
Batf	0.006	Use	0.007	Good	0.007
Come	0.006	Like	0.006	Right	0.006
Start	0.005	Public	0.006	Make	0.006
Roby	0.005	Country	0.006	Turn	0.005
Day	0.005	Say	0.006	Know	0.005

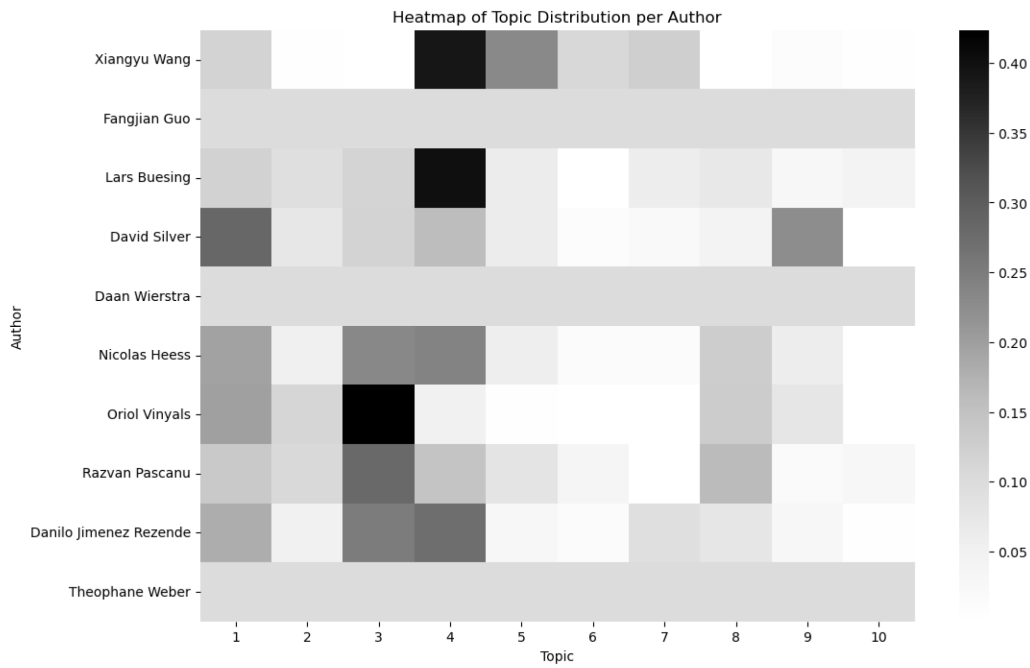

Figure 3.5: Heatmap of NIPS dataset.

Chapter 4

Conclusion

This thesis has addressed key challenges in the field of topic modeling by developing two novel probabilistic frameworks: Author Dirichlet Multinomial Allocation with Generalized Distribution (ADMAGD) and Author Beta-Liouville Multinomial Allocation (ABLiMA). These models were designed to overcome the limitations of traditional approaches, such as Latent Dirichlet Allocation (LDA) and Author-Topic Model (ATM), in handling sparsity, variability, and nuanced author-topic relationships.

The ADMAGD model incorporates the Generalized Dirichlet distribution, enabling it to capture complex dependencies between authors and topics. By leveraging this flexible distribution, ADMAGD enhances topic coherence and interpretability, making it particularly effective for datasets with intricate thematic relationships. The ABLiMA model, on the other hand, utilizes the Beta-Liouville distribution to address sparsity and variability in topic distributions. Its ability to represent absent or weakly represented topics makes it suitable for datasets with uneven thematic coverage, such as social media or short-form content.

Extensive experiments on benchmark datasets, including 20 Newsgroups and NIPS, demonstrated the superior performance of these models compared to traditional frameworks. Both ADMAGD and ABLiMA showed significant improvements in generating coherent topics, capturing nuanced thematic preferences, and managing sparsity. Visualizations of author-topic relationships further highlighted their interpretability and applicability to real-world scenarios, such as social media analysis, authorship attribution, and content recommendation systems.

The contributions of this thesis extend beyond the development of these models. By integrating flexible probabilistic distributions with author-specific modeling, this work lays a foundation for further research in flexible and robust topic modeling. Future directions include exploring hybrid approaches that combine the strengths of ADMAGD and ABLiMA, improving scalability for large datasets, and extending the models to multilingual and dynamic content analysis.

In conclusion, this thesis represents a significant step forward in advancing author-specific topic modeling, providing tools that are not only effective and interpretable but also adaptable to the complexities of modern textual datasets. The findings underscore the potential of integrating innovative probabilistic frameworks into topic modeling, paving the way for new applications and methodologies in the field of natural language processing.

References

Amirkhani, M., Manouchehri, N., & Bouguila, N. (2021). Birth-death MCMC approach for multivariate beta mixture models in medical applications. In H. Fujita, A. Selamat, J. C. Lin, & M. Ali (Eds.), *Advances and trends in artificial intelligence. artificial intelligence practices - 34th international conference on industrial, engineering and other applications of applied intelligent systems, IEA/AIE 2021, kuala lumpur, malaysia, july 26-29, 2021, proceedings, part I* (Vol. 12798, pp. 285–296). Springer.

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. doi: 10.1007/978-3-642-55032-4_29

Bakhtiari, A. S., & Bouguila, N. (2014b). A variational bayes model for count data learning and classification. *Engineering Applications of Artificial Intelligence*, *35*, 176–186. doi: 10.1016/j.engappai.2014.06.002

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 Systems with Applications*, *41*(4), 1218–1235. doi: 10.1016/j.eswa.2013.08.036

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python*. O'Reilly Media, Inc.

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. doi: 10.1145/2133806.2133826

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 doi: 10.1214/07-AOAS114

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

Bouguila, N. (2007). Spatial color image databases summarization. In *2007 ieee international conference on acoustics, speech and signal processing - icassp'07* (Vol. 1, p. I-953). IEEE. doi: 10.1109/ICASSP.2007.366116

Bouguila, N. (2012). 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 Systems with Applications*, *39*(5), 5946–5959. doi: 10.1016/j.eswa.2011.11.078

Bouguila, N., & Ziou, D. (2005a). Mml-based approach for finite dirichlet mixture estimation and selection. In P. Perner & A. Imiya (Eds.), *Machine learning and data mining in pattern recognition, 4th international conference, MLDM 2005, leipzig, germany, july 9-11, 2005, proceedings* (Vol. 3587, pp. 42–51). Springer.

Bouguila, N., & Ziou, D. (2005b). Mml-based approach for high-dimensional unsupervised learning using the generalized dirichlet mixture. In *2005 ieee computer society conference on computer vision and pattern recognition (cvpr'05)-workshops* (pp. 53–53). IEEE. doi: 10.1109/CVPR.2005.450

Bouguila, N., & Ziou, D. (2005c). On fitting finite dirichlet mixture using ECM and MML. In P. Wang, M. Singh, C. Apté, & P. Perner (Eds.), *Pattern recognition and data mining, third international conference on advances in pattern recognition, ICAPR 2005, bath, uk, august 22-25, 2005, proceedings, part I* (Vol. 3686, pp. 172–182). Springer.

Bouguila, N., & Ziou, D. (2006). Online clustering via finite mixtures of dirichlet and minimum message length. *Engineering Applications of Artificial Intelligence*, *19*(4), 371–379. doi:

10.1016/j.engappai.2005.11.004

Bouguila, N., & Ziou, D. (2007). Unsupervised learning of a finite discrete mixture: Applications to texture modeling and image databases summarization. *Journal of Visual Communication and Image Representation*, *18*(4), 295–309. doi: 10.1016/j.jvcir.2007.03.001

Boukhers, Z., & Staab, S. (2020). Analyzing the influence of hyperparameters and regularizers of topic modeling in terms of renyi entropy. *Entropy*, *22*(4), 394. doi: 10.3390/e22040394

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in neural information processing systems (nips)* (pp. 288–296).

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*, 391–407. doi: 10.1002/(SICI)1097-4571(199009)41:6⟨391::AID-ASI1⟩3.0.CO;2-9

Elguebaly, T., & Bouguila, N. (2010). Bayesian learning of generalized gaussian mixture models on biomedical images. In *Artificial neural networks in pattern recognition: 4th iapr tc3 workshop, annpr 2010, cairo, egypt, april 11-13, 2010, proceedings* (pp. 207–218). Springer Berlin Heidelberg. doi: 10.1007/978-3-642-12159-3_19

Ennajari, H., Bouguila, N., & Bentahar, J. (2021). Combining knowledge graph and word embeddings for spherical topic modeling. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(7), 3609–3623. doi: 10.1109/TNNLS.2021.3080285

Epaillard, E., & Bouguila, N. (2018). 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. doi: 10.1109/TNNLS.2018.2850724

Fan, W., & Bouguila, N. (2012). Online variational finite dirichlet mixture model and its applications. In *2012 11th international conference on information science, signal processing and their applications (isspa)* (pp. 448–453). IEEE. doi: 10.1109/ISSPA.2012.6310407

Fan, W., & Bouguila, N. (2013). Learning finite beta-liouville mixture models via variational bayes for proportional data clustering. In F. Rossi (Ed.), *Proceedings of the 23rd international joint conference on artificial intelligence (ijcai)* (pp. 1323–1329). Beijing, China: IJCAI/AAAI.

Fan, W., & Bouguila, N. (2015). Expectation propagation learning of a dirichlet process mixture of beta-liouville distributions for proportional data clustering. *Engineering Applications of Artificial Intelligence*, *43*, 1–14. doi: 10.1016/j.engappai.2015.03.007

Fan, W., Sallay, H., & Bouguila, N. (2016). Online learning of hierarchical pitman–yor process mixture of generalized dirichlet distributions with feature selection. *IEEE Transactions on Neural Networks and Learning Systems*, *28*(9), 2048–2061. doi: 10.1109/TNNLS.2016.2587822

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Suppl 1), 5228–5235. doi: 10.1073/pnas.0307752101

Ihou, K. E., & Bouguila, N. (2017). A new latent generalized dirichlet allocation model for image classification. In *2017 seventh international conference on image processing theory, tools and applications (ipta)* (pp. 1–6). IEEE. doi: 10.1109/IPTA.2017.8310132

Ihou, K. E., & Bouguila, N. (2019). Variational-based latent generalized dirichlet allocation model in the collapsed space and applications. *Neurocomputing*, *332*, 372–395.

Kaggle. (n.d.). *Nips papers dataset.* (Retrieved from https://www.kaggle.com/benhamner/nips-papers)

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *Proceedings of the 12th international conference on machine learning* (pp. 331–339).

Luo, Z., Amayri, M., Fan, W., Ihou, K. E., & Bouguila, N. (2024). Parallel inference for cross-collection latent generalized dirichlet allocation model and applications. *Expert Syst. Appl.*, *238*(Part A), 121720.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the conference on empirical methods in natural language processing* (pp. 262–272).

Najar, F., & Bouguila, N. (2022). Smoothed generalized dirichlet: A novel count-data model for detecting emotional states. *IEEE Transactions on Artificial Intelligence*, *3*(5), 685–698. doi: 10.1109/TAI.2021.3079354

Nguyen, H., Azam, M., & Bouguila, N. (2019). Data clustering using variational learning of finite scaled dirichlet mixture models. In *2019 ieee 28th international symposium on industrial electronics (isie)* (p. 1391-1396).

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on uncertainty in artificial intelligence (uai)* (pp. 487–494). AUAI Press.

Tahsin, F., Ennajari, H., & Bouguila, N. (2024). Author dirichlet multinomial allocation model with generalized distribution (ADMAGD). In *International symposium on networks, computers and communications, ISNCC 2024, washington, dc, usa, october 22-25, 2024* (pp. 1–7). IEEE.

Tahsin, F., Ennajari, H., & Bouguila, N. (2025). Author beta-liouville multinomial allocation model (ablima). In *Proceedings of the 27th international conference on enterprise information systems (iceis 2025).* (To appear)

Tang, J., & Chen, Q. (2019). Zero-inflated latent dirichlet allocation model for microbiome studies. *Frontiers in Microbiology*, *10*, 387. doi: 10.3389/fmicb.2019.00387

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. doi: 10.1109/TNNLS.2021.3059387

Zamzami, N., Alsuroji, R., Eromonsele, O., & Bouguila, N. (2020). Proportional data modeling via selection and estimation of a finite mixture of scaled dirichlet distributions. *Comput. Intell.*, *36*(2), 459–485.