Author Beta-Liouville Multinomial Allocation Model

Faiza Tahsin[a], Hafsa Ennajari[b] and Nizar Bouguila[c]
Concordia Institute for Information Systems Engineering (CIISE), Concordia University, Montreal, Canada
-

Keywords: Beta-Liouville, Bayesian Inference, Topic Modeling, Sparsity, Hierarchical Modeling.

Abstract: Conventional topic models usually presume that topics are evenly distributed among documents. Sometimes, this presumption may not be true for many real-world datasets characterized by sparse topic representation. In this paper, we present the Author Beta-Liouville Multinomial Allocation Model (ABLiMA), an innovative approach to topic modeling that incorporates the Beta-Liouville distribution to better capture the variability and sparsity of topic presence across documents. In addition to the prior flexibility our model also leverages the authorship information, leading to more coherent topic diversity.ABLiMA can represent topics that may be entirely absent or only partially present in specific documents, offering enhanced flexibility and a more realistic depiction of topic proportions in sparse datasets. Experimental results on the 20 Newsgroups and NIPS datasets demonstrate superior performance of ABLiMA compared to conventional models, suggesting its ability to model complex topics in various textual corpora. This model is particularly advantageous for analyzing text with uneven topic distributions, such as social media or short-form content, where conventional assumptions often fall short.

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 (Bouguila, 2009). In this context, we propose the Author Beta-Liouville Multinomial Allocation (ABLiMA) model, which integrates the Beta-Liouville distribution (Epaillard and Bouguila, 2016; Ali and Bouguila, 2019; Zamzami and Bouguila, 2020) 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 paper are as follows:

- We introduce the ABLiMA model, a novel approach to author-topic modeling that integrates the Beta-Liouville distribution, enabling more

[a] https://orcid.org/0009-0009-6156-1278
[b] https://orcid.org/0000-0001-8725-2638
[c] https://orcid.org/0000-0001-7224-7940

Tahsin, F., Ennajari, H. and Bouguila, N.
Author Beta-Liouville Multinomial Allocation Model.
DOI: 10.5220/0013288000003929

In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 1*, pages 251-258
ISBN: 978-989-758-749-8; ISSN: 2184-4992

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 paper is as follows: Section 2 provides an overview of the relevant literature on topic modeling and the Beta-Liouville distribution. Section 3 outlines the ABLiMA model, covering its generative process and mathematical formulation. Section 4 presents the experimental results obtained from various datasets, and Section 5 concludes with a discussion of findings and future research opportunities.

2 RELATED WORKS

In recent years, topic modeling has been receiving considerable attention, particularly due to the growth of probabilistic models such as Latent Dirichlet Allocation (LDA) (Blei et al., 2003). Documents are assumed to be mixtures of topics, and topics are assumed to be mixtures of words. Consequently, LDA has been frequently used for understanding latent thematic structures in text corpora. Although LDA has demonstrated usefulness in numerous applications, it encounters hardship in capturing sparsity and variations in thematic relevance across documents, especially with datasets with short or noisy texts, such as user-generated content and social media posts. (Blei and Lafferty, 2007) introduced correlated topic models to accommodate inter-topic dependencies; however, sparsity continued to be an obstacle. (Rosen-Zvi et al., 2004) introduced the Author-Topic Model (ATM), which builds upon LDA. This model integrates authorship information into the generative process, enabling it to identify topics based on both the authors of the documents and the text they contain. ATM presumes that an author is associated with a distribution of topics, and this distribution influences the documents they write. Sparse data and the varying importance of topics across various documents and authors were also challenges that ATM encountered, despite its advancements.

Several breakthroughs have been made by incorporating more flexible distributions to resolve these limitations. (Bouguila, 2012) introduced infinite Liouville mixture models to enhance text and texture categorization. The Beta-Liouville distribution has been implemented in numerous domains, such as high-dimensional data modeling and text clustering (Fan and Bouguila, 2013a). The Beta-Liouville distribution has demonstrated potential in handling sparsity and skewness in datasets, which are frequent challenges in real-world data, such as text corpora. (Fan and Bouguila, 2013b) Also proposed an approach for online learning using a Dirichlet process mixture of Beta-Liouville distributions.

(Fan and Bouguila, 2015; Luo et al., 2023) illustrated the Beta-Liouville distribution's efficiency in the context of document clustering and proportional data modeling when dealing with scarce and skewed data. This distribution is an appropriate choice for advanced topic modeling frameworks due to its ability to model intricate relationships among latent variables. (Bakhtiari and Bouguila, 2014) also introduced an online learning variant of topic models that utilizes Beta-Liouville priors, which allows for real-time changes to topic distributions. This online approach is appropriate to the requirements of contemporary dynamic datasets, including social media feeds and news articles, with thematic relevance that fluctuates over time. (Bakhtiari and Bouguila, 2016) introduced the Latent Beta-Liouville Allocation Model, which extends conventional topic modeling frameworks by incorporating Beta-Liouville priors to capture latent structure in count data. This model was recently proposed. This model demonstrated substantial enhancements in terms of interpretability and accuracy in high-dimensional and text datasets.

The ABLiMA model enhances these developments by incorporating Beta-Liouville priors into the author-topic modeling framework. In doing so, ABLiMA enhances previous models by addressing the challenge of sparsity and varying thematic relevance in author-specific documents. In summary, ABLiMA is a product of both classical models, such as LDA, and contemporary developments in the application of flexible priors, such as the Beta-Liouville distribution. With a combination of these ideas, the ABLiMA gives a far superior and more versatile approach to author-topic modeling that is capable of handling the present-day textual data set.

3 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.

3.1 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.

3.1.1 Generative Process

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

- Author-Level Topic Proportions: For each author $a \in \{1,\dots,A\}$, 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 $k \in \{1,\dots,K\}$, 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 $d \in \{1,\dots,D\}$ authored by an author a, and for each word position $n \in \{1,\dots,N_d\}$:

  – 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:

Algorithm 1: Generative Process of the ABLiMA Model.

for *each author* $a \in \{1,\dots,A\}$ **do**
  Draw author-level topic proportions
  $\theta_a \sim \text{Beta-Liouville}(\vec{\alpha},\vec{\beta})$;
end
for *each topic* $k \in \{1,\dots,K\}$ **do**
  Draw topic-word distribution
  $\phi_k \sim \text{Dirichlet}(\beta)$;
end
for *each document* $d \in \{1,\dots,D\}$ *authored by author* a **do**
  for *each word position* $n \in \{1,\dots,N_d\}$ **do**
    Draw topic $z_{d,n} \sim \text{Multinomial}(\theta_a)$;
    Draw word
    $w_{d,n} \sim \text{Multinomial}(\phi_{z_{d,n}})$;
  end
end

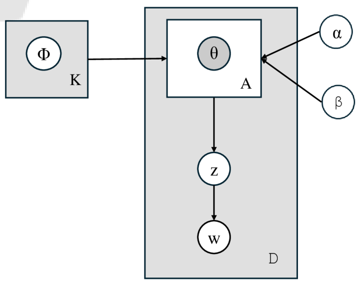

Figure 1: Graphical Model of ABLiMA.

3.2 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

Table 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.

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.

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

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

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

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

Here is the joint probability density function for ABLiMA:

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

The Gibbs Sampling function is given by:

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

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. For the Beta-Liouville author-level topic distribution hyperparameters ($\vec{\alpha}$ and $\vec{\beta}$) For the Beta-Liouville word distribution hyperparameters (a and b), we optimize them by maximizing the likelihood of the observed word distributions for 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:

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

where $\mathbb{E}$ 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.

Data: Training corpus, initial hyperparameters $\vec{\alpha}$, $\vec{\beta}$, and topic assignments Z
Result: Optimized hyperparameters $\vec{\alpha}^*$, $\vec{\beta}^*$
Initialization: Set initial values for α, β, and topic assignments Z;
repeat
  E-Step: Gibbs Sampling ;
  Perform Gibbs sampling to update the topic assignments Z;
  M-Step: Hyperparameter Maximization ;
  Maximize the likelihood $p(W, Z \mid \vec{\alpha}, \vec{\beta})$ with respect to $\vec{\alpha}$ and $\vec{\beta}$;
  Update $\vec{\alpha}$ and $\vec{\beta}$ based on the expected topic assignments Z;
until *convergence of* $\vec{\alpha}$, $\vec{\beta}$;
Return optimized hyperparameters $\vec{\alpha}^*$, $\vec{\beta}^*$

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

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

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.

4 EXPERIMENTAL RESULTS

4.1 Datasets

- 20 Newsgroups: This dataset contains documents from 20 different newsgroups, representing a wide variety of topics. It is commonly used for evaluating topic modeling techniques.
- NIPS Conference Papers: This dataset includes papers from 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 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 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 above tables present the results of the topic analysis conducted on the NIPS dataset. Table 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 fo-

Table 2: ABLiMA-Word Probabilities per Topic on 20 newsgroups dataset.

TOPIC 6		TOPIC 7	
WORD	**PROB.**	**WORD**	**PROB.**
God	0.0167	Game	0.0181
Christian	0.0111	Team	0.0152
Jesus	0.0086	Play	0.0116
Bible	0.0080	Player	0.0105
Believe	0.0066	Year	0.0105
Christ	0.0064	Win	0.0082
Church	0.0063	Season	0.0080
Life	0.0055	League	0.0072
People	0.0055	Score	0.0062
Word	0.0052	Fan	0.0060

TOPIC 10		TOPIC 12	
WORD	**PROB.**	**WORD**	**PROB.**
Space	0.0164	Work	0.0102
Launch	0.0077	Power	0.0094
Earth	0.0073	Good	0.0069
NASA	0.0071	Signal	0.0067
Year	0.0068	Design	0.0063
Orbit	0.0066	Wire	0.0062
Data	0.0059	Current	0.0061
Program	0.0055	Radio	0.0061
Project	0.0055	Device	0.0061
Large	0.0054	Low	0.0060

Table 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

Table 4: ABLiMA-Word Probabilities per Topic on NIPS.

TOPIC 2		TOPIC 3	
WORD	**PROB.**	**WORD**	**PROB.**
Node	0.0043	Layer	0.0057
Binary	0.0039	Architecture	0.0055
Graph	0.0038	Deep	0.0054
Assign	0.0038	Bengio	0.0052
Group	0.0036	Hinton	0.0051
Edge	0.0035	Convolutional	0.0043
Capture	0.0033	Sutskever	0.0041
Identify	0.0032	Unit	0.0039
Connect	0.0032	Activation	0.0035
Partition	0.0029	Lecun	0.0034

TOPIC 5		TOPIC 6	
WORD	**PROB.**	**WORD**	**PROB.**
IID	0.0040	Convex	0.0076
Sense	0.0034	Descent	0.0062
Family	0.0033	Minimization	0.0057
Finite	0.0033	Norm	0.0049
Uniform	0.0031	Regularization	0.0045
Turn	0.0031	Dual	0.0044
Literature	0.0029	Convexity	0.0043
Establish	0.0029	Smooth	0.0040
Implies	0.0029	Regularize	0.0039
Distance	0.0028	Program	0.0038

Table 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

cus on network structures. Topic 3 contains terms like ”layer” and ”deep,” indicating a focus on deep learning and neural network architecture. Table 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 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 6 shows the word probabilities across several topics for in the 20 Newsgroups for ATM (Author-Topic 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 7 displays the distribution of author

Table 6: ATM-Word Probabilities per Topic on 20 Newsgroups dataset.

TOPIC 1		TOPIC 2	
WORD	**PROB.**	**WORD**	**PROB.**
News	0.032	President	0.010
Reuters	0.016	Trump	0.008
Trump	0.010	Year	0.007
Business	0.008	New	0.007
World	0.008	House	0.006
Percent	0.007	State	0.006
State	0.007	Time	0.005
Market	0.007	City	0.005
President	0.006	Officials	0.005
Company	0.006	Include	0.005

TOPIC 4		TOPIC 9	
WORD	**PROB.**	**WORD**	**PROB.**
Trump	0.0037	Super	0.000
State	0.0012	Like	0.000
President	0.0011	Peak	0.000
Clinton	0.007	New	0.000
Campaign	0.006	Time	0.000
Vote	0.006	Play	0.000
Republican	0.006	Facebook	0.000
Party	0.005	Learn	0.000
House	0.005	Company	0.000
Republicans	0.005	Story	0.000

Table 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

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 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 8 outlines the LDA word probabilities for several topics in the 20 Newsgroups. 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

Table 8: LDA- Word Probabilities per Topic on 20 Newsgroups dataset.

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

TOPIC 4		TOPIC 6	
WORD	PROB.	WORD	PROB.
Need	0.009	Say	0.008
Use	0.008	Fbi	0.008
Gun	0.007	Child	0.008
State	0.007	Compound	0.007
Like	0.007	Make	0.007
Dod	0.006	Batf	0.006
Apr	0.006	Come	0.006
File	0.006	Start	0.005
Say	0.006	Roby	0.005
Make	0.005	Day	0.005

Gun, File, and Right, indicating a focus on rights and possibly legal or policy-related content.

4.2 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):

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

Figures 2 and 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 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., 2012), which has been shown to effectively reflect the semantic consistency of topics.

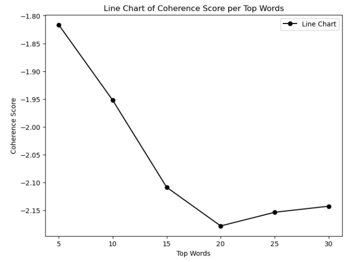

Figure 2: Coherence Score of 20 Newsgroups dataset.

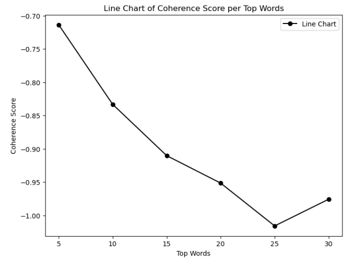

Figure 3: Coherence Score of NIPS dataset.

4.3 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

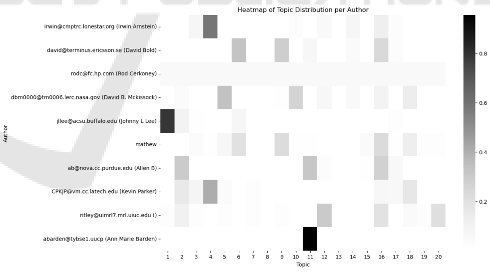

Figure 4: Coherence Score of NIPS dataset.

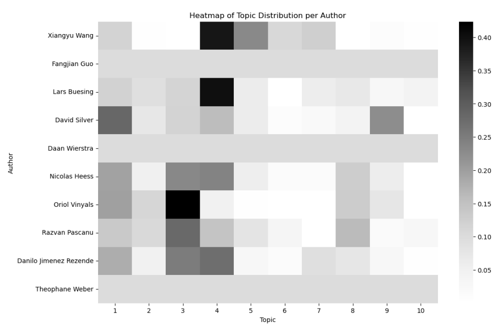

Figure 5: Coherence Score of NIPS dataset.

in figure 4 and 5 above 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. 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.

5 CONCLUSION

We proposed ABLiMA, an author-topic modeling approach, by integrating the Beta-Liouville, allowing greater flexibility in capturing the variability and sparsity of author-specific thematic focus. Through experiments, the model demonstrated its ability to extract meaningful topic distributions, reflected in coherent topic clusters and insightful author-topic relationships. Visualizations like heatmaps and coherence scores further validated the effectiveness of the model in distinguishing distinct topic preferences among authors. Future work could focus on optimizing hyperparameter estimation techniques and incorporating automatic inference of the optimal number of topics such as Dirichlet Process-based models.

REFERENCES

Ali, S. and Bouguila, N. (2019). Variational learning of beta-liouville hidden markov models for infrared action recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 898–906.

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

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

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

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

Bouguila, N. (2009). A model-based approach for discrete data clustering and feature weighting using map and stochastic complexity. *IEEE Transactions on Knowledge and Data Engineering*, 21(12):1649–1664.

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

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems (NIPS)*, pages 288–296.

Ennajari, H., Bouguila, N., and 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.

Epaillard, E. and Bouguila, N. (2016). Proportional data modeling with hidden markov models based on generalized dirichlet and beta-liouville mixtures applied to anomaly detection in public areas. *Pattern Recognit.*, 55:125–136.

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

Fan, W. and 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. and 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.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl 1):5228–5235.

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

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2012). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 262–272.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494.

Zamzami, N. and Bouguila, N. (2020). High-dimensional count data clustering based on an exponential approximation to the multinomial beta-liouville distribution. *Inf. Sci.*, 524:116–135.