Statistical Models for Short Text Clustering

SAMAR HANNACHI

A Thesis in The

Concordia Institute for Information Systems Engineering

Presented in Partial Fulfillment of the Requirements

for the Degree of

Master of Applied Science (Quality Systems Engineering) at

Concordia University

Montréal, Québec, Canada

January 2022

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **SAMAR HANNACHI**

Entitled: **Statistical Models for Short Text Clustering**

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

Master of Applied Science (Quality Systems Engineering)

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

Signed by the Final Examining Committee:

Chair and Internal Examiner
Dr. Ayda Basyouni

Internal Examiner
Dr. Arash Mohammadi

Supervisor
Dr. Nizar Bouguila

Approved by
Dr. Mohammad Mannan, Graduate Program Director

2022

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

Abstract

Statistical Models for Short Text Clustering

SAMAR HANNACHI

A notable rise in the amounts of data collected, which were made available to the public, is witnessed. This allowed the emergence of many research problems among which extracting knowledge from short texts and their different related challenges. In this thesis, we elaborate new approaches to enhance short text clustering results obtained through the use of mixture models. We deployed the collapsed Gibbs sampling algorithm previously used with the Dirichlet Multinomial mixture model on our proposed statistical models. In particular, we proposed the collapsed Gibbs sampling generalized Dirichlet Multinomial (CGSGDM) and the collapsed Gibbs sampling Beta-Liouville Multinomial (CGSBLM) mixture models to cope with the challenges that come with short texts. We demonstrate the efficiency of our proposed approaches on the Google News corpora. We compared the experimental results with related works that made use of the Dirichlet distribution as a prior. Finally, we scaled our work to use infinite mixture models namely collapsed Gibbs sampling infinite generalized Dirichlet Multinomial mixture model (CGSIGDMM) and collapsed Gibbs sampling infinite Beta-Liouville Multinomial mixture model (CGSIBLMM). We also evaluate our proposed approaches on the Tweet dataset additionally to the previously used Google News dataset. An improvement of the work is also proposed through an online clustering process demonstrating good performance on the same used datasets. A final application is presented to assess the robustness of the proposed framework in the presence of outliers.

Acknowledgments

I would like to express my deepest gratitude to my supervisor, Dr. Nizar Bouguila, who knew how to guide me and advise me during this journey leading to this important Master's degree. Per his experience, he knew how to adapt to the conditions of the pandemic and keep guiding me. He took into account these special circumstances while still expecting high-quality work from me. His patience, passion for his work and human qualities will remain with me throughout my life hoping I can pay it forward.

I would like to also thank Dr. Fatma Najar who was deeply involved with me throughout this journey of mine. She supported me mentally, guided my research work and always pushed me to exceed my limits never letting me give up. I am deeply grateful !

I am also expressing my deepest gratitude to my parents who, like for each step I take in my life, were ready to give me the necessary support to live throughout this journey successfully. Accomplishing this degree is as important to them as it is to me. Finally, I want to thank my sisters Ines and Camilia who helped me push through hard times.

Contents

List of Figures vii

List of Tables viii

1 Introduction 1

1.1 Multinomial Distribution 2

1.2 Multinomial Dirichlet Distribution 3

1.3 Multinomial Generalized Dirichlet Distribution 4

1.4 Multinomial Beta-Liouville Distribution 5

1.5 Contributions 6

1.6 Thesis Overview 7

2 Collapsed Gibbs Sampling using Discrete Mixture for Short Text Clustering 8

2.1 Background 8

2.1.1 Latent Dirichlet Allocation 8

2.1.2 Gibbs Sampling for Dirichlet Multinomial Model 9

2.2 Proposed Models 11

2.2.1 Collapsed Gibbs Sampling for Generalized Dirichlet Multinomial Mixture Model 11

2.2.2 Collapsed Gibbs Sampling Beta-Liouville Multinomial Model 15

2.3 Experimental results 17

2.3.1 Short-text Datasets and Preprocessing 17

2.3.2 Evaluation Metrics 18

2.3.3 CGSGDMM Results 19

2.3.4 CGSBLMM Results 21

3 Online Short Text Clustering using Infinite Extensions of Discrete Mixture Models 26

3.1 Mixture Models 26

3.2 Collapsed Gibbs Sampling 28

3.2.1 Existing Cluster 28

3.2.2 Choosing new cluster 33

3.3 Proposed Algorithm 34

3.4 Online Clustering Initialization 35

3.5 Experimental Results 36

3.5.1 Short-Text Datasets and Preprocessing 36

3.5.2 Evaluation Metrics 37

3.5.3 Comparison between Infinite Mixture Models 37

3.5.4 Performance of the infinite GSGDMM 39

3.5.5 Performance of the infinite GSBLMM 40

3.5.6 Convergence of the proposed Approaches 42

3.5.7 Performance with Online Clustering 42

3.5.8 Speed of the approaches 45

3.5.9 Outlier Detection 46

4 Conclusion 47

Bibliography 49

List of Figures

Figure 2.1 Performance of the Approaches on the SnippetSet 21

Figure 2.2 Performance of GSGDMM with different numbers of K on the TitleSnippetSet 21

Figure 2.3 Number of clusters found by GSGDMM for different number of iterations . 21

Figure 2.4 NMI for different values of delta for TSSet dataset 21

Figure 2.5 Comparison of the approaches on the TitleSnippet set 22

Figure 2.6 Comparison of the approaches on the SSet 22

Figure 2.7 Impact of K on the performance 23

Figure 2.8 Number of clusters found by CGSBLM on SSet and TSSet Datasets 24

Figure 2.9 C, H and NMI for different values of α for TSSet Dataset 25

Figure 3.1 Comparison between Infinite Mixture Models 38

Figure 3.2 Performance of infinite GSGDMM given δ 39

Figure 3.3 Number of clusters found by GSGDMM given δ 40

Figure 3.4 Performance of infinite GSBLMM given α_k 41

Figure 3.5 Number of clusters found given α_k 41

Figure 3.6 Number of Clusters Found over the 10 first Iterations 43

Figure 3.7 Comparision between the proposed algorithms 45

Figure 3.8 Time used in seconds with different number of iterations 45

Figure 3.9 outliers detection 46

List of Tables

Table 2.1 Performance of the Approaches on the TSSet 19

Table 3.1 Performance of the Approaches on the datasets 37

Table 3.2 Performance of the approaches with online clustering on the datasets 44

Chapter 1

Introduction

During the last years, big corporates started making use of the amounts of data they collected through the decades, which were digitized by the use of computers and the rise of the Internet. Depending on the source, the data can be in the form of texts, images, or videos addressing different issues depending on the context of application. The specific sub-field that deals with the data in the form of texts is called Natural Language Processing (NLP) as it grasps the interaction between humans and computers using the human natural language. Some common tools are usually used to address this kind of problems such as generative statistical models which are commonly used for documents classification such as latent Dirichlet allocation (LDA) [1]. The LDA generative process allows the extraction of similarities between different documents and assigning them to unobserved groups using the Dirichlet distribution [2]. This same process is used to estimate the parameters of the multinomial distributions which describe the distribution of the topics and the words of the vocabulary [1]. Indeed, short texts are used to express opinions on social media using Twitter and Facebook posts which increases the amounts of data available. Nevertheless, this type of data presents its own challenges. Due to their nature, short texts have many constraints such as sparsity which hinders an accurate modeling of the language [3]. They also present a large-volume of characteristics which increases the calculations and enhances the complexity of the problems to be addressed. For short texts challenges, authors in [4] focused on using different data representation methods such as Bag of Words (BOW) and Term Frequency-Inverse Document Frequency

(TF-IDF). Others in [5] proposed a term weighting scheme and the author in [6] proposed classification and feature weighting using MAP and stochastic complexity. In [7], the authors proposed a topic memory network. Others introduced external knowledge [8] while deep learning methods are applied as short text classifiers in [9], [10]. Another very well-known technique to deal with short texts is the use of statistical generative models [11]. Those models are generally based on the Dirichlet distribution to extract the latent topics for which short texts can be assigned. In that sense, similar texts will be grouped under the same topics and different ones will be assigned different topics. Authors in [12] investigated the problem of discrete data by applying finite mixture models. Others in [13] worked on spam categorization using support vector machines for training. Authors in [14] and [15] used the principle of MML (Minimum Message length). Another work in [16] used a generative model based on the multinomial Dirichlet mixture while [17] considered a variational Bayes learning approach to learn a topic model. In [18], authors used the leave-one-out likelihood when estimating the parameters of the statistical model. Authors in [19] used an expandable hierarchical statistical framework for modeling count data. The works in [20] and [21] proposed a new distribution replacing the commonly used Dirichlet distribution by the Scaled Dirichlet distribution for text modeling.

The ambitious work presented in this thesis and which will be detailed later requires a number of statistical tools as follow :

1.1 Multinomial Distribution

Identifying subgroups among a collection of objects represented as count vectors is equivalent to grouping objects with sufficient similarities between them. It is generally assumed that these count vectors follow a multinomial distribution [22] and that the different sub-categories or components must be represented by a probability density function. Formally, we have a set of documents $D = (d_1, \dots, d_N)$ where each document d_i has a representation of the number of times a word appears in it $\vec{X}_i = (X_{i1}, \dots, X_{iD+1})$. These words are coming from a defined vocabulary of fixed length $V = (x_1, \dots, x_V)$. We assume that the count vector $\vec{X}_i$ follows a multinomial distribution with parameter $\vec{P}$ = (P_1,...,P_D):

$$p(\vec{X}_i|\vec{P}) = \frac{(\sum_{d=1}^{D+1} X_{id})!}{X_{i1}!\dots X_{iD+1}!} \prod_{d=1}^{D+1} P_d^{X_{id}} \tag{1}$$

where $P_{D+1} = 1 - \sum_{d=1}^{D} P_d$.

It is known that this same distribution has its own limitations, especially when the data are sparse, as is the case for short texts. Previous studies have addressed this problem by introducing priors to the multinomial distribution when building the statistical model.

1.2 Multinomial Dirichlet Distribution

The Dirichlet distribution is a multivariate generalization of the Beta distribution. Commonly used in Bayesian statistics as a prior to the multinomial distribution, it has interesting properties which reduces the complexity of the calculations [23]. We set the parameter of the multinomial distribution to be estimated as $\vec{P}$ = (P_1,...,P_{D+1}) a vector with $D + 1$ components where $P_d \geq 0$ and $\sum_{d=1}^{D+1} P_d$ = 1 and α = (α_1,...,α_{D+1}) as the vector of parameters of the Dirichlet distribution where $\alpha_d \geq 0$. The Dirichlet probability density function is given as follows:

$$p(\vec{P}|\alpha) = \frac{\Gamma(\sum_{d=1}^{D+1} \alpha_d)}{\prod_{d=1}^{D+1} \Gamma(\alpha_d)} \prod_{d=1}^{D+1} P_d^{\alpha_d - 1} \tag{2}$$

where $\alpha = (\alpha_1, \dots, \alpha_{D+1})$ is the shape parameter of the Dirichlet distribution. The Multinomial Dirichlet distribution is a mixture of unigrams that adds a prior to the multinomial distribution to estimate its parameters. This adds flexibility when describing the structure of the data. The density function called Dirichlet Multinomial model can be obtained through integrating the joint probability of the vector of occurences $\vec{X}_i$ of a document i and the vector of parameters of the multinomial distribution $\vec{P}$ as in [24]:

$$\begin{aligned} p(\vec{X}_i|\alpha) &= \int p(\vec{X}_i, \vec{P}|\vec{\alpha}) d\vec{P} \\ &= \frac{\Gamma(\sum_{d=1}^{D+1} X_{id} + 1)\Gamma(\sum_{d=1}^{D+1} \alpha_d)}{\Gamma(\sum_{d=1}^{D+1} X_{id} + \sum_{d=1}^{D+1} \alpha_d)} \prod_{d=1}^{D+1} \frac{\Gamma(X_{id} + \alpha_d)}{\Gamma(\alpha_d)\Gamma(X_{id} + 1)} \end{aligned} \tag{3}$$

1.3 Multinomial Generalized Dirichlet Distribution

Some studies have suggested that the use of more general priors improves the accuracy of the clustering process and takes into account important parts of the data structure such as the correlation between data points. Introduced by Connor and Mosimann in [25], the generalized Dirichlet distribution was introduced to overcome the limitations associated with the use of the Dirichlet distribution. In the case of the Dirichlet distribution, all input points must have the same variance, add up to one, and all be negatively correlated. Allowing for more general covariance, the generalized Dirichlet distribution used as a prior to the multinomial distribution takes into account both positive and negative correlations. It also allows sampling of each proportion from the vector of probabilities that come from independent Beta distributions. This key point gives the generalized Dirichlet distribution its flexibility compared to the more restrictive Dirichlet distribution. The probability density function of the generalized Dirichlet distribution is written as follows:

$$p(\vec{P}|\vec{\alpha},\vec{\beta}) = \prod_{d=1}^{D} \frac{\Gamma(\alpha_d+\beta_d)}{\Gamma(\alpha_d)\Gamma(\beta_d)} P_d^{\alpha_d-1}(1-\sum_{l=1}^{d} P_l)^{\gamma_d} \tag{4}$$

where $\gamma_d = \beta_d - \alpha_d + \beta_{d+1}$ for $d = 1,\ldots,D-1$ and $\gamma_d = \beta_D - 1$, $\vec{\alpha} = (\alpha_1,\ldots,\alpha_D)$, $\vec{\beta} = (\beta_1,\ldots,\beta_{D+1})$ are the parameters of the generalized Dirichlet distribution.

For modeling count data, it is common to use the previously described Multinomial Dirichlet mixture model. But, such structure has its limitations like its restrictive negative covariance when describing certain types of data. Bouguila in [24] showed the efficiency of the use of the generalized Dirichlet when modeling count data due to its more general covariance structure which allows the description of different types of data. Like the Dirichlet, the generalized Dirichlet distribution is a conjugate prior to the multinomial distribution which allows the integration over the joint distribution of $\vec{X}_i$ and $\vec{P}$ to be written as in [24]:

$$\begin{aligned} P(\vec{X}_i|\vec{\alpha},\vec{\delta}) &= \int p(\vec{X}_i,\vec{P}|\vec{\alpha},\vec{\delta})d\vec{P} \\ &= \frac{\Gamma((\sum_{d=1}^{D+1} X_{id})+1)}{\prod_{d=1}^{D+1}\Gamma(X_{id}+1)} \prod_{d=1}^{D} \frac{\Gamma(\alpha_d+\delta_d)}{\Gamma(\alpha_d)\Gamma(\delta_d)} \prod_{d=1}^{D} \frac{\Gamma(\alpha_d^{'})\Gamma(\delta_d^{'})}{\Gamma(\alpha_d^{'}+\delta_d^{'})} \end{aligned} \tag{5}$$

where $\Gamma(.)$ is the Gamma function, α_d and δ_d are the parameters of the generalized Dirichlet and

$\alpha_d^{'} = \alpha_d + X_{id}$ and $\delta_d^{'} = \delta_d + X_{id+1} + \cdots + X_{iD+1}$.

1.4 Multinomial Beta-Liouville Distribution

In another work [26], Bouguila demonstrated that the liouville family of distributions can be introduced as a prior to the Multinomial when modeling count data. More specifically, the Beta-Liouville distribution was used for such reason and showed some good results.
While the generalized Dirichlet distribution shows its effectiveness in overcoming the limitations of the Dirichlet distribution, it has the disadvantage of having twice as many parameters. The Beta-Liouville distribution has a smaller number of parameters than the generalized Dirichlet distribution while remaining flexible compared to the Dirichlet distribution. Its probability density function is given as follows [27]:

$$p(\vec{P}|\theta) = \frac{\Gamma(\sum_{d=1}^{D+1} \alpha_d)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} u^{\alpha - \sum_{d=1}^{D+1} \alpha_d}(1-u)^{\beta-1} \prod_{d=1}^{D+1} \frac{P_d^{\alpha_d - 1}}{\Gamma(\alpha_d)} \tag{6}$$

where $u = \sum_{d=1}^{D} P_d$ and $\theta = (\alpha_1, \ldots, \alpha_{D+1}, \alpha, \beta)$ is the vector parameter of the Beta-Liouville distribution.

The marginal distribution for this model can be written as in [26]:

$$\begin{aligned} p(\vec{X}_i|\vec{\theta}) &= \int_{\vec{P}} p(\vec{X}_i, \vec{P}|\theta) d\vec{P} \\ &= \frac{\Gamma((\sum_{d=1}^{D+1} X_{id}) + 1)}{\prod_{d=1}^{D+1} \Gamma(X_{id}+1)} \frac{\Gamma(\sum_{d=1}^{D} \alpha_d)\Gamma(\alpha+\beta)\Gamma(\alpha^{'})\Gamma(\beta^{'}) \prod_{d=1}^{D} \Gamma(\alpha_d^{'})}{\Gamma(\sum_{d=1}^{D} \alpha_d^{'})\Gamma(\alpha^{'}+\beta^{'})\Gamma(\alpha)\Gamma(\beta) \prod_{d=1}^{D} \Gamma(\alpha_d)} \end{aligned} \tag{7}$$

where $\alpha_1, \ldots, \alpha_{D+1}$, α and β are the parameters of the Beta-Liouville distribution, where $\alpha_1^{'}, \ldots, \alpha_{D+1}^{'}$, $\alpha^{'} = \alpha + \sum_{d=1}^{D} X_{id}$ and $\beta^{'} = \beta + X_{iD+1}$ are the updated parameters.

1.5 Contributions

(1) **Short Text Clustering using Generalized Dirichlet Multinomial Mixture Model:**

We use the collapsed Gibbs Sampling algorithm on the generalized Dirichlet multinomial mixture model to overcome the limitation brought by the Dirichlet distribution when classifying textual data. Our approach proved its efficiency compared to the one that use the Dirichlet as a prior to the multinomial distribution in the mixture model. The efficacy of our approach was proved on the challenging task of short texts classification.

This work has been published in the 13th Asian Conference on Intelligent Information and Database Systems (ACIIDS 2021) [28].

(2) **Collapsed Gibbs Sampling of Beta-Liouville Multinomial for Short text clustering:**

In this work, we used in our approach the Beta-Liouville distribution as a prior in the mixture model. This enhanced the classification results on the same dataset used in the previous work. This work has been published in the 34th International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (IEA/AIE 2021) [29].

(3) **Short Text Clustering using Infinite Extensions of Discrete Mixture Models:**

We proposed another alternative scaling the mixture models to their infinite version. We introduce the collapsed Gibbs Sampling for Infinite Mixture models using the generalized Dirichlet and the Beta-Liouville as priors to the multinomial distribution.

(4) **Online Short Text Clustering :**

To further improve our work, we did something analogous to the Fast Gibbs Sampling Dirichlet Multinomial Mixture+ (FGSDMM+) algorithm by introducing our prior to the model. This improved the classification results even better while reducing the computational cost. To finalize this work, we applied it on the task of outlier detection.

The contributions (3) and (4) were combined in one paper that was submitted to the International Journal : Computational Intelligence.

1.6 Thesis Overview

- In chapter 1, we introduce in detail the background of our work. We present then the contributions that stemmed from our thesis.

- In chapter 2, we present the generalized Dirichlet and Beta-Liouville priors used in the mixture model for the multinomial distribution. We explain their integration into the collapsed Gibbs sampling algorithm to estimate the parameter of the multinomial distribution. We explain the improved results obtained and compare them to the baseline work that uses the Dirichlet prior in the estimation process.

- In chapter 3, we explain the infinite mixture models in details and introduce our approaches that use the generalized Dirichlet and the Beta-Liouville as priors. We compare our approaches namely Infinite GSGDMM and Infinite GSBLMM to a related work using the GS-DPMM approach. We improve our approaches by introducing an online clustering initialization. We end this work by applying the different approaches to outliers detection.

- In conclusion, we briefly summarize our contributions and present some potential future research works.

Chapter 2

Collapsed Gibbs Sampling using Discrete Mixture for Short Text Clustering

In this chapter, we detail two approaches that use the generalized Dirichlet and the Beta-Liouville distributions as priors to the multinomial. The use of the collapsed Gibbs sampling algorithm is considered for the clustering process of short texts. Our work is applied on the Google News dataset which proved its efficiency compared to a more restrictive prior like the Dirichlet distribution.

2.1 Background

2.1.1 Latent Dirichlet Allocation

LDA is a statistical generative model used in text modeling that generates documents according to a fixed number of latent topics [1]. Each document is represented as a distribution over the topics and each topic is represented as a multinomial distribution over the words in the vocabulary. We can generate a document by sampling a mixture of topics from which we sample words. The generation of a document starts by randomly choosing one of the distributions over topics and assigning it to a document. Then, to each word in that document, a topic is assigned randomly from the previously chosen distribution. To assign the new topic to the word, the topics present in a document are monitored and the number of times that same word was assigned a certain topic across all the other documents is counted. This process is repeated for all the words in the different documents [1]. The distribution that assigns to each word of a document a topic is generated by a multinomial distribution with parameter θ. This parameter is estimated using Dirichlet prior with parameter α. In our mixture model, we have a second Dirichlet distribution with parameter β that helps into the estimation of the counts of the number of times topics are assigned to words across all the documents in the form of a multinomial with parameter ϕ.

So, estimating the parameters θ and ϕ comes down to estimating which are the words that compose a certain topic and which are the topics that can be the most representative of a document. Those parameters θ and ϕ being intractable, can be approximated using the Gibbs sampling algorithm. It is a Markov chain Monte Carlo algorithm effectively used to estimate the posterior distribution in probabilistic models [30].

2.1.2 Gibbs Sampling for Dirichlet Multinomial Model

When working with mixture models, it is common use to rely on the Markov chain Monte Carlo algorithm called the Gibbs sampler [31]. The model that we will follow on our work is known as GSDMM (Gibbs Sampling Dirichlet Multinomial Mixture) which was designed as a model for short text clustering [32]. This model can be seen as a rectified LDA given that it assumes that each document can be assigned only one topic. A very known analogy to this model is the ”Movie Group Approach”. The documents are assimilated to students having each a list of favorite movies representing the words. At first, the students are randomly assigned to K tables. The instruction while shuffling from one table to another is to always take into consideration two factors. The first one is to always choose a table with the highest number of students. The second one is that the film interests of the people in the same table must coincide. This is repeated until the number of clusters becomes unchanged. This same process can be found under the name of ”Chinese Restaurant Process” [33].

Algorithm 1 CGSDM

```
Set $m_k$ = 0, $n_k$ = 0 and $n_k^w$ = {}
for all the Documents do
    Sample $k_d \sim Multinomial(1/\text{K})$
    Incerement the variables values by 1, $L$ and $N$ respectively
end for
for all the Iterations do
    Assign $k_d$
    Set $m_k$ = $m_k$ - 1 , $n_k$ = $n_k$ - $L$ and $n_k^w$ = $n_k^w$ - $N$
    Sample $k_d \sim P(k|k_{\neg d}, \vec{d})$
    Incerement the variables values by 1, $L$ and $N$ respectively
end for
```

Algorithm 1 shows the functioning of the collapsed GSDMM where as a first step the variables m_k: the number of documents in cluster k, n_k: the number of words in cluster k and n_k^w: the number of occurrence of word w in cluster k are initialized to zero. Then, each document will be assigned a cluster randomly while the variables previously mentioned will be incremented respectively by 1, N_d: the number of words in document d and N_d^w: the number of occurrences of word w in document d. Then, a number of iterations will be chosen to iterate the operation of recording the actual cluster of a document, decrement the parameters by the same amounts mentioned, generate new cluster to each document following the conditional probability using the generalized Dirichlet multinomial distribution, and then increment the variables again. As shown in algorithm 1, the sampling of a document d follows two major steps :

(1) Selection of an initial cluster to be assigned to a document using the multinomial distribution.

(2) Sampling of the cluster of a document d from the conditional distribution $P(k|k_{\neg d}, \vec{d})$.

$P(k|k_{\neg d}, \vec{d})$ is derived from the mentioned Dirichlet multinomial mixture model which confirms two assumptions about the movie group process analogy. The first assumption is that tables having a lot of students will get more students and the second one is that students in the same table will share the same interests as the number of iterations grows. In that sense, only a portion of the K clusters, which will gather the students having same interests, will remain full.

While using the same collapsed Gibbs Sampling algorithm, the mixture model will change according to the approach that we will be testing. Eventually, we will be having two proposed models where one details the use of generalized Dirichlet as a prior to the multinomial while the other details the use of the Beta-Liouville as a prior.

2.2 Proposed Models

2.2.1 Collapsed Gibbs Sampling for Generalized Dirichlet Multinomial Mixture Model

This subsection will introduce the use of the generalized Dirichlet distribution when estimating the parameters of the multinomial distribution using the Gibbs sampling algorithm. Indeed, the generalized Dirichlet distribution as presented in the previous chapter results from the multinomial over the latent parameter of the multinomial distribution giving the probability of selecting a cluster k_i characterized by a generalized Dirichlet distribution as :

$$P(k_i|\alpha, \delta) = \int P(k_i|c)P(c|\alpha, \delta)dc \tag{8}$$

where $P(k_i|c)$ is a multinomial distribution and $P(c|\alpha, \delta)$ is a generalized Dirichlet distribution.

As shown in the algorithm 1, the hidden cluster of a document d is estimated using the conditional probability given the parameters of the Dirichlet. Indeed, the first one which is a generalized Dirichlet having the parameters α and δ will approximate the parameter θ of the multinomial that will give the distribution of the documents. The second one which is a simple Dirichlet with parameter β will give the distribution of the topics.

It is derived from the joint probability which can be written for the document $\vec{d}$ and the cluster k as :

$$P(\vec{d}, k|\alpha, \delta, \beta) = P(\vec{d}|k, \beta)P(k|\alpha, \delta) \tag{9}$$

We have :

$$P(\vec{d}|k, \beta) = \int P(\vec{d}|k, \phi)P(\phi|\beta)d\phi \tag{10}$$

$P(\vec{d}|k,\phi)$ is a multinomial distribution given by [34] :

$$P(\vec{d}|k,\phi) = \prod_{k=1}^{K}\prod_{w=1}^{V}\phi_{k,w}^{n_k^{(w)}} \tag{11}$$

and $P(\phi|\beta)$ is a Dirichlet given by [34] :

$$P(\phi|\beta) = \frac{\Gamma(\sum_{k=1}^{K}\beta_k)}{\prod_{k=1}^{K}\Gamma(\beta_k)}\prod_{w=1}^{V}\phi_{k,w}^{\beta_k-1} \tag{12}$$

From (11) and (12) we have:

$$\begin{aligned} P(\vec{d}|k,\beta) &= \int\prod_{k=1}^{K}\prod_{w=1}^{V}\phi_{k,w}^{n_k^{(w)}}\frac{\Gamma(\sum_{k=1}^{K}\beta_k)}{\prod_{k=1}^{K}\Gamma(\beta_k)}\prod_{w=1}^{V}\phi_{k,w}^{\beta_k-1}d\phi_k \\ &= \frac{\Gamma(\sum_{k=1}^{K}\beta_k)}{\prod_{k=1}^{K}\Gamma(\beta_k)}\int\prod_{k=1}^{K}\prod_{w=1}^{V}\phi_{k,w}^{n_k^{(w)}}\phi_{k,w}^{\beta_k-1}d\phi_k \\ &= \frac{\Gamma(\sum_{k=1}^{K}\beta_k)}{\prod_{k=1}^{K}\Gamma(\beta_k)}\int\prod_{k=1}^{K}\prod_{w=1}^{V}\phi_{k,w}^{n_k^{(w)}+\beta_k-1}d\phi_k \end{aligned} \tag{13}$$

We have $\Delta(\beta)$ is the dirichlet integral of the first kind for the summation function given by :

$$\Delta(\beta) = \frac{\Gamma(\sum_{k=1}^{K}\beta_k)}{\prod_{k=1}^{K}\Gamma(\beta_k)} \tag{14}$$

We have integrating over the probability density function equals to 1 :

$$\int\frac{1}{\Delta(\beta'_k)}\prod_{k=1}^{K}\prod_{w=1}^{V}\phi_{k,w}^{n_k^{(w)}+\beta-1}d\phi_k = 1 \tag{15}$$

where $\beta'_k = \beta + n_k^{(w)}$

$$\implies \int_{\phi\in\Phi}\prod_{k=1}^{K}\prod_{w=1}^{V}\phi_{k,w}^{n_k^{(w)}+\beta-1}d\phi_k = \prod_{k=1}^{K}\Delta(\beta'_k) \tag{16}$$

$$\Longrightarrow \prod_{k=1}^{K} \Delta(\vec{n}_k + \beta) = \int \prod_{k=1}^{K} \prod_{w=1}^{V} \phi_{k,w}^{n_k^{(w)}+\beta_k-1} d\phi_k = \prod_{k=1}^{K} \frac{\Gamma(\sum_{w=1}^{V}(n_k^{(w)} + \beta))}{\prod_{w=1}^{V}\Gamma(n_k^{(w)} + \beta)} \tag{17}$$

From (14) and (17) we have :

$$P(\vec{d}|k, \beta) = \prod_{k=1}^{K} \frac{\Delta(\vec{n}_k + \beta)}{\Delta(\beta)} \tag{18}$$

Now we will follow the same procedure for $P(\vec{z}|\alpha, \delta)$ where :

$$P(k|\alpha, \delta) = \int P(k|\theta)P(\theta|\alpha, \delta)d\theta \tag{19}$$

We have $P(k|\theta)$ is a multinomial given by :

$$P(k|\theta) = \prod_{k=1}^{K} \theta_k^{m_k} \tag{20}$$

and $P(\theta|\alpha, \delta)$ is a generalized Dirichlet distribution given by [24] :

$$P(\theta|\alpha, \delta) = \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + \delta_k)}{\Gamma(\alpha_k)\Gamma(\delta_k)} \theta_k^{\alpha_k - 1}(1 - \sum_{j=1}^{l} \theta_j)^{\gamma_l} \tag{21}$$

From (20) and (21) we have :

$$\begin{aligned} P(k|\alpha, \delta) &= \int P(k|\theta)P(\theta|\alpha, \delta)d\theta \\ &= \int \prod_{k=1}^{K} \theta_k^{m_k} \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + \delta_k)}{\Gamma(\alpha_k)\Gamma(\delta_k)} \theta_k^{\alpha_k - 1}(1 - \sum_{j=1}^{l} \theta_j)^{\gamma_l} d\theta \\ &= \prod_{k=1}^{K} \frac{\Gamma(\alpha_k + \delta_k)}{\Gamma(\alpha_k)\Gamma(\delta_k)} \int \prod_{k=1}^{K} \theta_k^{\alpha_k - 1 + m_k}(1 - \sum_{j=1}^{l} \theta_j)^{\gamma_l} d\theta \end{aligned} \tag{22}$$

For the case of generalized Dirichlet we have :

$$\Delta(\alpha, \delta) = \prod_{k=1}^{K} \frac{\Gamma(\alpha_k)\Gamma(\delta_k)}{\Gamma(\alpha_k + \delta_k)} \tag{23}$$

We have :

$$\int \frac{1}{\Delta(\alpha', \delta')} \prod_{k=1}^{K} \theta_k^{\alpha-1+m_k} (1 - \sum_{j=1}^{l} \theta_j)^{\gamma_l'} d\theta = 1 \tag{24}$$

where α' = α + m_k and δ' = δ + $\sum_{l=k+1}^{K} m_l$

$$\Longrightarrow \int \prod_{k=1}^{K} \theta_k^{\alpha-1+m_k} (1 - \sum_{j=1}^{l} \theta_j)^{\gamma_l'} d\theta = \Delta(\alpha', \delta') = \prod_{k=1}^{K} \frac{\Gamma(\alpha + m_k)\Gamma(\delta + \sum_{l=k+1}^{K} m_l)}{\Gamma(\alpha + m_k + \delta + \sum_{l=k+1}^{K} m_l)} \tag{25}$$

$$\Delta(\alpha + \vec{m}, \delta + \sum_{l=k+1}^{K} m_l) = \int \prod_{k=1}^{K} \theta_k^{\alpha-1+m_k} (1 - \sum_{j=1}^{l} \theta_j)^{\gamma_l'} d\theta = \prod_{k=1}^{K} \frac{\Gamma(m_k + \alpha)\Gamma(\delta + \sum_{l=k+1}^{K} m_l)}{\Gamma(m_k + \alpha + \delta + \sum_{l=k+1}^{K} m_l)} \tag{26}$$

From (23) and (26) we have :

$$P(\vec{d}|k, \alpha, \delta) = \frac{\Delta(\vec{m} + \alpha, \delta + \sum_{l=k+1}^{K} m_l)}{\Delta(\alpha, \delta)} \tag{27}$$

The conditional probability that will give us the hidden cluster will be derived as follows :

$$\begin{aligned} P(z_d = k | k_{\neg d}, \vec{d}) &\propto \frac{P(\vec{d}, k | \alpha, \beta, \delta)}{P(\vec{d}_{\neg d}, k_{\neg d} | \alpha, \beta, \delta)} \\ &\propto \frac{P(\vec{d}|k, \beta) P(k|\alpha, \delta)}{P(\vec{d}_{\neg d}|k_{\neg d}, \beta) P(k_{\neg d}|\alpha, \delta)} \\ &\propto \frac{\frac{\Delta(\vec{m}+\alpha, \delta+\sum_{l=k+1}^{K} m_l)}{\Delta(\alpha,\delta)}}{\frac{\Delta(\vec{m}_\neg d+\alpha, \delta+\sum_{l=k+1}^{K} m_{l,\neg d})}{\Delta(\alpha,\delta)}} \frac{\frac{\Delta(\vec{n}_k+\beta)}{\Delta(\beta)}}{\frac{\Delta(\vec{n}_{k,\neg d}+\beta)}{\Delta(\beta)}} \end{aligned} \tag{28}$$

$$P(z_d = k | k_{\neg d}, \vec{d}) \propto \frac{\Delta(\vec{m} + \alpha, \delta + \sum_{l=k+1}^{K} m_l)}{\Delta(\vec{m}_{\neg d} + \alpha, \delta + \sum_{l=k+1}^{K} m_{l,\neg d})} \frac{\Delta(\vec{n}_k + \beta)}{\Delta(\vec{n}_{k,\neg d} + \beta)}, \tag{29}$$

where $\vec{n}_k$ = $\{n_k^{(w)}\}_{w=1}^{V}$

To elaborate on this conditional probability, we will rely on three major properties :

(1) The property of the Gamma function : $\frac{\Gamma(x+m)}{\Gamma(x)} = \prod_{i=1}^{m}(x+i-1)$

(2) The proposition that : $m_k = m_{k,\neg d} + 1$

(3) The assumption that each word can appear at most once in each document

This results as follows :

$$\begin{aligned}
P(z_d = k|\vec{k}_{\neg d}, \vec{d}) &\propto \frac{\Gamma(\alpha + m_k)\Gamma(\delta + \sum_{l=k+1}^{K} m_{l,\neg d} + \alpha + m_{k,\neg d})\Gamma(\delta + \sum_{l=k+1}^{K} m_l)}{\Gamma(\alpha + \delta + \sum_{l=k+1}^{K} m_l + m_k)\Gamma(m_{k,\neg d} + \alpha)\Gamma(\delta + \sum_{l=k+1}^{K} m_{l,\neg d})} \frac{\prod_{w=1}^{V}\Gamma(n_k^{(w)} + \beta)\Gamma(n_{k,\neg d} + V\beta)}{\prod_{w=1}^{V}\Gamma(n_{k,\neg d}^{(w)} + \beta)\Gamma(n_k + V\beta)} \\
&\propto \frac{\Gamma(\alpha + m_{k,\neg d} + 1)\Gamma(\delta + \sum_{l=k+1}^{K} m_{l,\neg d} + \alpha + m_{k,\neg d})\Gamma(\delta + \sum_{l=k+1}^{K} m_{l,\neg d} + K - k)}{\Gamma(\alpha + \delta + \sum_{l=k+1}^{K} m_{l,\neg d} + m_{k,\neg d} + 1 + K - k)\Gamma(m_{k,\neg d} + \alpha)\Gamma(\delta + \sum_{l=k+1}^{K} m_{l,\neg d})} \frac{\frac{\prod_{w=1}^{V}\Gamma(n_k^{(w)}+\beta)}{\prod_{w=1}^{V}\Gamma(n_{k,\neg d}^{(w)}+\beta)}}{\frac{\Gamma(n_k+V\beta)}{\Gamma(n_{k,\neg d}+V\beta)}} \\
&\propto \frac{(m_{k,\neg d} + \alpha)\prod_{i=1}^{K-k}(\delta + \sum_{l=k+1}^{K} m_{l,\neg d} + i - 1)}{\prod_{i=1}^{K-k+1}(m_{k,\neg d} + \alpha + \delta + \sum_{l=k+1}^{K} m_{k,\neg d} + i - 1)} \frac{\prod_{w=1}^{V}(n_{k,\neg d}^{(w)} + \beta)}{\prod_{i=1}^{N_d}(n_{k,\neg d} + V\beta + i - 1)}
\end{aligned} \tag{30}$$

$$P(z_d = k|\vec{k}_{\neg d}, \vec{d}) \propto \frac{(m_{k,\neg d} + \alpha)\prod_{i=1}^{K-k}(\delta + \sum_{l=k+1}^{K} m_{l,\neg d} + i - 1)}{\prod_{i=1}^{K-k+1}(m_{k,\neg d} + \alpha + \delta + \sum_{l=k+1}^{K} m_{k,\neg d} + i - 1)} \frac{\prod_{w=1}^{V}(n_{k,\neg d}^{(w)} + \beta)}{\prod_{i=1}^{N_d}(n_{k,\neg d} + V\beta + i - 1)} \tag{31}$$

where α, δ are the two parameters of the generalized Dirichlet, β is the parameter of the Dirichlet, V size of vocabulary, $m_{k,\neg d}$ the number of documents in cluster k except for the document d, $n_{k,\neg d}^{(w)}$ number of occurence of word w in the cluster k without considering the document d and $n_{k,\neg d}$ number of words in cluster k without considering the cluster of document d.

From equation (31), we can see that the parameters α and δ determine the prior probability of a student choosing a table while the parameter β regulates the factor of sharing the interests in the same table.

2.2.2 Collapsed Gibbs Sampling Beta-Liouville Multinomial Model

This subsection details the use of the Beta-Liouville distribution when estimating the parameter of the multinomial distribution that assigns a topic to each word present in a document. The different computations to estimate the different parameters will be done using the collapsed Gibbs sampling algorithm.

The collapsed Gibbs Sampling Beta-Liouville Multinomial (CGSBLM) follows the same structure described in the algorithm 1 where the assignment of each document to a cluster follows two steps. An initial cluster is first randomly assigned to each document as an initialization step. Then, each document is assigned, over a certain number of iterations, a cluster derived from the Dirichlet and Beta-Liouville distribution.

$p(k|k_{\neg d},\vec{d})$ will give us the latent cluster of a document d by introducing Dirichlet and Beta-Liouville Multinomial distribution. It will allow the estimation of the parameters of the multinomial distributions using the joint probability $p(\vec{d},k|\alpha_1,\dots,\alpha_k,\alpha,\delta,\beta)$ composed of two probabilities.

$p(\vec{d}|k,\beta)$ is the marginalisation over the product of the probability of the multinomial distribution by the conditional probability of the Dirichlet distribution over its parameter β which gives after calculations :

$$p(\vec{d}|k,\beta)=\prod_{k=1}^{K}\frac{\Delta(\vec{n}_k+\beta)}{\Delta(\beta)} \tag{32}$$

where Δ is the function used in [34] and $\vec{n_k}=\{n_k^w\}_{w=1}^V$ with n_k^w the number of occurrence of the word w in the cluster k. The contribution of this paper comes in hand with the second probability that marginalizes over the probability of the multinomial with parameter θ and the probability density function of the Beta-Liouville distribution given by :

$$\begin{aligned}
p(k|\alpha_1,\dots,\alpha_k,\alpha,\delta) &= \int p(k|\theta)p(\theta|\alpha_1,\dots,\alpha_k,\alpha,\delta)d\theta\\
&= \int\prod_{k=1}^{K}\theta_k^{m_k}\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)\Gamma(\alpha+\delta)}{\Gamma(\alpha)\Gamma(\delta)}\prod_{k=1}^{K}\frac{\theta_k^{\alpha_k-1}}{\Gamma(\alpha_k)}(\sum_{k=1}^{K}\theta_k)^{\alpha-\sum_{k=1}^{K}\alpha_k}\\
&\quad(1-\sum_{k=1}^{K}\theta_k)^{\delta-1}d\theta\\
&= \frac{\Gamma(\sum_{k=1}^{K}\alpha_k)\Gamma(\alpha+\delta)}{\Gamma(\alpha)\Gamma(\delta)}\int\prod_{k=1}^{K}\frac{\theta_k^{m_k+\alpha_k-1}}{\Gamma(\alpha_k)}(\sum_{k=1}^{K}\theta_k)^{\alpha-\sum_{k=1}^{K}\alpha_k}\\
&\quad(1-\sum_{k=1}^{K}\theta_k)^{\delta-1}d\theta
\end{aligned} \tag{33}$$

Marginalizing the probability density function of the Beta-Liouville distribution over the parameter θ with updated parameters corresponding to the remaining integral in the equation (33) will allow us to express it in function of a fraction of Gamma functions. Following the work in [26], we will have the updated parameters as follow :

$$\begin{cases}
\alpha^{'}=\alpha+\sum_{k=1}^{K-1}m_k\\
\alpha_k^{'}=\alpha+m_k\\
\delta^{'}=\delta+m_K
\end{cases}$$

The equation (33) is then equivalent to :

$$\begin{aligned}p(k|\alpha_1,\ldots,\alpha_k,\alpha,\delta) = &\frac{\Gamma(\sum_{k=1}^{K}\alpha_k)\Gamma(\alpha+\delta)}{\Gamma(\alpha)\Gamma(\delta)\prod_{k=1}^{K}\Gamma(\alpha_k)}\frac{\Gamma(\alpha+\sum_{k=1}^{K-1}m_k)\Gamma(\delta+m_K)}{\Gamma(\sum_{k=1}^{K}(\alpha_k+m_k))}\\ &\frac{\prod_{k=1}^{K}\Gamma(\alpha_k+m_k)}{\Gamma(\alpha+\sum_{k=1}^{K-1}m_k+\delta+m_K)}\end{aligned} \tag{34}$$

The conditional distribution that samples a cluster to a document from the Dirichlet and Beta-Liouville distributions will be derived as follows :

$$\begin{aligned}p(z_d=k|\vec{k}_{\neg d},\vec{d}) &\propto \frac{p(\vec{d},k|\alpha_1,\ldots,\alpha_K,\alpha,\beta,\delta)}{p(\vec{d}_{\neg d},k_{\neg d}|\alpha_1,\ldots,\alpha_K,\alpha,\beta,\delta)}\\ &\propto \frac{\prod_{i=1}^{K-1}(\alpha+\sum_{k=1}^{K-1}m_{k,\neg d}+i-1)(\delta+m_{K,\neg d})}{\prod_{i=1}^{K}(\sum_{k=1}^{K}(\alpha_k+m_{k,\neg d})+i-1)\prod_{i=1}^{N_d}(n_{k,\neg d}+V\beta+i-1)}\\ &\quad\frac{\prod_{k=1}^{K}(\alpha_k+m_{k,\neg d}))\prod_{w=1}^{V}(n_{k,\neg d}^{(w)}+\beta)}{\prod_{i=1}^{K}(\alpha+\sum_{k=1}^{K-1}m_{k,\neg d}+\delta+m_{K,\neg d}+i-1)}\end{aligned} \tag{35}$$

where α_1,…,α_K, α, δ are the parameters of the Beta-Liouville distribution, β is the parameter of the Dirichlet distribution, $m_{k,\neg d}$ is the number of documents in the cluster k without including the document d, V size of vocabulary, $n_{k,\neg d}^{(w)}$ number of occurences of word w in the cluster k without considering the document d and $n_{k,\neg d}$ number of words in cluster k without considering the cluster of document d.

2.3 Experimental results

2.3.1 Short-text Datasets and Preprocessing

The Google News dataset was extracted from the Google News website of November, 27, 2013 where the titles and snippets of 11,109 articles were collected and associated to one of the 152 clusters. This dataset was previously used in [35]. The validity of the dataset was examined manually and was divided to different sets. Our work will focus on the SnippetSet which consists of short texts containing the main information from the articles and on the TitleSnippetSet which contains both the titles and snippets of the short texts.

The data preprocessing of the texts included lowercasing all the words, removing non-latin characters and stop words, using the WordNet Lemmatizer of NLTK to apply the stemming, keeping only sentences ranging between 2 and 15 words and removing words which frequency is less than 2.

2.3.2 Evaluation Metrics

To assess the effectiveness of our contribution to cluster short texts, we used the same metrics as in [32] : Homogeneity (H), Completeness (C), Adjusted Rand Index (ARI), Normalized Mutual Information (NMI) and Adjusted Mutual Information (AMI) [36]. Homogeneity and Completeness are two metrics that give a comparison between the ground truth and the inferred information. The Homogeneity is a cluster-wise metric where it insights if each cluster contains only observations belonging to the same ground truth. Completeness is a data-wise metric where it informs whether all the data points from the same ground truth cluster were assigned to the same cluster. The Normalized Mutual Information (NMI), which gives the same result as the V-measure, is defined as the harmonic mean between the completeness and the homogeneity [37]. The Adjusted Rand Index (ARI) measures the similarity between two data clusterings [38]. The Adjusted Mutual Information (AMI) quantifies the amount of information obtained on one random variable through observing another random variable.

The theoretical definition of the main metrics that will be used to assess this work are as follows :

$$H = 1 - \frac{H1(C1|K)}{H1(C1)}$$

where

$$H1(C1|K) = -\sum_{c,k} \frac{n_{ck}}{N} log(\frac{n_{ck}}{n_k})$$

with $\frac{n_{ck}}{n_k}$ represents the ratio between the number of samples labelled c in cluster k and the total number of samples in cluster k and

$$H1(C1) = -\sum_{c=1}^{C1} \frac{\sum_{k=1}^{K} n_{ck}}{C1} log(\frac{\sum_{k=1}^{K} n_{ck}}{C1}) \tag{36}$$

Completeness (C) indicates whether all data points in the same ground truth belong to the same cluster and is written as :

$$C = 1 - \frac{H1(K|C1)}{H1(K)}$$

We also rely on the normalized mutual information (NMI) [39] which represents the harmonic mean

	Metrics	GSDMM	GSGDMM
TSSet	NMI	0.928	0.933
	H	0.911	0.925
	C	0.945	0.941
	ARI	0.789	0.832
	AMI	0.897	0.913

Table 2.1: Performance of the Approaches on the TSSet

between C and H which is defined as :

$$NMI = 2 * \frac{H * C}{H + C}$$

2.3.3 CGSGDMM Results

Comparison of Gibbs Sampling algorithms

In this subsection, we will show the performance of our approach compared to the GSDMM approach. As given in [32], we set the initial number of clusters to 500, the number of iterations to 30, α = 0.1, δ = 0.1 and $\beta = 0.1$ for the working datasets. Figure 2.1 shows that our approach gives better results than the GSDMM approach. We can see that the GSGDMM approach improved the NMI, H, ARI and AMI metrics while having the completeness quite the same for the SSet dataset. From table 2.1, we can see the dataset TSSet for which all the mentioned metrics were increased except for the completeness metric which value slightly decreased. We can also see that both of the GSDMM and GSGDMM models perform better on longer texts. This can open room to many improvements as of giving a better representation to the short texts making it longer through different techniques.

Influence of K

In this part, we assess the influence of the initial number of clusters K on the performance of the GSGDMM model. For that, we set α = 0.1, β = 0.1 and the number of iterations to 30. Figure 2.2 displays the performance of the TitleSnippetSet for different values of K. We can see that with a small number of clusters, it gets easy for the model to assign similar documents to the same cluster which gives a very high completeness. But, this same fact gives a low value of homogeneity as it gets hard for the model to separate between the different documents. As we increase the value of K, we start to reach a certain equilibrium between the value of the homogeneity and the value of the completeness. The latter will start decreasing while the former will increase obviously. As the harmonic mean between the completeness and homogeneity, we can rely on the value of NMI to give us the best number of clusters to start with. The highest value for NMI is given for a value of K equal to 400.

Influence of the number of iterations

In this subsection, we analyze the effect of the number of iterations on the number of clusters found on the two datasets. We set the number of initial clusters K to 400, α = 0.1, β = 0.1 and δ = 0.1. From Figure 2.3, we can see that the number of clusters found for both datasets drops quickly from 400 to 182 after only 5 iterations. This observation affirms the initial concept of MGP where the most popular tables will get more popular and the less popular ones will get empty quickly. This is why we see the number of clusters found dropping. We can also see that the final number of clusters found reached is a bit above the actual number of clusters of the Google News dataset. From [32], we can see that the number of clusters found by the GSDMM for the TSSet is very near the actual number of clusters for Google News reaching 161 clusters while the one for the SSet went below reaching 148. In that aspect, GSDMM may seem to be performing better but GSGDMM has a better clustering quality since its homogeneity and completeness are higher. Also, it is predicted that going further 30 iterations will improve the number of clusters found by GSGDMM.

Performance given the parameter δ

In this part, we try to find which value of delta can give us the best results. For that, we set K = 300, α = 0.1, β = 0.1 for the TSSet dataset. We set the number of iterations to 10 and do computations for different values of delta ranging from 0.01 to 0.4. The performance is tracked through the NMI metric as it gives a good idea on how well the model is performing. From Figure 2.4, we can see that the highest value for NMI is reached for δ = 0.2.

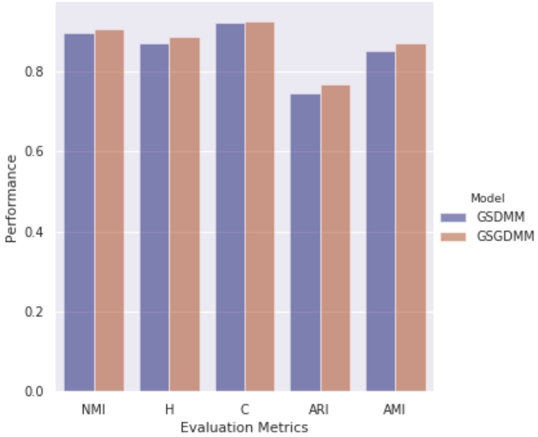


Figure 2.1: Performance of the Approaches on the SnippetSet

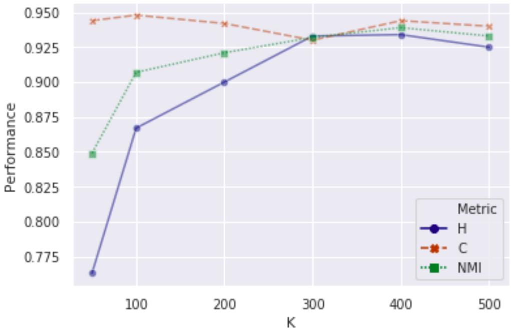


Figure 2.2: Performance of GSGDMM with different numbers of K on the TitleSnippetSet

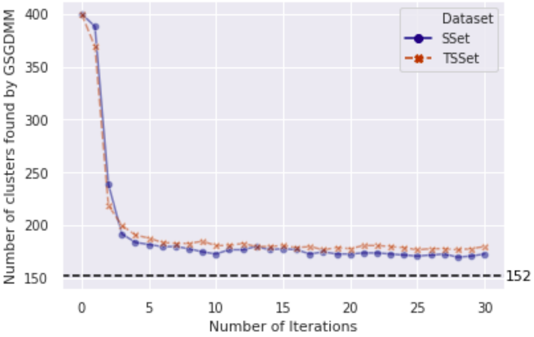


Figure 2.3: Number of clusters found by GSGDMM for different number of iterations

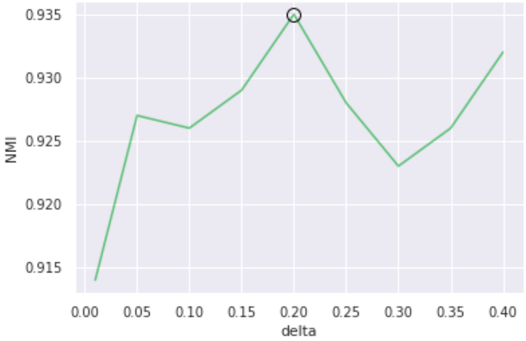


Figure 2.4: NMI for different values of delta for TSSet dataset

2.3.4 CGSBLMM Results

In this section, we prove the efficiency of our approach compared to the original approach used in [32]. We will do so by using three datasets from [32]. We run the different tests 20 times over each dataset to get the averaged results presented.

Approaches Comparison

In this subsection, we will compare the results of the GSDMM approach with our approach. For both approaches, we set the initial number of clusters K to 500, the number of iterations to

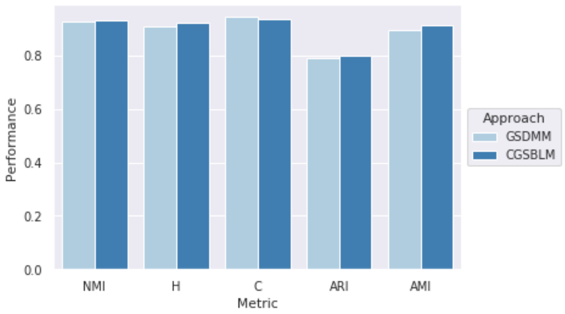


Figure 2.5: Comparison of the approaches on the TitleSnippet set

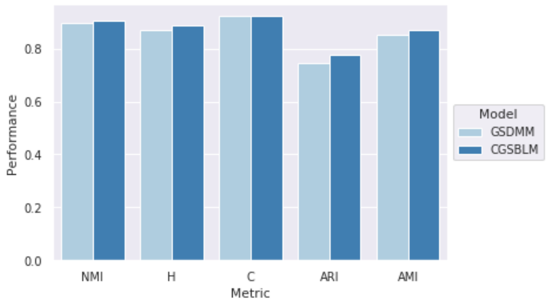
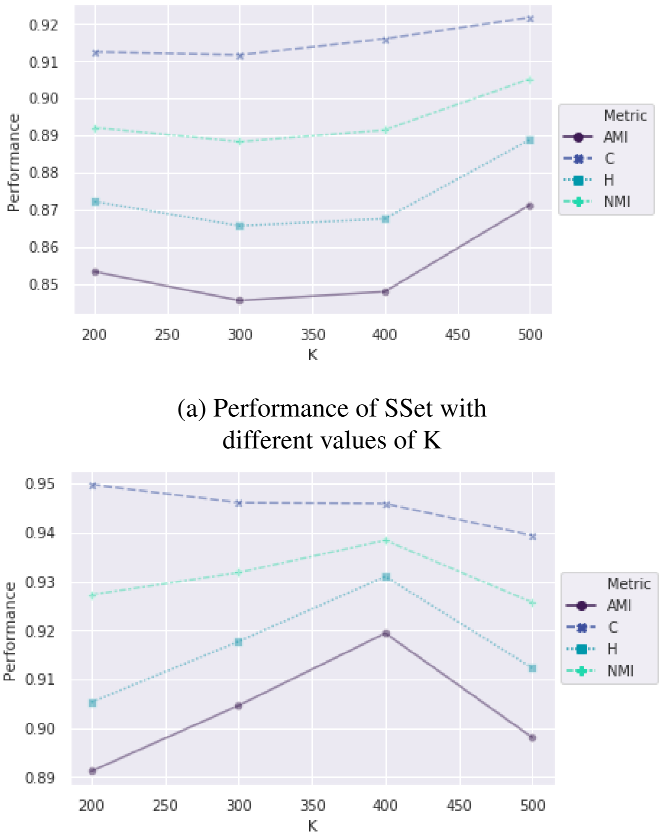


Figure 2.6: Comparison of the approaches on the SSet

30 and the value of the parameters to 0.1. Figure 2.5 shows that our approach improved all the used metrics except for the Completeness where the GSDMM model performed slightly better on the TitleSnippet set. For the Snippet set, Figure 2.6 showed some better results than the original approach after 15 iterations.



(a) Performance of SSet with different values of K

(b) Performance of TSSet with different values of K

Figure 2.7: Impact of K on the performance

Influence of the initial number of clusters

In this part, we will try to assess the impact of the initial number of clusters on the performance of our approach on both datasets. From Figure 2.7, we can see that the highest number of clusters equivalent to 500 gave the highest values for the different metrics for the Snippet set. The TitleSnippet set showed its best performance with an initial number of clusters 400 for the metrics AMI, H and NMI. In that sense, we can't conclude on a correlation between the initial number of clusters and the performance of the approach.

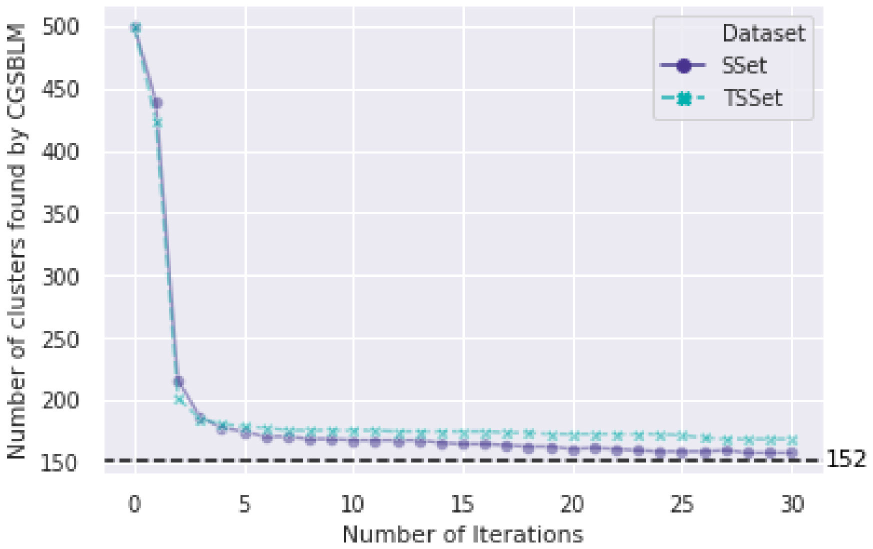


Figure 2.8: Number of clusters found by CGSBLM on SSet and TSSet Datasets

Influence of the number of iterations

In this subsection, we will describe the impact of the number of iterations on the number of clusters found by the two approaches on Snippet and TitleSnippet datasets. From Figure 2.8, we can see that the number of clusters found drops from 500 to 200 in less than 5 iterations. This corborates the assumption that the most-populated clusters will be chosen first which will lead the less popular clusters to become empty. We can also see that the final number of clusters found for both datasets is very near the original number of clusters of the Google News dataset. This shows how efficient our approach is as it estimates very well the final number of clusters.

Influence of the parameter α

In this subsection, we will see how the model performs when we change the value of the parameter α. We set the initial number of clusters K to 200 and the number of iterations to 15. We tried different values of α : 0.001,0.01,0.02,0.1,0.2. From Figure 2.9, we can see how the value of NMI which is the harmonic mean between the homogeneity and completeness is at its best for α=0.01.

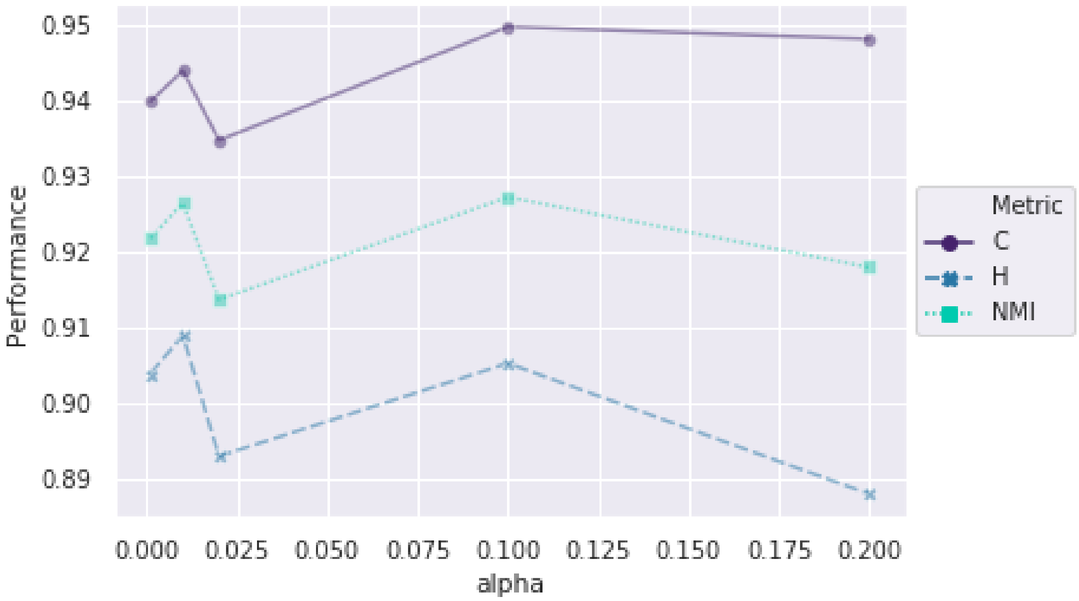


Figure 2.9: C, H and NMI for different values of α for TSSet Dataset

Chapter 3

Online Short Text Clustering using Infinite Extensions of Discrete Mixture Models

In this chapter, we detail the use of infinite extensions of discrete mixture models on short text clustering [40] [41]. We propose two approaches that use the more general priors : generalized Dirichlet and Beta-Liouville and compare them to the Dirichlet distribution. To improve this work, we also propose two approaches that make use of the online clustering in the initialization process of our algorithm. Our work was evaluated on two main datasets : Google News and Tweet. We finish our work by an outliers detection application.

3.1 Mixture Models

The distributions presented in the previous section are the main elements of the probabilistic model known as the mixture model [42]. Used to detect patterns within a group of elements, its main advantage is that it does not require prior knowledge of the subgroup to which an element may belong to. Formally, a finite mixture model representing $\vec{X}$ can be expressed as follow:

$$p(\vec{X}|\vec{\pi},\vec{\theta}) = \sum_{j=1}^{K} \pi_j p(\vec{X}|\theta_j) \tag{37}$$

where $\vec{\pi} = (\pi_1, \ldots, \pi_K)$ represents the vector of mixing probabilities which are positive and sum to one and $\vec{\theta} = (\theta_1, \ldots, \theta_K)$ represents the parameters of probability density functions associated to the different mixture components.

More advanced work bring finite mixture models to the infinite landscape. Indeed, when working with finite mixture models, an important limitation is to fix an appropriate number of components K describing the data. Infinite mixture models alleviate this problem by considering that the number of mixtures is infinite.

- **Infinite Dirichlet Multinomial Mixture Model:** which takes a Dirichlet process as a prior. The infinite Dirichlet prior that estimates the multinomial parameter is constructed by a stick-breaking construction [43]. In this framework, the Dirichlet process has two parameters, a basic distribution H and a concentration parameter Ψ. This Dirichlet formulation can be written as $DP(\Psi,H)$. The second change is that K becomes non-fixed and tends to infinity, as shown in [44].

- **Infinite generalized Dirichlet Multinomial Mixture Model:** the use of the generalized Dirichlet distribution as a prerequisite for the mixture model in an infinite setting requires the use of a mathematical property of the generalized Dirichlet distribution. Indeed, the generalized Dirichlet distribution has a structural property called neutrality which allows the mutual independence of the vector of proportions of the data points. This property is presented and discussed in more detail in [45]. The property used requires a change in the space of the original data point X into another data point ξ where the features are conditionally independent. The generalized Dirichlet distribution is written as follows:

$$GD(\vec{\xi_i}|\alpha, \delta) = \prod_{w=1}^{V} Beta(\vec{\xi_{iw}}|\alpha_w, \delta_w) \tag{38}$$

where $\vec{\xi_i} = (\xi_{i1}, \ldots, \xi_{iV})$, $\xi_{i1} = X_{i1}$, $\xi_{iw} = X_{iw}/(1-\sum_{f=1}^{w-1} X_{if})$ for $l > 1$ and $Beta(\xi_{iw}|\alpha_w, \delta_w)$ is a Beta distribution defined with parameters (α_w,δ_w).

- **Infinite Beta-Liouville Multinomial Mixture Model:** as for the priors presented earlier, the Beta-Liouville distribution can be considered in an infinite framework obtained when we use the stick-breaking framework and make the number of components infinite [46] [47].

3.2 Collapsed Gibbs Sampling

Collapsed Gibbs sampling is a Markov chain Monte Carlo algorithm used in many works such as in [32] to approximate observations of multivariate probability distributions. Given the particular nature of infinite mixture models where the number of components is considered to tend to infinity, at each clustering step, the document is either assigned to an existing cluster among the set of already known clusters or assigned to a newly created cluster. The clustering approach is then divided between the choice of an existing cluster or the choice of a new cluster as in [48].

3.2.1 Existing Cluster

For words in documents generated by mixture models, the clustering assignment is derived from the probability that a document d chooses a cluster knowing the assignment of other documents and their information. As shown in [48], this conditional probability is:

$$p(z_d = k|z_{\neg d}, \vec{d}, \alpha, \beta) \propto p(z_d = k|z_{\neg d}, \alpha)p(d|z_d = k, d_{k,\neg d}, \beta) \tag{39}$$

where k is the latent cluster that is sampled from the distribution, $z_{\neg d}$ is the assignment to the cluster of all documents excluding document d , $\vec{d}$ are the observed documents, $d_{k,\neg d}$ are the other documents currently assigned to the k cluster, and α and β are the parameters of the Dirichlet distributions in the mixture model that will help into estimating the parameters of the multinomial distributions with parameters θ and ϕ.

In equation (38) we have :

- $p(z_d = k|\vec{z}_{\neg d}, \alpha)$ is the probability that a document chooses a cluster given the clustering assignments of the other documents. It is obtained by integration on the multinomial parameter θ.

- $p(d|z_d = k, \vec{d}_{k,\neg d}, \beta)$ is the predictive probability of a document given all the other documents that are already assigned to cluster k. It is obtained by integration on the parameter of the second multinomial distribution ϕ in our mixture model.

In our case study, the second term remains the same for all the working priors we use because this part of the mixture model will always be generated by an infinite multinomial Dirichlet mixture model following the work in [48]. It is as follows:

$$p(d|z_d = k, \vec{d}_{k,\neg d}, \beta) = \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} (n_{k,\neg d}^w + \beta + j - 1)}{\prod_{i=1}^{N_d} (n_{k,\neg d} + V\beta + i - 1)} \tag{40}$$

where N_d^w is the number of occurrences of the word w in the document d, $n_{k,\neg d}^w$ is the number of occurrences of the word w in the cluster z, N_d is the number of words in the document d, $n_{k,\neg d}$ is the number of words in the cluster k, and V is the vocabulary size. Our contribution focuses on the first term $p(z_d = k|\vec{z}_{\neg d}, \alpha)$ which is generated in [48] by an infinite Dirichlet mixture. In our work, we compute the equivalent of this probability for the infinite generalized Dirichlet mixture model and the infinite Beta-Liouville mixture model.

- **Infinite Generalized Dirichlet Multinomial Mixture Model :**

  The first term of equation (38) is obtained by integration over the θ parameter of the multinomial distribution. This leads to the use of the Sum Rule of Probability, the Product Rule of Probability and the properties of the D-Separation. We integrate the posterior distribution of θ which is multiplied by the multinomial distribution. To derive the posterior distribution, we use Bayes' rule as follows:

$$
\begin{aligned}
p(\theta|z\vec{_{\neg d}}, \alpha/K, \delta) &= \frac{p(\theta|\alpha/K, \delta)p(\vec{z}_{\neg d}|\theta)}{\int p(\theta|\alpha/K, \delta)p(\vec{z}_{\neg d}|\theta)d\theta} \\
&= \frac{\prod_{k=1}^{K} Beta(\xi|\alpha/K, \delta)\prod_{k=1}^{K} \xi^{m_{k,\neg d}}}{\prod_{k=1}^{K} \int Beta(\xi|\alpha/K, \delta)\xi^{m_{k,\neg d}}d\xi}
\end{aligned}
\tag{41}
$$

where ξ is the transformed data point, α/K and δ are the parameters of the Beta distribution and $m_{k,\neg d}$ is the number of documents present in the k cluster. We have:

$$
\int \frac{\xi^{m_{k,\neg d}+\alpha/K-1}(1-\xi)^{\delta-1}}{B(\alpha/K + m_{k,\neg d}, \delta)}d\xi = 1
$$

Then,

$$
\int \xi^{m_{k,\neg d}+\alpha/K-1}(1-\xi)^{\delta-1}d\xi = B(\alpha/K + m_{k,\neg d}, \delta)
\tag{42}
$$

Replacing in equation (40) gives :

$$
\begin{aligned}
p(\theta|z\vec{_{\neg d}}, \alpha/K, \delta) &= \frac{\prod_{k=1}^{K} \frac{\xi^{m_{k,\neg d}+\alpha/K-1}(1-\xi)^{\delta-1}}{B(\alpha/K,\delta)}}{\prod_{k=1}^{K} \frac{B(\alpha/K+m_{k,\neg d},\delta)}{B(\alpha/K,\delta)}} \\
&= \prod_{k=1}^{K} \frac{\xi^{m_{k,\neg d}+\alpha/K-1}(1-\xi)^{\delta-1}}{B(\alpha/K + m_{k,\neg d}, \delta)} \\
&= \prod_{k=1}^{K} Beta(\xi|\alpha/K + m_{k,\neg d}, \delta)
\end{aligned}
\tag{43}
$$

With these calculations, we can derive the first term of the conditional probability in equation (38) :

$$
\begin{aligned}
p(z_d = k|z\vec{_{\neg d}}, \alpha/K, \delta) &= \int Beta(\xi|\alpha/K + m_{k,\neg d}, \delta)p(z_d = k|\xi)d\xi \\
&= \frac{B(\alpha/K + m_k, \delta)}{B(\alpha/K + m_{k,\neg d}, \delta)} = \frac{\frac{\Gamma(\alpha/K+m_k)\Gamma(\delta)}{\Gamma(\alpha/K+m_k+\delta)}}{\frac{\Gamma(\alpha/K+m_{k,\neg d})\Gamma(\delta)}{\Gamma(\alpha/K+m_{k,\neg d}+\delta)}}
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{\Gamma(\alpha/K + m_{k,\neg d} + 1)\Gamma(\alpha/K + m_{k,\neg d} + \delta)}{\Gamma(\alpha/K + m_{k,\neg d})\Gamma(\alpha/K + m_{k,\neg d} + \delta + 1)} \\
&= \frac{\alpha/K + m_{k,\neg d}}{\alpha/K + m_{k,\neg d} + \delta}
\end{aligned} \tag{44}
$$

In our case of study, K tend to infinity which makes α/K tend to 0. The conditional probability in (38) is as follow :

$$
p(z_d = k | z_{\neg d}, \vec{d}, \alpha, \beta) \propto \frac{m_{k,\neg d}}{m_{k,\neg d} + \delta} \frac{\prod_{w \in d} \prod_{j=1}^{N_d^w} (n_{k,\neg d}^w + \beta + j - 1)}{\prod_{i=1}^{N_d} (n_{k,\neg d} + V\beta + i - 1)} \tag{45}
$$

- **Infinite Beta-Liouville Multinomial Mixture Model :**

The procedure and rules described above can be used for the infinite Beta-Liouville approach where the posterior distribution is derived as follows :

$$
\begin{aligned}
p(\theta | \vec{z_{\neg d}}, \alpha_i, \alpha/K, \delta) &= \frac{p(\theta|\alpha_k, \alpha/K, \delta) p(\vec{z}_{\neg d}|\theta)}{\int p(\theta|\alpha_k, \alpha/K, \delta) p(\vec{z}_{\neg d}|\theta) d\theta} \\
&= \frac{\prod_{k=1}^{K} \theta_k^{\alpha_k + m_{k,\neg d} - 1} (\sum_{k=1}^{K} \theta_k)^{\alpha/K - \sum_{k=1}^{K} \alpha_k}}{\int \prod_{k=1}^{K} \theta_k^{\alpha_k + m_{k,\neg d} - 1} (\sum_{k=1}^{k} \theta_k)^{\alpha/K - \sum_{k=1}^{K} \alpha_k}} \\
&\quad \frac{(1 - \sum_{k=1}^{K} \theta_k)^{\delta - 1}}{(1 - \sum_{k=1}^{K} \theta_k)^{\delta - 1} d\theta}
\end{aligned} \tag{46}
$$

where θ is the parameter of the multinomial distribution and α_k, α/K and δ are the parameters of the Beta-Liouville distribution [26].

To obtain the equivalent of the integral in (45), we need to update the parameters following the work of [26] :

$$
\begin{cases}
\alpha^{'} = \alpha/K + \sum_{k=1}^{K-1} m_k \\
\alpha_k^{'} = \alpha_k + m_k \\
\delta^{'} = \delta + m_K
\end{cases}
$$

which results in :

$$\begin{aligned}
&\int \prod_{k=1}^{K} \theta_k^{\alpha_k + m_{k,\neg d} - 1} (\sum_{k=1}^{K} \theta_k)^{\alpha/K - \sum_{k=1}^{K} m_k} (1 - \sum_{k=1}^{K} \theta_k)^{\delta + m_K - 1} d\theta \\
&\quad = \frac{\Gamma(\alpha/K + \sum_{k=1}^{K-1} m_k)\Gamma(\delta + m_K)\Gamma(\alpha_k + m_k)}{\Gamma(\sum_{k=1}^{K}(\alpha_k + m_{k,\neg d}))\Gamma(\alpha/K + \sum_{k=1}^{K} m_k + \delta)}
\end{aligned} \tag{47}$$

We have :

$$\begin{aligned}
&p(\theta|\vec{z_{\neg d}}, \alpha_i, \alpha/K, \delta) \\
&= \frac{\prod_{k=1}^{K} \theta_k^{\alpha_k + m_{k,\neg d} - 1} (\sum_{k=1}^{K} \theta_k)^{\alpha/K - \sum_{k=1}^{K} \alpha_k} (1 - \sum_{k=1}^{K} \theta_k)^{\delta - 1}}{\frac{\Gamma(\alpha/K + \sum_{k=1}^{K-1} m_{k,\neg d})\Gamma(\delta + m_{K,\neg d})\Gamma(\alpha_k + m_{k,\neg d})}{\Gamma(\sum_{k=1}^{K}(\alpha_k + m_{k,\neg d}))\Gamma(\alpha/K + \sum_{k=1}^{K} m_{k,\neg d} + \delta)}} \\
&= \frac{\Gamma(\sum_{k=1}^{K}(\alpha_k + m_{k,\neg d}))\Gamma(\alpha/K + \sum_{k=1}^{K} m_{k,\neg d} + \delta)}{\Gamma(\alpha/K + \sum_{k=1}^{K-1} m_{k,\neg d})\Gamma(\delta + m_{K,\neg d})\Gamma(\alpha_k + m_{k,\neg d})} \\
&\quad \prod_{k=1}^{K} \theta_k^{\alpha_k + m_{k,\neg d} - 1} (\sum_{k=1}^{K} \theta_k)^{\alpha/K - \sum_{k=1}^{K} \alpha_k} (1 - \sum_{k=1}^{K} \theta_k)^{\delta - 1} \\
&= BL(\theta|\alpha_k + m_{k,\neg d}, \alpha/K + \sum_{k=1}^{K-1} m_{k,\neg d}, \delta + m_{K,\neg d})
\end{aligned} \tag{48}$$

where $m_{K,\neg d}$ is the number of documents present in the cluster K without considering the document d.

The derivation of the conditional probability is as follows:

$$\begin{aligned}
&p(z_d = k|\vec{z_{\neg d}}, \alpha/K, \delta) \\
&= \int BL(\theta|\alpha_k + m_{k,\neg d}, \alpha/K + \sum_{k=1}^{K-1} m_{k,\neg d}, \delta + m_{K,\neg d}) p(z_d = k|\theta) d\theta \\
&= \frac{\Gamma(\sum_{k=1}^{K}(\alpha_k + m_{k,\neg d}))\Gamma(\alpha/K + \sum_{k=1}^{K} m_{k,\neg d} + \delta)}{\Gamma(\alpha/K + \sum_{k=1}^{K-1} m_{k,\neg d})\Gamma(\delta + m_{K,\neg d})\Gamma(\alpha_k + m_{k,\neg d})} \\
&\quad \frac{\Gamma(\alpha/K + \sum_{k=1}^{K-1} m_k)\Gamma(\delta + m_K)\Gamma(\alpha_k + m_k)}{\Gamma(\sum_{k=1}^{K}(\alpha_k + m_k))\Gamma(\alpha/K + \sum_{k=1}^{K} m_k + \delta)} \\
&= \frac{\prod_{k=1}^{K-1}(\alpha/K + \sum_{k=1}^{K-1} m_{k,\neg d} + k - 1)(\delta + m_{K,\neg d} + 1)}{\prod_{k=1}^{K}(\sum_{k=1}^{K}(\alpha_k + m_{k,\neg d} + k - 1))} \\
&\quad \frac{(\alpha_k + m_{k,\neg d} + 1)}{\prod_{k=1}^{K}(\alpha/K + \sum_{k=1}^{K} m_{k,\neg d} + \delta + k - 1)}
\end{aligned} \tag{49}$$

The conditional probability in (38) is as follows :

$$\begin{aligned} & p(z_d = k|z_{\neg d}, \vec{d}, \alpha, \beta) \\ & \propto \frac{\prod_{k=1}^{K-1}(\sum_{k=1}^{K-1} m_{k,\neg d} + k - 1)(\delta + m_{K,\neg d} + 1)}{\prod_{k=1}^{K}(\sum_{k=1}^{K}(\alpha_k + m_{k,\neg d} + k - 1))} \\ & \quad \frac{(\alpha_k + m_{k,\neg d} + 1)\prod_{w\in d}\prod_{j=1}^{N_d^w}(n_{k,\neg d}^w + \beta + j - 1)}{\prod_{i=1}^{N_d}(n_{k,\neg d} + V\beta + i - 1)\prod_{k=1}^{K}(\sum_{k=1}^{K} m_{k,\neg d} + \delta + k - 1)} \end{aligned} \tag{50}$$

3.2.2 Choosing new cluster

Since K is an undefined value in infinite mixture models, it makes sense to compute the probability of a document d to be assigned to a newly defined cluster $K + 1$ as follows in [48]:

$$p(z_d = K + 1|z_{\neg d}, \vec{d}, \alpha, \delta, \beta) \propto p(z_d = K + 1|z_{\neg d}, \alpha, \delta)p(d|z_d = K + 1, \beta) \tag{51}$$

where $K + 1$ is the newly created cluster. The term containing the β parameter remains unchanged as in [48] :

$$p(d|z_d = K + 1, \beta) = \frac{\prod_{w\in d}\prod_{j=1}^{N_d^w}(\beta + j - 1)}{\prod_{i=1}^{N_d}(V\beta + i - 1)} \tag{52}$$

In this work, we introduce two different infinite mixture models for which we compute the conditional probability that a document d chooses a new cluster $K + 1$:

- **Infinite Generalized Dirichlet Multinomial Mixture Model :**

$$\begin{aligned} p(z_d = K + 1|z_{\neg d}, \alpha, \delta) &= 1 - \sum_{k=1}^{K} p(z_d = k|z_{\neg d}, \alpha, \delta) \\ &= 1 - \sum_{k=1}^{K} \frac{m_{k,\neg d}}{m_{k,\neg d} + \delta} \end{aligned} \tag{53}$$

The equation (50) is equivalent to :

$$p(z_d = K + 1|z_{\neg d}, \vec{d}, \alpha, \delta, \beta) \propto \frac{\prod_{w\in d}\prod_{j=1}^{N_d^w}(\beta + j - 1)}{\prod_{i=1}^{N_d}(V\beta + i - 1)}(1 - \sum_{k=1}^{K} \frac{m_{k,\neg d}}{m_{k,\neg d} + \delta}) \tag{54}$$

- **Infinite Beta-Liouville Multinomial Mixture Model :**

$$\begin{aligned}
&p(z_d = K+1|z_{\neg d}, \alpha_i, \alpha, \delta) \\
&= 1 - \sum_{k=1}^{K} p(z_d = k|z_{\neg d}, \alpha_i, \alpha, \delta) \\
&= 1 - \sum_{k=1}^{K} (\frac{\prod_{k=1}^{K-1}(\sum_{k=1}^{K-1} m_{k,\neg d} + k - 1)(\delta + m_{K,\neg d} + 1)}{\prod_{k=1}^{K}(\sum_{k=1}^{K}(\alpha_k + m_{k,\neg d} + k - 1))} \\
&\quad \frac{(\alpha_k + m_{k,\neg d} + 1)}{\prod_{k=1}^{K}(\sum_{k=1}^{K} m_{k,\neg d} + \delta + k - 1)})
\end{aligned} \tag{55}$$

The equation (50) is then equivalent to :

$$\begin{aligned}
&p(z_d = K+1|z_{\neg d}, \vec{d}, \alpha_k, \alpha, \delta, \beta) \\
&\propto \frac{\prod_{w \in d}\prod_{j=1}^{N_d^w}(\beta + j - 1)}{\prod_{i=1}^{N_d}(V\beta + i - 1)} (1 - \sum_{k=1}^{K} (\frac{\prod_{k=1}^{K-1}(\sum_{k=1}^{K-1} m_{k,\neg d} + k - 1)}{\prod_{k=1}^{K}(\sum_{k=1}^{K}(\alpha_k + m_{k,\neg d} + k - 1))} \\
&\quad \frac{(\delta + m_{K,\neg d} + 1)(\alpha_k + m_{k,\neg d} + 1)}{\prod_{k=1}^{K}(\sum_{k=1}^{K} m_{k,\neg d} + \delta + k - 1)}))
\end{aligned} \tag{56}$$

3.3 Proposed Algorithm

The collapsed Gibbs sampling algorithm for infinite mixture models is presented in Algorithm 1. The main objective of this algorithm is to assign each of the available D documents to a given cluster. The main variables of the algorithm are:

(1) m_k : number of documents in cluster k.

(2) n_k : number of words in cluster k.

(3) n_k^w : number of time the word w appeared in cluster k.

The presented parameters are initialized to zero and the initial number of clusters K_{init} is set to 1. After this initialization step, each document is randomly assigned to a cluster. Then, the main variables of the algorithm are incremented by 1 for m_k, by L for n_k and by N for n_k^w. After this initialization step, we start applying our collapsed Gibbs sampling algorithm for a predefined number of iterations I. At each iteration and for each document, the currently assigned cluster is

Algorithm 2 CGS for Infinite Multinomial Mixture

```
Set $m_k$ = 0, $n_k$ = 0, $n_k^w$ = $\{\}$ and $K$ = 1
for all the Documents do
  Sample $k_d \sim Multinomial(1/\mathrm{K})$
  $m_k {+}{=}\, 1, n_k {+}{=}\, L \; and \; n_k^w {+}{=}\, N$
end for
for all the Iterations do
  Assign $k_d$
  Set $m_k$ = $m_k$ - 1 , $n_k$ = $n_k$ - $L$ and $n_k^w$ = $n_k^w$ - $N$
  if $n_z$ == 0 then
  Decrease $K$ by 1
  Remove inactive clusters
  Sample $k_d \sim p(k|k_{\neg d}, \vec{d})$ and $p(K+1|K+1_{\neg d}, \vec{d})$
  if $k_d$ is active then
   $m_k {+}{=}\, 1, n_k {+}{=}\, L$ and $n_k^w {+}{=}\, N$
  else
   Create a new cluster
   Initialize its main variables to 0
end for
```

recorded and the main variables are decremented with the same amounts they were incremented with. If a cluster is considered empty, it is deleted and the indices of the non-empty clusters is reordered. Then, each document is assigned to a new cluster according to the probabilities obtained from equations (44) and (53) if we are dealing with the infinite Generalized Dirichlet mixture model or from equations (49) and (55) if we are dealing with the infinite Beta-Liouville mixture model. A new index is then sampled from the calculated probabilities. If the calculated index already exists, the main variables are incremented accordingly. Otherwise, a new cluster is created by initializing its principal variables to zero.

3.4 Online Clustering Initialization

In the proposed Algorithm 1, we can see that the collapsed Gibbs sampling algorithm assigns each document a cluster randomly, which adds to the uncertainty of the clustering process. A solution to this problem was proposed in [44] in the form of an online clustering scheme for initialization. The full name that has been given to the algorithm is Fast GSDMM+ when it uses the

Dirichlet mixture model. This algorithm is an improvement of the FGSDMM algorithm also described in [44].

In Algorithm 3, we can see how the initialization step of the algorithm works. First, all documents are assigned to a common unique cluster. Then, one by one, the documents start choosing between the already existing clusters or the newly created cluster. At each step, the new clusters are created one by one, which lets the documents choose between non-empty clusters and an empty cluster. The decision is made based on the results of the probabilities, as explained in the previous algorithm. When a decision about a new cluster is made, the algorithm creates this new cluster by storing this document. After this initialization process, our collapsed Gibbs sampling algorithm for infinite mixture models is run for a number of iterations until a fixed number of clusters are obtained. This algorithm has proven in a previous work to reduce the spatial complexity and randomness of the initialization process.

Algorithm 3 CGS for Infinite Multinomial Mixture with Online Initialization

```
Set $m_k$ = 0, $n_k$ = 0, $n_k^w$ = {} and $K$ = 0
for all the Documents do
  Sample $k_d \sim p(k|k_{\neg d}, \vec{d})$ and $p(K+1|K+1_{\neg d}, \vec{d})$
  if $k_d$ is active then
    Assign k to d, $m_k$+= 1, $n_k$+= $L$ and $n_k^w$+= $N$
  else
    Increment $K$
    Assign $K+1$ to d, $m_{K+1} = 0$, $n_{K+1} = 0$ and $n_{K+1}^w = \{\}$
end for
Apply Algorithm 2
```

3.5 Experimental Results

3.5.1 Short-Text Datasets and Preprocessing

We evaluate the proposed approaches on two datasets used in [48]. The first dataset is the Google News containing 11,109 news articles with 152 topics. The dataset was divided into three datasets (TSet, SSet, TSSet). Two of them contain very short texts, TSet containing the titles of the articles and SSet containing the excerpts of the articles. TSSet contains both the titles of the articles and their excerpts. The second dataset is the TweetSet [48] which contains 2,472 tweets belonging

Table 3.1: Performance of the Approaches on the datasets

Datasets	Metric	GSDPMM	Infinite GSGDMM	Infinite GSBLMM
TSSet	c	0.94	0.95	0.95
	h	0.84	0.86	0.91
	NMI	0.89	0.90	0.93
TweetSet	c	0.82	0.82	0.8
	h	0.88	0.86	0.93
	NMI	0.85	0.84	0.87
TSet	c	0.89	0.89	0.88
	h	0.83	0.86	0.90
	NMI	0.86	0.88	0.89
SSet	c	0.91	0.91	0.92
	h	0.79	0.83	0.84
	NMI	0.85	0.87	0.88

to 89 queries. Text preprocessing includes lowercasing all words, removing non-Latin characters and stop words, using NLTK's WordNet Lemmatizer to apply uprooting, keeping sentences between 2 and 15 words, and removing words with a frequency of less than 2.

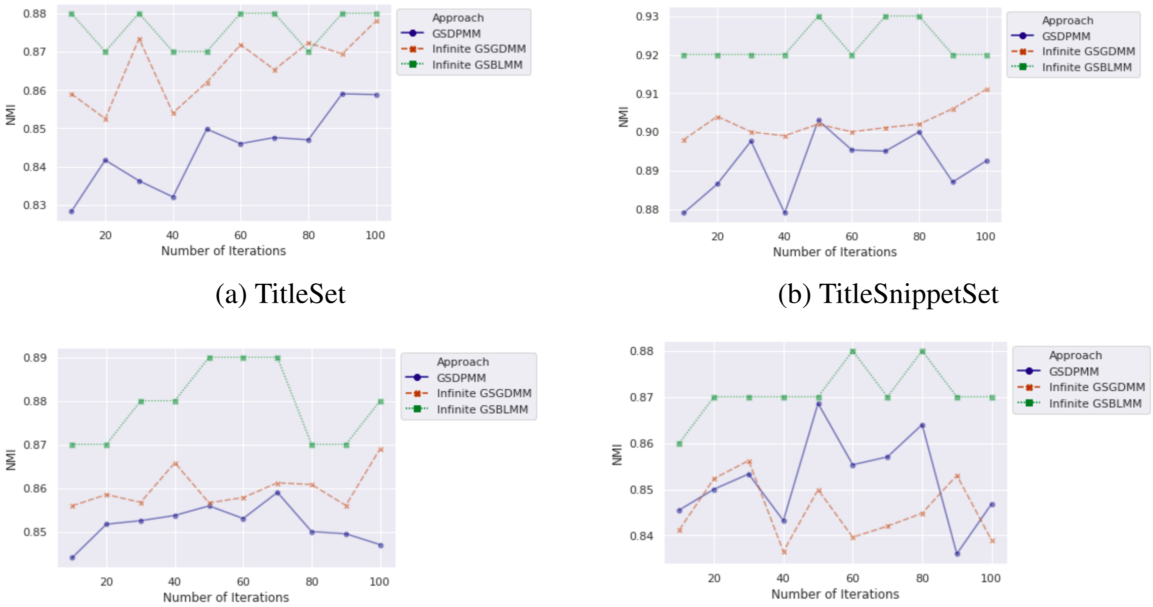
3.5.2 Evaluation Metrics

For a fair evaluation of the work, the same metrics used in [48] considered as we compare our approaches with the results listed in [48]. Those metrics are detailed in chapter 2 where c is the completeness and h is the homogeneity.

3.5.3 Comparison between Infinite Mixture Models

In this subsection, we discuss the comparison between the different infinite mixture models. We set β to 0.02, δ to 0.5 and α_k to $5 * 10^{-4}$. We set the number of iterations to 100 and the initial number of clusters K to 1. We list in table 1, the results obtained when applying the GSDPMM and the two proposed approaches (Infinite GSGDMM, Infinite GSBLMM) on the four previously mentioned datasets.

We prove from Table 3.1 that using the infinite generalized Dirichlet model as the prior performed better than the GSDPMM when considering the NMI metric. We can also see that the Beta-Liouville infinite multinomial mixture model performs better than the two previously mentioned approaches. This shows that the use of priors with more flexible covariance is effective for



(a) TitleSet

(b) TitleSnippetSet

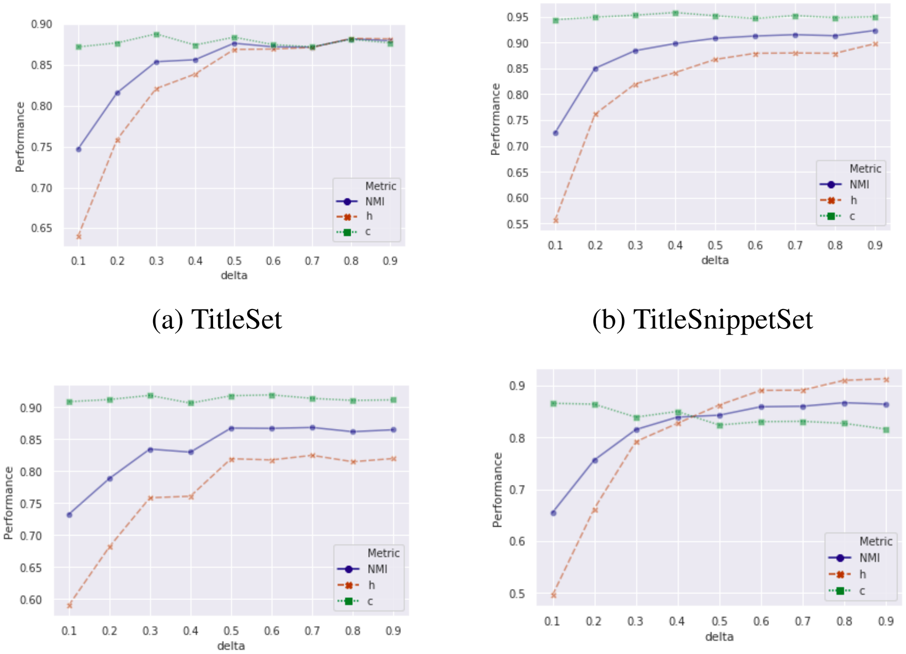
(c) SnippetSet

(d) TweetSet

Figure 3.1: Comparison between Infinite Mixture Models

classifying short texts and agrees with the theoretical assumptions. The other two proposed metrics, completeness and homogeneity, showed overall better results than the GSDPMM approach. Considering the homogeneity metric for all datasets, we notice that the results are improved by at least 5% while taking into account the completeness, there is no great improvement. This shows that both proposed approaches improve the results of clustering by clusters, where they were more successful in detecting whether each cluster contains only observations belonging to the same ground truth.

In Figure 3.1, we can clearly see how the infinite Beta-Liouville multinomial mixture model approach performed better than the other two approaches over a large set of iterations ranging from 10 to 100. This is especially clear for the SnippetSet short text dataset. We can also see that the approach performed well on the TitleSnippetSet dataset. Infinite GSGDMM also performed well over the same range of iterations compared to GSDPMM on TitleSet, TitleSnippetSet and SnippetSet. However, we can see from Figure 3.1 (d) that the infinite GSGDMM approach did not outperform the GSDPMM approach when used on the TweetSet dataset. It is clear that the infinite GSGDMM approach did not perform well for almost any number of iterations compared to the GSDPMM approach. But, we can see better results when the number of iterations is set to 20, 30



(a) TitleSet

(b) TitleSnippetSet

(c) SnippetSet

(d) TweetSet

Figure 3.2: Performance of infinite GSGDMM given δ

and 90.

3.5.4 Performance of the infinite GSGDMM

In this subsection, we evaluate the effect of the δ parameter which is the newly introduced parameter compared to the GSDPMM approach. In Figure 3.2, we see how a larger value of δ gives a better value for the h and NMI metrics. The improvement is evident on all datasets where the best values were obtained for a value of the δ parameter of 0.9. On the other hand, the c metric did not obtain as much improvement for almost all tested values of the δ parameter. The curve is almost constant over the different values of the δ parameter for the TitleSnippetSet and SnippetSet datasets. This same c metric decreases at a very slight rate for TweetSet. This confirms the observation made in the previous subsection where we deduced that our proposed approaches only improved the clustering performance cluster-wise, leaving the data-wise clustering unchanged. We also tried to evaluate the role of the parameter δ on the approximation of the number of clusters found for the four datasets. Figure 3.3 shows that a higher value of δ gives a higher number of

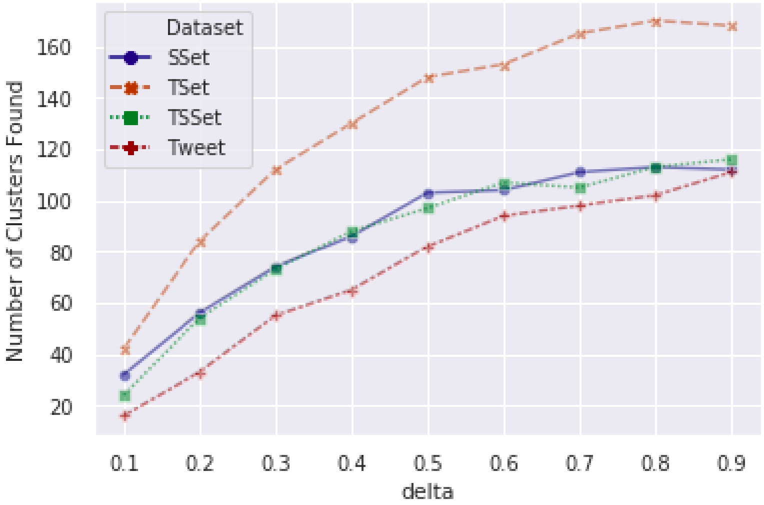


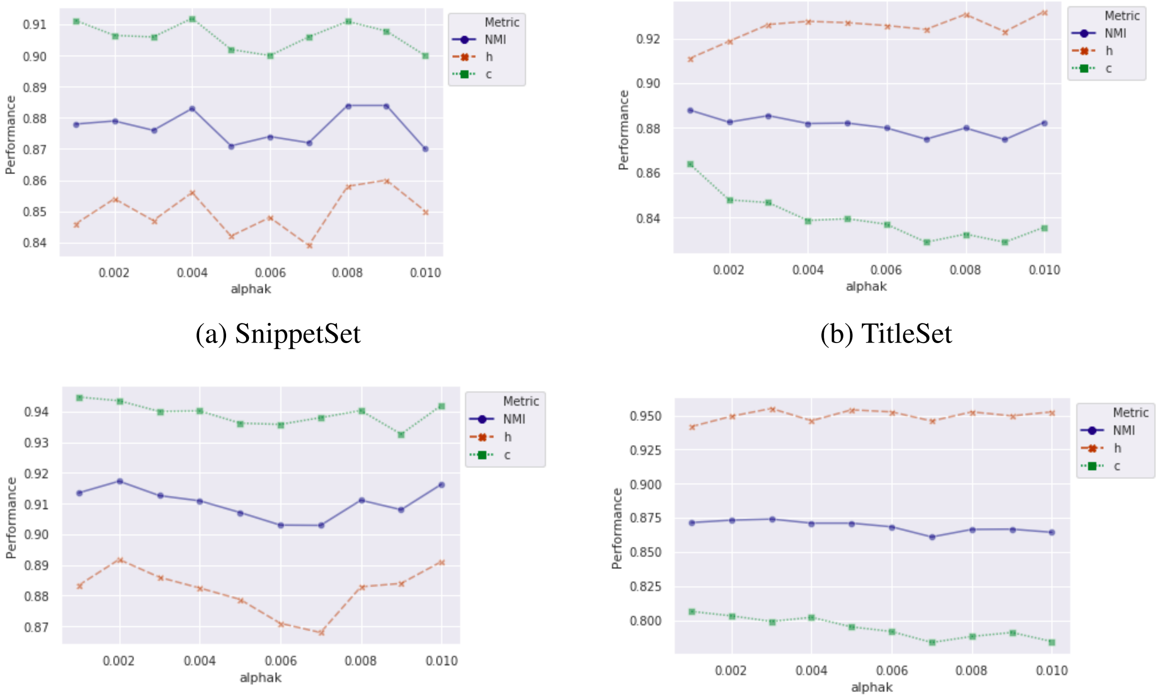
Figure 3.3: Number of clusters found by GSGDMM given δ

clusters found. We can see that for the SSet and TSSet datasets, the maximum number of clusters found is about 120 clusters for a value of δ equal to $0, 9$ while the actual number of clusters assigned is 152. Digging deeper into the graph, we see that for this same value of δ, TSet converges to a larger number of clusters around 163 clusters. For TweetSet, the number of clusters found for this same value of δ is slightly less than 120 while the real number of clusters is supposed to be 89. This means that it would be wiser to opt for a lower δ value for both the T and Tweet datasets. We can see that a value of δ between 0.5 and 0.6 should give a value closer to the actual number of clusters found for the Tweet dataset while a value between 0.6 and 0.7 should be more appropriate for the T dataset. Further experiments with higher values of δ should be conducted for the remaining two datasets. This should not be at the expense of the values given by the different metrics which should be closely monitored, while taking into account some trade-offs.

3.5.5 Performance of the infinite GSBLMM

As explained in the previous sections, the Beta-Liouville priority introduces a new parameter α_k whose effect will be studied in this subsection.

In Figure 3.4, we evaluated the NMI, C, and H metrics on the four datasets for α_k values ranging



(a) SnippetSet

(b) TitleSet

(c) TitleSnippetSet

(d) TweetSet

Figure 3.4: Performance of infinite GSBLMM given α_k

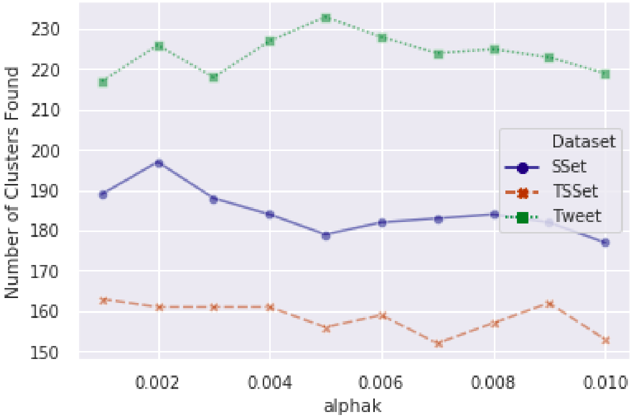


Figure 3.5: Number of clusters found given α_k

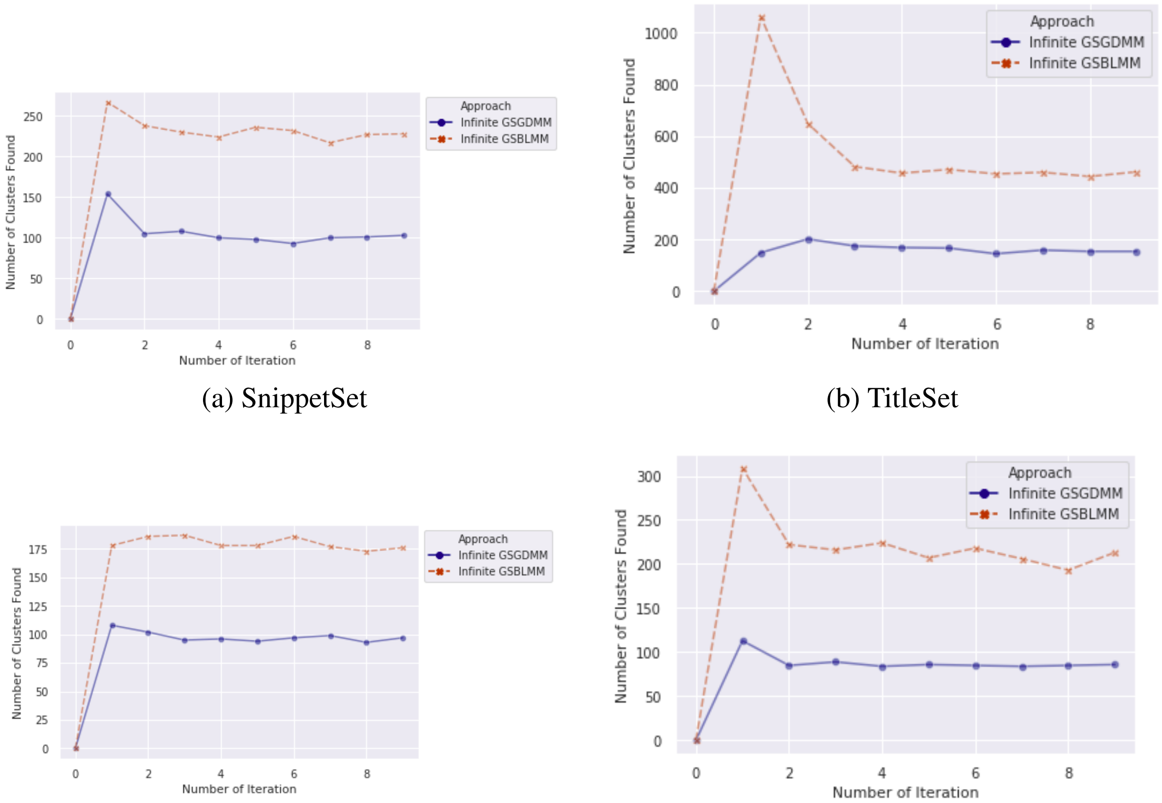
from 0.001 to 0.01. For the TweetSet, we can see that the lower the value of the α_k parameter, the better the results obtained for the NMI and c metrics while the curve of the h metric remains constant for the different values of α_k. This observation is the same for TSet, except that we can see an improvement of the value of H. Indeed, the larger the value of α_k is, the better the performance of the h metric is. For the other two datasets (TSSet, SSet), there is no clear pattern to infer from its behavior for our proposed approach. We can see that for TSSet, there is first a decrease in performance but then a clear increase from a value of α_k equal to 0.008. We also tried to see the impact of the α_k parameter on the final number of clusters found. From Figure 3.5, we found that compared to the actual number of clusters, the number of clusters found is high for both datasets used: SSet and TweetSet. However, it should be noted that the higher the value of α_k, the more likely we are to get a low number of found clusters. The TSSet dataset performs well on the whole range of α_k tested. Indeed, we can see that for the different values, we are in a realistic range of the number of clusters found between 150 and 165.

3.5.6 Convergence of the proposed Approaches

In this subsection, we want to study the number of clusters found for the first 10 iterations using the two proposed approaches. Figure 3.6 shows that there is some convergence in the number of clusters found starting from the iteration number 2. This is especially true for the first infinite GSGDMM approach which also gives a number of clusters found closer to the actual number of clusters. The second proposed approach infinite GSBLMM gives a number of clusters found for TSSet closer to the real number of clusters while it gives values much higher than the actual number of clusters for the other datasets. Computing more iterations and better fine-tuning should reduce this discrepancy.

3.5.7 Performance with Online Clustering

In this subsection, we evaluate the performance of our approaches and the GSDPMM approach with a change in the initialization step as previously presented in Algorithm 3. Indeed, we presented the FGSDMM+ algorithm that uses an online initialization clustering procedure to assign documents to initial clusters. This reduces the noise in the clustering process. The experimental



(a) SnippetSet

(b) TitleSet

(c) TitleSnippetSet

(d) TweetSet

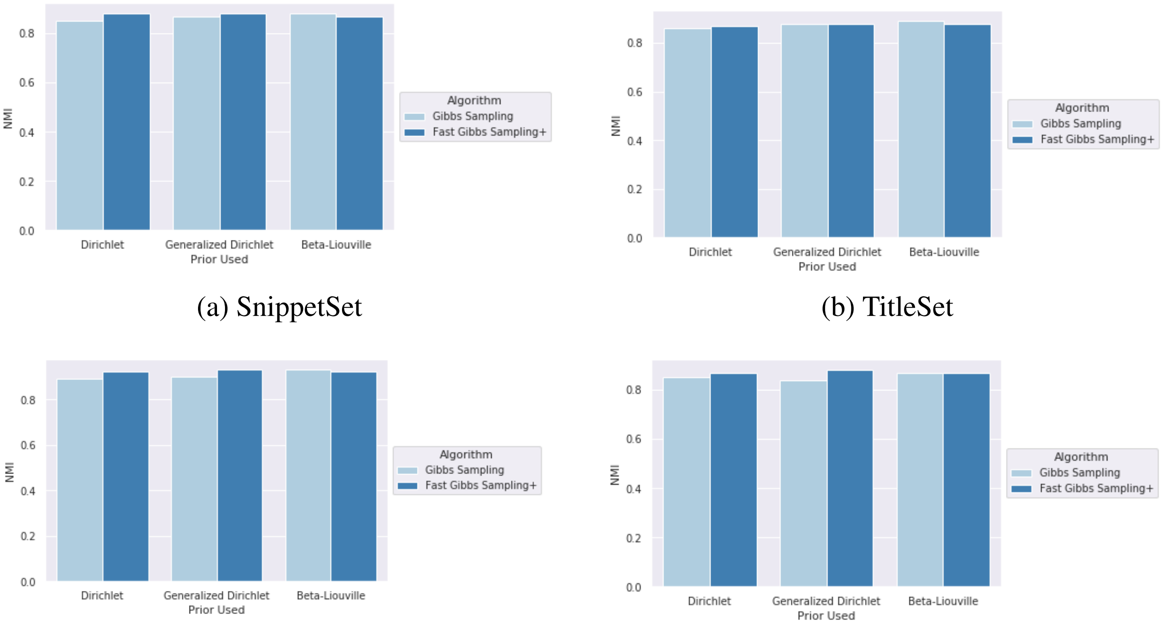
Figure 3.6: Number of Clusters Found over the 10 first Iterations

Table 3.2: Performance of the approaches with online clustering on the datasets

Datasets	Metric	GSDPMM	Infinite GSGDMM	Infinite GSBLMM
TSSet	c	0.95	0.95	0.95
	h	0.88	0.91	0.9
	NMI	0.92	0.93	0.92
TweetSet	c	0.82	0.81	0.86
	h	0.93	0.95	0.93
	NMI	0.87	0.88	0.87
TSet	c	0.87	0.87	0.87
	h	0.88	0.9	0.89
	NMI	0.87	0.88	0.88
SSet	c	0.92	0.91	0.91
	h	0.85	0.84	0.83
	NMI	0.88	0.88	0.87

results are shown in Table 3.2 where, following the progression of the NMI metric, it is clear that the infinite FGSGDMM+ gives the best results. Looking at the c metric in more detail, we see that the infinite FGSBLMM+ performs better in cluster-wise. Overall, both of our approaches perform better than the FGSDPMM+ approach.

To address the main objective of this subsection, we compare the performance of the Gibbs sampling approach with the approach proposed by the Fast Gibbs Sampling+ algorithm. Figure 3.7 gives a comparison between the algorithm with online clustering initialization and the one without initialization. On all datasets, the algorithm using online clustering initialization performs better than the one without online clustering initialization. The infinite FGSBLMM+ approach lost 0.01 in performance on the TS, T and S datasets while giving the same results on the Tweet dataset compared to the baseline approach.



(a) SnippetSet

(b) TitleSet

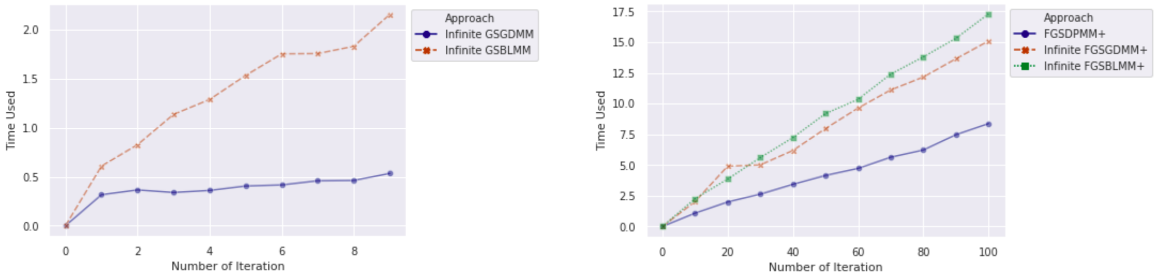
(c) TitleSnippetSet

(d) TweetSet

Figure 3.7: Comparision between the proposed algorithms

3.5.8 Speed of the approaches

In this subsection, we assess the speed of the approaches for a different number of iterations. We can see in Figure 3.8 (a) how the infinite GSBLMM is taking more computational time to give the clustering result. Figure 3.8 (b) shows the same when using the algorithms with online clustering. It also shows how the FGSDPMM+ algorithm is performing faster than our two proposed approaches.



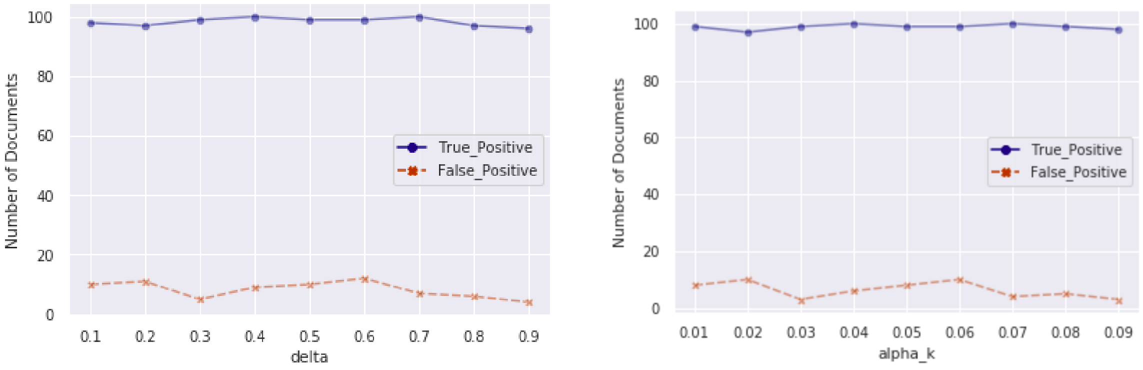
(a)

(b)

Figure 3.8: Time used in seconds with different number of iterations

3.5.9 Outlier Detection

In this subsection, we assess the performance of our proposed approaches on detecting outliers. We added manually to the TS dataset 100 documents stemming from different datasets collected from different website resources. We call the new dataset Outlier TS. As presented in [48], we run the proposed algorithms with the proposed approaches on the OutlierTS dataset with different values of δ for the generalized Dirichlet prior and different values of α_k for the Beta-Liouville prior. We fixed the number of iterations to 5 and fixed β to 0.02. We consider the clusters with only one documents as outliers.



(a) Infinite FGSGDMM+

(b) Infinite FGSBLMM+

Figure 3.9: outliers detection

Figure 3.9 shows that both of the proposed approaches are giving a good performance on the outlier detection task using the created dataset. We find that the infinite FGSBLMM+ performs slightly better. We can also see that both of the approaches detect only a maximum of 12 documents as false positive. In this perspective, infinite FGSBLMM+ is doing a better job as it is detecting a number of false positive less than 10. It is also performing better when it comes to detecting the actual outliers.

Chapter 4

Conclusion

In this thesis, we developed a number of approaches for short text clustering to alleviate on the challenges presented by such data. We proposed two different approaches using collapsed Gibbs sampling for mixture models and two different approaches using collapsed Gibbs sampling for infinite extensions of discrete mixture models. We also elaborated two other approaches, Infinite FGSGDMM+ and Infinite FGSBLMM+, making use of online clustering.

In chapter 2, we presented the functioning of the baseline approach Gibbs sampling for Dirichlet Multinomial Model. Then, we proposed our approaches that make use in the first place of the generalized Dirichlet distribution (CGSGDMM) then of the Beta-Liouville distribution (CGSBLMM). Our proposed approaches proved to be more efficient when classifying short texts while being able to cope with the challenges that they present.

In chapter 3, we elaborated on the previously presented work by presenting approaches that use the infinite extensions of discrete mixture models. This work led to the proposal of two approaches : Infinite Gibbs Sampling Generalized Dirichlet Multinomial Mixture Model (Infinite GSGDMM) and Infinite Gibbs Sampling Beta-Liouville Multinomial Mixture Model (Infinite GSBLMM). They proved to be more efficient than the GSDPMM approach on which our work is based. We topped this work by tuning the initialization process using an online approach. This led to two approaches named : Infinite FGSGDMM+ and Infinite FGSBLMM+. These new approaches presented better performance not only as compared to the baseline approach but also as compared to our proposed approaches Infinite GSGDMM and Infinite GSBLMM. We finished our work by an outliers detection application.

All the presented approaches outperformed the models using the Dirichlet distribution as a prior and coped extremely well with the challenges presented by short texts.

Future work might include mainly the proposal of more advanced priors to the multinomial distribution such as the Scaled Dirichlet distribution and the Shifted-Scaled distribution while considering other learning techniques [49].

Bibliography

[1] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[2] Diane J Hu. Latent dirichlet allocation for text, images, and music. *University of California, San Diego. Retrieved April*, 26:2013, 2009.

[3] Charu C Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer, 2012.

[4] Nizar Bouguila and Tarek Elguebaly. A fully bayesian model based on reversible jump MCMC and finite beta mixtures for clustering. *Expert Syst. Appl.*, 39(5):5946–5959, 2012.

[5] Issa Alsmadi and Gan Keng Hoon. Term weighting scheme for short-text classification: Twitter corpuses. *Neural Computing and Applications*, 31(8):3819–3831, 2019.

[6] Nizar Bouguila. A model-based approach for discrete data clustering and feature weighting using MAP and stochastic complexity. *IEEE Trans. Knowl. Data Eng.*, 21(12):1649–1664, 2009.

[7] Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R Lyu, and Irwin King. Topic memory networks for short text classification. *arXiv preprint arXiv:1809.03664*, 2018.

[8] Ou Jin, Nathan N Liu, Kai Zhao, Yong Yu, and Qiang Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 775–784, 2011.

[9] Cicero Dos Santos and Maira Gatti. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78, 2014.

[10] Ji Young Lee and Franck Dernoncourt. Sequential short-text classification with recurrent and convolutional neural networks. *arXiv preprint arXiv:1603.03827*, 2016.

[11] Nizar Bouguila and Walid ElGuebaly. On discrete data clustering. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 503–510. Springer, 2008.

[12] Nizar Bouguila and Walid ElGuebaly. Discrete data clustering using finite mixture models. *Pattern Recognit.*, 42(1):33–42, 2009.

[13] Nizar Bouguila and Ola Amayri. A discrete mixture-based kernel for svms: Application to spam and image categorization. *Inf. Process. Manag.*, 45(6):631–642, 2009.

[14] Nizar Bouguila and Djemel Ziou. Mml-based approach for finite dirichlet mixture estimation and selection. In Petra Perner and Atsushi Imiya, editors, *Machine Learning and Data Mining in Pattern Recognition, 4th International Conference, MLDM 2005, Leipzig, Germany, July 9-11, 2005, Proceedings*, volume 3587 of *Lecture Notes in Computer Science*, pages 42–51. Springer, 2005.

[15] Nuha Zamzami and Nizar Bouguila. Model selection and application to high-dimensional count data clustering. *Applied Intelligence*, 49(4):1467–1488, 2019.

[16] Nizar Bouguila and Walid ElGuebaly. A generative model for spatial color image databases categorization. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP 2008, March 30 - April 4, 2008, Caesars Palace, Las Vegas, Nevada, USA*, pages 821–824. IEEE, 2008.

[17] Ali Shojaee Bakhtiari and Nizar Bouguila. A variational bayes model for count data learning and classification. *Eng. Appl. Artif. Intell.*, 35:176–186, 2014.

[18] Nizar Bouguila and Mukti Nath Ghimire. Discrete visual features modeling via leave-one-out likelihood estimation and applications. *J. Vis. Commun. Image Represent.*, 21(7):613–626, 2010.

[19] Ali Shojaee Bakhtiari and Nizar Bouguila. An expandable hierarchical statistical framework for count data modeling and its application to object classification. In *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, pages 817–824. IEEE, 2011.

[20] Nuha Zamzami and Nizar Bouguila. Text modeling using multinomial scaled dirichlet distributions. In Malek Mouhoub, Samira Sadaoui, Otmane Aït Mohamed, and Moonis Ali, editors, *Recent Trends and Future Technology in Applied Intelligence - 31st International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems, IEA/AIE 2018, Montreal, QC, Canada, June 25-28, 2018, Proceedings*, volume 10868 of *Lecture Notes in Computer Science*, pages 69–80. Springer, 2018.

[21] Nuha Zamzami and Nizar Bouguila. A novel scaled dirichlet-based statistical framework for count data modeling: Unsupervised learning and exponential approximation. *Pattern Recognition*, 95:36–47, 2019.

[22] Nizar Bouguila. On multivariate binary data clustering and feature weighting. *Comput. Stat. Data Anal.*, 54(1):120–134, 2010.

[23] M Farrow. Mas3301 bayesian statistics. *Newcastle University*, 2017.

[24] Nizar Bouguila. Clustering of count data using generalized dirichlet multinomial distributions. *IEEE Transactions on Knowledge and Data Engineering*, 20(4):462–474, 2008.

[25] Robert J Connor and James E Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

[26] Nizar Bouguila. Count data modeling and classification using finite mixtures of distributions. *IEEE Transactions on Neural Networks*, 22(2):186–198, 2010.

[27] Md Hafizur Rahman and Nizar Bouguila. Efficient feature mapping in classifying proportional data. *IEEE Access*, 9:3712–3724, 2020.

[28] Samar Hannachi, Fatma Najar, and Nizar Bouguila. Short text clustering using generalized dirichlet multinomial mixture model. In *ACIIDS (Companion)*, pages 149–161, 2021.

[29] Samar Hannachi, Fatma Najar, Koffi Eddy Ihou, and Nizar Bouguila. Collapsed gibbs sampling of beta-liouville multinomial for short text clustering. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*, pages 564–571. Springer, 2021.

[30] Rajeeva L Karandikar. On the markov chain monte carlo (mcmc) method. *Sadhana*, 31(2):81–104, 2006.

[31] Tarek Elguebaly and Nizar Bouguila. Bayesian learning of generalized gaussian mixture models on biomedical images. In Friedhelm Schwenker and Neamat El Gayar, editors, *Artificial Neural Networks in Pattern Recognition, 4th IAPR TC3 Workshop, ANNPR 2010, Cairo, Egypt, April 11-13, 2010. Proceedings*, volume 5998 of *Lecture Notes in Computer Science*, pages 207–218. Springer, 2010.

[32] Jianhua Yin and Jianyong Wang. A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242, 2014.

[33] Wentao Fan and Nizar Bouguila. Online learning of a dirichlet process mixture of beta-liouville distributions via variational inference. *IEEE Trans. Neural Networks Learn. Syst.*, 24(11):1850–1862, 2013.

[34] Gregor Heinrich. Parameter estimation for text analysis. Technical report, Technical report, 2005.

[35] Somnath Banerjee, Krishnan Ramanathan, and Ajay Gupta. Clustering short texts using wikipedia. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 787–788, 2007.

[36] Jiawei Han, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[37] Hila Becker. *Identification and characterization of events in social media*. PhD thesis, Columbia University, 2011.

[38] Shaohong Zhang and Hau-San Wong. Arimp: A generalized adjusted rand index for cluster ensembles. In *2010 20th International Conference on Pattern Recognition*, pages 778–781. IEEE, 2010.

[39] Aaron F McDaid, Derek Greene, and Neil Hurley. Normalized mutual information to evaluate overlapping community finding algorithms. *arXiv preprint arXiv:1110.2515*, 2011.

[40] Nizar Bouguila. Infinite liouville mixture models with application to text and texture categorization. *Pattern Recognition Letters*, 33(2):103–110, 2012.

[41] Nizar Bouguila and Djemel Ziou. A nonparametric bayesian learning model: Application to text and image categorization. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 463–474. Springer, 2009.

[42] Nizar Bouguila and Wentao Fan. *Mixture models and applications*. Springer, 2020.

[43] Hemant Ishwaran and Lancelot F James. Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, 96(453):161–173, 2001.

[44] Jianhua Yin and Jianyong Wang. A text clustering algorithm using an online clustering scheme for initialization. In *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1995–2004, 2016.

[45] Sabri Boutemedjet, Nizar Bouguila, and Djemel Ziou. A hybrid feature extraction selection approach for high-dimensional non-gaussian data clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(8):1429–1443, 2008.

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

[47] Wentao Fan and Nizar Bouguila. Expectation propagation learning of a dirichlet process mixture of beta-liouville distributions for proportional data clustering. *Engineering Applications of Artificial Intelligence*, 43:1–14, 2015.

[48] Jianhua Yin and Jianyong Wang. A model-based approach for text clustering with outlier detection. In *2016 IEEE 32nd International Conference on Data Engineering (ICDE)*, pages 625–636. IEEE, 2016.

[49] Koffi Eddy Ihou and Nizar Bouguila. Variational-based latent generalized dirichlet allocation model in the collapsed space and applications. *Neurocomputing*, 332:372–395, 2019.