Novel Topic Models for Content Based Recommender Systems

Kamal Maanicshah, Manar Amayri and Nizar Bouguila

Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Quebec, Canada

Keywords: Topic Biterm Models, Generalized Dirichlet Distribution, Beta-Liouville Distribution, Mixture Allocation, Recommender Systems.

Abstract: Content based recommender systems play a vital role in applications related to user suggestions. In this paper, we introduce novel topic models which help tackle the recommendation task. Being one of the prominent approaches in the field of natural language processing, topic models like latent Dirichlet allocation (LDA) try to identify patterns of topics across multiple documents. Due to the proven efficiency of generalized Dirichlet allocation and Beta-Liouville allocation in recent times, we use these models for better performance. In addition, since it is a known fact that co-occurences of words are commonplace in text documents, the models have been designed with this reality in mind. Our models follow a mixture based design to achieve better topic quality. We use variational inference for estimating the parameters. Our models are validated with two different datasets for recommendation tasks.

1 INTRODUCTION

Recommendation systems have become an inseparable part of a variety of online services like web search, news articles, movies, etc. in recent years (Pazzani and Billsus, 2007). Most of the recent advancement in this field is centred on collaborative filtering (Bobadilla et al., 2011) and content based filtering (Pazzani and Billsus, 2007). While the former method is based on modelling the activities of users with similar behaviour in a platform, the later works on modelling the likes of an individual user in the platform. Both approaches have their own merits and are used depending on the task at hand. In this article, we explore a content based recommender system based on novel topic models. Topic modelling refers to an unsupervised learning approach that extracts topics from documents and groups the words belonging to each topic. Latent Dirichlet allocation (LDA) is one of the most famous topic models used for this purpose (Blei et al., 2003). Since the introduction of LDA a number of research ideas have been proposed to improve the vanilla model to suit different applications. For example, there is demonstration of LDA being used to model multilingual topics simultaneously (Mimno et al., 2009). This would help in tagging documents appropriately irrespective of language. Another improvement over LDA which can be used for supervised classification of images has also proved to be effective (Chong et al., 2009). LDA has also been used for creating recommendation systems (Nagori and Aghila, 2011). LDA can extract topics from the description of user activity which could help suggest new items that the user might be interested in. It is well known that models which take into account, the co-occurrences of words, tend to give a boost for topic modelling tasks (Opper and Saad, 2001). This made us to choose a design which incorporates the possibility of bigram words such as ‘Thank you’, ‘high school’, etc. Recent research has shown that using mixture models in conjunction with LDA helps in extracting better topics (Opper and Saad, 2001). In our models we will integrate this idea to improve recommendations. There has also been studies, involving the use of alternative priors other than Dirichlet for the prior of topic proportions in a document. Generalized Dirichlet (GD) and Beta-Liouville (BL) distributions have proved to be efficient substitutes for Dirichlet (Bakhtiari and Bouguila, 2016; Bakhtiari and Bouguila, 2014). Generalized Dirichlet (GD) distribution has a general covariance structure which might help to better fit the data, as compared to Dirichlet which has a negative covariance structure. The drawback of GD however, is that twice the number of parameters have to be estimated than Dirchlet distribution. BL distribution helps to overcome this drawback and also sports a general covariance matrix. Based on these theo-

Maanicshah, K., Amayri, M. and Bouguila, N.
Novel Topic Models for Content Based Recommender Systems.
DOI: 10.5220/0011826700003467
In *Proceedings of the 25th International Conference on Enterprise Information Systems (ICEIS 2023) - Volume 1*, pages 138-145
ISBN: 978-989-758-648-4; ISSN: 2184-4992

retical grounds and experimental proofs, we decided to use these distributions as priors in our model to provide a better fit to the data. Parameters estimation plays a crucial role in machine learning models. Most of the approaches on LDA based models mention variational inference and Gibbs sampling as effective methods for estimating the parameters (Blei et al., 2003; Liu et al., 2020). However, in the case of pure Bayesian approaches such us Gibbs sampling, computations are not always tractable for complex priors (Attias, 1999). Variational inference on the other hand, approximates the posterior probability instead of calculating it which gives guaranteed convergence (Hu et al., 2019). Hence, we choose variational inference as our parameter estimation method. Furthermore, as opposed to the frequently used method for inferring the variational solutions as established in (Blei et al., 2003), we employed the method used for mixture models in (Fan et al., 2012). This makes the mathematical computation of variational solutions easier. We evaluate our model, with two challenging datasets. One of them is for anime recommendation and the other is for recommendation of movies from netflix. We estimate the performance of the model based on coherence score for both datasets. In addition, since we had enough ground truth data to validate the netflix dataset, we estimate the accuracy of predictions as well. The rest of the paper is organized as follows: The description of the proposed models is given in Section 2. This is followed by the variational algorithm to estimate the parameters in Section 3. The experiments performed on the datasets with our models are detailed in Section 4. We finally conclude in Section 5 with our findings.

2 MODEL DESCRIPTION

Let us assume that we have a set of D documents in a corpus. We denote the number of words in a document $d = 1, 2, ..., D$ by N_d. For the n^{th} word among N_d words in a document d, w_{dn} can be represented as a V dimensional indicator vector, where $w_{dnv} = 1$ if the word w_{dn} is the v^{th} word in the vocabulary and 0 otherwise. We also have another latent variable $\mathcal{Z} = \{\vec{z}_{dn}\}$ which specifies, to which among the K topics the word has maximum affinity. z_{dnk} follows a similar convention that $z_{dnk} = 1$ if the word belongs to the k^{th} topic of the K different topics and 0 otherwise. The distribution of words for a topic k is given by a multinomial with parameters $\vec{\beta}_k$ which has a Dirichlet prior with parameter $\vec{\lambda}_k$. $p(\vec{\theta} \mid \Phi)$ is the distribution of the prior for the topic proportions for the documents and takes the form of GD or BL distributions with parameter Φ. In addition to this, we also have another latent variable $\mathcal{Y} = (\vec{y}_1, \vec{y}_2, ..., \vec{y}_D)$ corresponding to the mixture model which is an L dimensional one hot encoded indicator vector showing which component the document belongs to. Hence, $y_{dl} = 1$ if the document is sampled from the l^{th} component and 0 if not. $\mathcal{Y}$ in turn is sampled from a multinomial distribution with parameters $\vec{\pi} = (\pi_1, \pi_2, ..., \pi_L)$ with the constraints, $0 \leq \pi_l \leq 1$ and $\sum_{l=1}^{L} \pi_l$. According to these assumptions, for a corpus $\mathcal{W}$ containing D documents, we can write the marginal as,

$$p(\mathcal{W} \mid \vec{\pi}, \vec{\Phi}, \vec{\beta}) = \prod_{d=1}^{D} \int \Bigg[\Big(\sum_{y_d} p(\vec{\theta}_d \mid y_d, \vec{\Phi}) p(y_d \mid \vec{\pi}) \Big) \prod_{n=1}^{N_d} \sum_{z_{dn}} p(w_{dn}, w_{d(n-1)} \mid z_{dn}, \vec{\beta}) \; p(z_{dn} \mid \vec{\theta}_d) \Bigg] d\vec{\theta}_d \tag{1}$$

where $\vec{\Phi}, \vec{\beta}$ are the parameters of the prior distribution for document topic proportions and topic word proportions respectively. Here, $w_{d(n-1)} = v_{n-1}$ and $w_{dn} = v_n$ incorporates the dependency of adjacent words to the topic latent variable. Based on this general structure, we can define the priors based on GD and BL distributions as mentioned in the following subsections.

2.1 Latent Generalized Dirichlet Bi-Term Mixture Allocation (Bi-LGDMA)

In the case of Bi-LGDMA, the prior for the topic proportions is generated from a GD distribution. Let us consider a GD distribution with parameters $(\sigma_{l1}, \sigma_{l2}, ..., \sigma_{lK}, \tau_{l1}, \tau_{l2}, ..., \tau_{lK})$. The probability density function of the topic proportions can be written as,

$$p(\vec{\theta}_d \mid \vec{\sigma}_l, \vec{\tau}_l) = \prod_{k=1}^{K} \frac{\Gamma(\tau_{lk} + \sigma_{lk})}{\Gamma(\tau_{lk})\Gamma(\sigma_{lk})} \theta_{dk}^{\sigma_{lk}-1} \left(1 - \sum_{j=1}^{k} \theta_{dj}\right)^{\gamma_{lk}} \tag{2}$$

where, $\gamma_{lk} = \tau_{lk} - \tau_{l(k+1)} - \sigma_{l(k+1)}$ for $k = 1, 2, ..., K-1$ and $\gamma_{lk} = \sigma_{lk} - 1$ for $k = K$. As mentioned earlier, owing to the fact that using a mixture model over the topic proportions helps improve the model (Chien et al., 2018), we introduce mixture models as,

$$p(\vec{\theta}_d \mid \vec{y}_d, \vec{\sigma}, \vec{\tau}) = \prod_{l=1}^{L} \left[\prod_{k=1}^{K} \frac{\Gamma(\tau_{lk} + \sigma_{lk})}{\Gamma(\tau_{lk})\Gamma(\sigma_{lk})} \theta_{dk}^{\sigma_{lk}-1} \left(1 - \sum_{j=1}^{k} \theta_{dj}\right)^{\gamma_{lk}} \right]^{y_{dl}} \tag{3}$$

The latent variable $\mathcal{Y}$ is governed by a multinomial distribution with parameter $\vec{\pi}$ which is the mixing weights for the mixture model. This is denoted by $p(y_d \mid \vec{\pi}) = \prod_{l=1}^{L} \pi_l^{y_{dl}}$. Contrary to bigrams where the probability of two words occurring together is considered, we take into account that logically these bigrams end up belonging to the same topic and consider them as bi-terms associated with the same topic. This gives us,

$$p(w_{d(n-1)}, w_{dn} \mid z_{dn}, \vec{\beta}) = \prod_{k=1}^{K} \Big(\prod_{v=1}^{V} \beta_{kv}^{w_{d(n-1)(v-1)}+w_{dnv}} \Big)^{z_{dnk}} \tag{4}$$

The relation between the topic proportions and latent variable $\vec{z}_d$ is given by the multinomial $p(z_{dn} \mid \vec{\theta}_d) = \prod_{k=1}^{K} \theta_{dk}^{z_{dnk}}$. In general, it is well known that introducing conjugate prior over the undetermined parameters helps improve parameter estimation (Fan et al., 2012). However, in the case of GD the conjugate priors are intractable. Due to this reason, we introduce Gamma prior for the parameter $\vec{\sigma}$ as $p(\sigma_{lk}) = \mathcal{G}(\sigma_{lk} \mid \upsilon_{lk}, \nu_{lk}) = \frac{\nu_{lk}^{\upsilon_{lk}}}{\Gamma(\upsilon_{lk})} \sigma_{lk}^{\upsilon_{lk}-1} e^{-\nu_{lk}\sigma_{lk}}$, where $\mathcal{G}(\cdot)$ indicates a Gamma distribution. Similarly, following the same convention, the prior for $\vec{\tau}$ is given by, $p(\tau_{lk}) = \mathcal{G}(\tau_{lk} \mid s_{lk}, t_{lk})$. In addition, we also apply variational smoothing as mentioned in (Blei et al., 2003) which helps to eliminate problems that arise due to sparsity in data. Assuming a Dirichlet prior over $\vec{\beta}$ we can define,

$$p(\vec{\beta}_k \mid \vec{\lambda}_k) = \frac{\Gamma(\sum_{v=1}^{V} \lambda_{kv})}{\prod_{v=1}^{V} \Gamma(\lambda_{kv})} \prod_{v=1}^{V} \beta_{kv}^{\lambda_{kv}-1} \tag{5}$$

We assume a GD distribution over $\vec{\theta}_d$ given by the equation,

$$p(\vec{\theta}_d \mid \vec{g}_d, \vec{h}_d) = \prod_{k=1}^{K} \frac{\Gamma(g_{dk}+h_{dk})}{\Gamma(g_{dk})\Gamma(h_{dk})} \theta_{dk}^{g_{dk}-1} \Big(1 - \sum_{j=1}^{k} \theta_{dj} \Big)^{\zeta_{dk}} \tag{6}$$

where, $\zeta_{dk} = h_{dk} - g_{d(k-1)} - h_{d(k-1)}$ while $k \leq K-1$ and $\zeta_{dk} = h_{dk} - 1$ when $k = K$. This helps us in deriving the variational solutions. Thus, considering the parameters $\Theta = \{\mathcal{Z}, \vec{\beta}, \vec{\theta}, \vec{\sigma}, \vec{\tau}, \vec{y}\}$, the joint posterior distribution can be written as,

$$\begin{aligned} p(W,\Theta) = & p(\vec{W} \mid \mathcal{Z}, \vec{\beta}) p(\vec{z} \mid \vec{\theta}) p(\vec{\theta} \mid \vec{\sigma}, \vec{\tau}, \vec{y}) p(\vec{y} \mid \vec{\pi}) \\ & \times p(\vec{\theta} \mid \vec{g}, \vec{h}) p(\vec{\beta} \mid \vec{\lambda}) p(\vec{\sigma} \mid \vec{\upsilon}, \vec{\nu}) p(\vec{\tau} \mid \vec{s}, \vec{t}) \end{aligned} \tag{7}$$

2.2 Latent Beta-Liouville Bi-Term Mixture Allocation (Bi-LBLMA)

By following similar assumptions, we can construct our Bi-LBLMA model with some changes. The basic idea here is to replace the prior for topic proportions with BL distribution. Considering a BL distribution with parameters $(\mu_{l1}, \mu_{l2}, ..., \mu_{lK}, \sigma_l, \tau_l)$, we can write the prior as,

$$\begin{aligned} p(\vec{\theta}_d \mid \vec{\mu}, \vec{\sigma}, \vec{\tau}) = & \prod_{l=1}^{L} \prod_{k=1}^{K} \frac{\Gamma(\sum_{k=1}^{K} \mu_{lk})}{\prod_{k=1}^{K} \Gamma(\mu_{lk})} \frac{\Gamma(\sigma_l + \tau_l)}{\Gamma(\sigma_l)\Gamma(\tau_l)} \theta_{dk}^{\mu_{lk}-1} \\ & \times \Big[\sum_{k=1}^{K} \theta_{dk} \Big]^{\sigma_l - \sum_{k=1}^{K} \mu_{lk}} \Big[1 - \sum_{k=1}^{K} \theta_{dk} \Big]^{\tau_l - 1} \end{aligned} \tag{8}$$

Assuming Gamma priors for the parameters quoting the same reasons for Bi-LGDMA, the priors ar given by, $\mathcal{G}(\mu_{lk} \mid \upsilon_{lk}, \nu_{lk})$, $\mathcal{G}(\sigma_l \mid s_l, t_l)$ and $\mathcal{G}(\tau_l \mid \Omega_l, \Lambda_l)$ respectively. The variational distribution for the topic proportions in the case of Bi-LBLMA consequently takes the form,

$$\begin{aligned} p(\vec{\theta}_d \mid \vec{f}_d, g_d, h_d) = & \prod_{k=1}^{K} \frac{\Gamma(\sum_{k=1}^{K} f_{dk})}{\prod_{k=1}^{K} \Gamma(f_{dk})} \frac{\Gamma(g_d + h_d)}{\Gamma(g_d)\Gamma(h_d)} \theta_{dk}^{f_{dk}-1} \\ & \times \Big[\sum_{k=1}^{K} \theta_{dk} \Big]^{g_d - \sum_{k=1}^{K} f_{dk}} \\ & \times \Big[1 - \sum_{k=1}^{K} \theta_{dk} \Big]^{h_d - 1} \end{aligned} \tag{9}$$

The rest of the equations are the same as mentioned in previous subsection. Making these changes, we can write the posterior joint probability for Bi-LBLMA as,

$$\begin{aligned} p(W,\Theta) = & p(\vec{W} \mid \mathcal{Z}, \vec{\beta}) p(\vec{z} \mid \vec{\theta}) p(\vec{\theta} \mid \vec{\mu}, \vec{\sigma}, \vec{\tau}, \vec{y}) p(\vec{y} \mid \vec{\pi}) \\ & \times p(\vec{\theta} \mid \vec{f}, \vec{g}, \vec{h}) p(\vec{\beta} \mid \vec{\lambda}) p(\vec{\mu} \mid \vec{\upsilon}, \vec{\nu}) \\ & \times p(\vec{\sigma} \mid \vec{s}, \vec{t}) p(\vec{\tau} \mid \vec{\Omega}, \vec{\Lambda}) \end{aligned} \tag{10}$$

where $\Theta = \{\mathcal{Z}, \vec{\beta}, \vec{\theta}, \vec{\mu}, \vec{\sigma}, \vec{\tau}, \vec{y}\}$ represents the parameters of the model.

3 VARIATIONAL INFERENCE

Having defined the model, the next step is to estimate the parameters. In this article we use the variational method used in (Fan et al., 2012). The basic idea of variational inference is to assume a distribution $Q(\Theta)$ which is bound to be the approximation of the true posterior $p(W \mid \Theta)$ and then minimize the difference between the two distributions until they are similar. This is done by calculating the Kullback-Leibler (KL) divergence between the two distributions. The equation to find KL divergence between $Q(\Theta)$ and $p(W \mid \Theta)$ can be written as,

$$KL(Q \,||\, P) = -\int Q(\Theta) \ln \Big(\frac{p(W \mid \Theta)}{Q(\Theta)} \Big) d\Theta \tag{11}$$

We can simplify this equation as,

$$KL(Q\,||\,P) = \ln p(W) - \mathcal{L}(Q) \tag{12}$$

where, $\mathcal{L}(Q) = \int Q(\Theta) \ln \left(\frac{p(W,\Theta)}{Q(\Theta)} \right) d\Theta$ is the lower bound. Theoretically, when $KL(Q\,||\,P)$ is 0 the distributions must be identical. Hence, maximizing the lower bound $\mathcal{L}(Q)$ will minimize the value of KL divergence and consequently bring the value closer to 0. By following mean-field theory (Opper and Saad, 2001) we consider the parameters to be independent of each other since the true posterior becomes intractable otherwise. $Q(\Theta)$ can now be written as the product of the individual parameters as $Q(\Theta) = \prod_{j=1}^{J}$, with J being the total number of parameters.The optimal solution for each of the parameters can be found by calculating the expectations of all the parameters except the current parameter. This can be expressed as,

$$Q_j(\Theta_j) = \frac{\exp \langle \ln p(W,\Theta) \rangle_{\neq j}}{\int \exp \langle \ln\, p(W,\Theta) \rangle_{\neq j} d\Theta} \tag{13}$$

Once initiated with some random values, the variational solutions are updated iteratively and thus lowering the lower bound. The optimal variational solutions for all the parameters are obtained at convergence. The variational solutions for our models are given in the following subsections.

3.1 Variational Solutions for Bi-LGDMA

Calculating the variational solutions for Eq. 7 yields the following equations:

$$Q(\mathcal{Y}) = \prod_{d=1}^{D} \prod_{l=1}^{L} r_{dl}^{y_{dl}}, Q(\mathcal{Z}) = \prod_{d=1}^{D} \prod_{N=1}^{N_d} \prod_{k=1}^{K} \phi_{dnk}^{z_{dnk}} \tag{14}$$

$$Q(\vec{\sigma}) = \mathcal{G}(\vec{\sigma} \mid \vec{\upsilon}^*, \vec{\nu}^*), Q(\vec{\tau}) = \mathcal{G}(\vec{\tau} \mid \vec{s}^*, \vec{t}^*) \tag{15}$$

$$Q(\vec{\beta}) = \prod_{k=1}^{K} \prod_{v=1}^{V} \frac{\Gamma(\sum_{v=1}^{V} \lambda_{kv}^*)}{\prod_{v=1}^{V} \Gamma(\lambda_{kv}^*)} \beta_{kv}^{\lambda_{kv}^* - 1} \tag{16}$$

$$Q(\vec{\theta}) = \prod_{d=1}^{D} \prod_{k=1}^{K} \frac{\Gamma(g_{dk}^* + h_{dk}^*)}{\Gamma(g_{dk}^*)\Gamma(h_{dk}^*))} \theta_{dk}^{g_{dk}^* - 1} \left(1 - \sum_{j=1}^{k} \theta_{dj}\right)^{\zeta_{dk}^*} \tag{17}$$

where,

$$r_{dl} = \frac{\rho_{dl}}{\sum_{l=1}^{L} \rho_{dl}}, \phi_{dnk} = \frac{\delta_{dnk}}{\sum_{k=1}^{K} \delta_{dnk}}, \pi_l = \frac{1}{D} \sum_{d=1}^{D} r_{dl} \tag{18}$$

$$\rho_{dl} = \exp\left\{ \ln \pi_l + \mathcal{R}_l + \sum_{k=1}^{K} (\sigma_{lk} - 1) \langle \ln \theta_{dk} \rangle + \gamma_{lk} \left\langle 1 - \sum_{j=1}^{k} \theta_{dj} \right\rangle \right\} \tag{19}$$

$$\delta_{dnk} = exp\Big(\big[w_{d(n-1)(v-1)} + w_{dnv} \big] \langle \ln \beta_{kv} \rangle + \langle \ln \theta_{dk} \rangle \Big) \tag{20}$$

Here, $\mathcal{R}$ is the taylor series approximations of $\langle \ln \frac{\Gamma(\sigma+\tau)}{\Gamma(\sigma)\Gamma(\tau)} \rangle$ and is given by,

$$\begin{aligned}\mathcal{R} = & \ln \frac{\Gamma(\overline{\sigma} + \overline{\tau})}{\Gamma(\overline{\sigma})\Gamma(\overline{\tau})} + \overline{\sigma} \big[\Psi(\overline{\sigma} + \overline{\tau}) - \Psi(\overline{\sigma}) \big] (\langle \ln \sigma \rangle - \ln \overline{\sigma}) \\ & + \overline{\tau} \big[\Psi(\overline{\sigma} + \overline{\tau}) - \Psi(\overline{\tau}) \big] (\langle \ln \tau \rangle - \ln \overline{\tau}) \\ & + 0.5 \overline{\sigma}^2 \big[\Psi'(\overline{\sigma} + \overline{\tau}) - \Psi'(\overline{\sigma}) \big] \langle (\ln \sigma - \ln \overline{\sigma})^2 \rangle \\ & + 0.5 \overline{\tau}^2 \big[\Psi'(\overline{\sigma} + \overline{\tau}) - \Psi'(\overline{\tau}) \big] \langle (\ln \tau - \ln \overline{\tau})^2 \rangle \\ & + \overline{\sigma} \overline{\tau} \Psi'(\overline{\sigma} + \overline{\tau}) (\langle \ln \sigma \rangle - \ln \overline{\sigma}) (\langle \ln \tau \rangle - \ln \overline{\tau})\end{aligned} \tag{21}$$

$$\begin{aligned}\upsilon_{lk}^* = & \upsilon_{lk} + \sum_{d=1}^{D} \langle y_{dl} \rangle \Big[\Psi(\overline{\sigma}_{lk} + \overline{\tau}_{lk}) - \Psi(\overline{\sigma}_{lk}) \\ & + \overline{\tau}_{lk} \Psi'(\overline{\sigma}_{lk} + \overline{\tau}_{lk}) (\langle \ln \tau_{lk} \rangle - \ln \overline{\tau}_{lk}) \Big] \overline{\sigma}_{lk}\end{aligned} \tag{22}$$

$$\begin{aligned}s_{lk}^* = & s_{lk} + \sum_{d=1}^{D} \langle y_{dl} \rangle \Big[\Psi(\overline{\tau}_{lk} + \overline{\sigma}_{lk}) - \Psi(\overline{\tau}_{lk}) \\ & + \overline{\sigma}_{lk} \Psi'(\overline{\tau}_{lk} + \overline{\sigma}_{lk}) (\langle \ln \sigma_{lk} \rangle - \ln \overline{\sigma}_{lk}) \Big] \overline{\tau}_{lk}\end{aligned} \tag{23}$$

$$\nu_{lk}^* = \nu_{lk} - \sum_{d=1}^{D} \langle y_{dl} \rangle \langle \ln \theta_{dk} \rangle \tag{24}$$

$$t_{lk}^* = t_{lk} - \sum_{d=1}^{D} \langle y_{dl} \rangle \left\langle \ln \left[1 - \sum_{j=1}^{K} \theta_{dj} \right] \right\rangle \tag{25}$$

$$g_{dk}^* = g_{dk} + \sum_{n=1}^{N_d} \langle z_{dnk} \rangle + \sum_{l=1}^{L} \langle y_{dl} \rangle \sigma_{lk} \tag{26}$$

$$h_{dk}^* = h_{dk} + \sum_{l=1}^{L} \langle y_{dl} \rangle \tau_{lk} + \sum_{kk=k+1}^{K} \phi_{dn(kk)} \tag{27}$$

$$\lambda_{kv}^* = \lambda_{kv} + \sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{v=1}^{V} \phi_{dnk} \left[w_{d(n-1)v} + w_{dnv} \right] \tag{28}$$

$$\pi_l = \frac{1}{D} \sum_{d=1}^{D} r_{dl} \tag{29}$$

In the above equations, $\langle \cdot \rangle$ indicates the expectation of the variable, whose values are detailed in (Maanicshah et al., 2023). We calculate equations 14 - 17 by repetitively updating the parameters until there is no considerable change in the lower bound estimates. At this point of convergence, we will have the optimal values for the variational solutions.

3.2 Variational Solutions for LBLMA

Similar to the previous section, we can derive the following variational solutions for Eq. 10. The only difference is the apparent change in $Q(\vec{\theta})$ and some definitions of related variables. The variational solutions are hence given by,

$$Q(\mathcal{Y}) = \prod_{d=1}^{D}\prod_{l=1}^{L} r_{dl}^{y_{dl}}, Q(\mathcal{Z}) = \prod_{d=1}^{D}\prod_{N=1}^{N_d}\prod_{k=1}^{K} \phi_{dnk}^{z_{dnk}} \quad (30)$$

$$Q(\vec{\mu}) = \mathcal{G}(\vec{\mu}^* \mid \vec{\upsilon}*, \vec{\nu}^*), Q(\sigma_l) = \mathcal{G}(\sigma_l \mid s_l, t_l) \quad (31)$$

$$Q(\tau_l)=\mathcal{G}(\tau_l \mid \Omega_l, \Lambda_l), Q(\vec{\beta})=\prod_{k=1}^{K}\prod_{v=1}^{V} \frac{\Gamma(\sum_{v=1}^{V}\lambda_{kv}^*)}{\prod_{v=1}^{V}\Gamma(\lambda_{kv}^*)} \beta_{kv}^{\lambda_{kv}^*-1} \quad (32)$$

$$Q(\vec{\theta}) = \prod_{d=1}^{D}\prod_{k=1}^{K} \frac{\Gamma(\sum_{k=1}^{K} f_{dk}^*)}{\Gamma(f_{dk}^*)} \frac{\Gamma(g_d^* + h_d^*)}{\Gamma(g_d^*)\Gamma(h_d^*)} \theta_{dk}^{f_{dk}^*-1} \times \Big[\sum_{k=1}^{K}\theta_{dk}\Big]^{g_d^* - \sum_{k=1}^{K} f_{dk}^*} \Big[1 - \sum_{k=1}^{K}\theta_{dk}\Big]^{h_d^*-1} \quad (33)$$

where,

$$r_{dl} = \frac{\rho_{dl}}{\sum_{l=1}^{L}\rho_{dl}}, \phi_{dnk} = \frac{\delta_{dnk}}{\sum_{k=1}^{K}\delta_{dnk}}, \pi_l = \frac{1}{D}\sum_{d=1}^{D} r_{dl} \quad (34)$$

$$\rho_{dl} = exp\Bigg\{ \ln\pi_l + \mathcal{R}_l + \mathcal{S}_l + (\mu_{lk} - 1)\big\langle \ln\theta_{dk} \big\rangle + \Big(\sigma_l - \sum_{k=1}^{K}\mu_{lk}\Big)\Big\langle \ln\Big[\sum_{k=1}^{K}\theta_{dk}\Big]\Big\rangle + (\tau_l - 1)\Big\langle \ln\Big[1 - \sum_{k=1}^{K}\theta_{dk}\Big]\Big\rangle \Bigg\} \quad (35)$$

Due to intractability, we use Taylor series expansions for $\big\langle \frac{\Gamma(\sum_{k=1}^{K}\sigma_{lk})}{\Gamma(\sigma_{lk})} \big\rangle$ and $\big\langle \ln\frac{\Gamma(\sigma+\tau)}{\Gamma(\sigma)\Gamma(\tau)} \big\rangle$ denoted by $\mathcal{R}$ and $\mathcal{S}$ respectively. The approximations are given as,

$$\begin{aligned}\mathcal{R}_l = &\ln\frac{\Gamma(\sum_{k=1}^{K}\mu_{lk})}{\prod_{k=1}^{K}\Gamma(\mu_{lk})} \\ &+ \sum_{k=1}^{K}\bar{\mu}_{lk}\Big[\Psi\Big(\sum_{k=1}^{K}\bar{\mu}_{lk}\Big) - \Psi(\bar{\mu}_{lk})\Big]\big[\big\langle \ln\mu_{lk}\big\rangle - \ln\bar{\mu}_{lk}\big] \\ &+ \frac{1}{2}\sum_{k=1}^{K}\bar{\mu}_{lk}^2\Big[\Psi'\big(\sum_{k=1}^{K}\bar{\mu}_{lk}\big) - \Psi'(\bar{\mu}_{lk})\Big] \\ &\times \big\langle (\ln\mu_{lk} - \ln\bar{\mu}_{lk})^2\big\rangle + \frac{1}{2}\sum_{a=1}^{K}\sum_{b=1,a\neq b}^{K}\bar{\mu}_{la}\bar{\mu}_{lb} \\ &\times \Big[\Psi'\Big(\sum_{k=1}^{K}\bar{\mu}_{lk}\Big)\big(\big\langle\ln\mu_{la}\big\rangle - \ln\bar{\mu}_{la}\big)\big(\big\langle\ln\mu_{lb}\big\rangle - \ln\bar{\mu}_{lb}\big)\Big]\end{aligned}$$

$$\begin{aligned}\mathcal{S} = &\ln\frac{\Gamma(\bar{\sigma}+\bar{\tau})}{\Gamma(\bar{\sigma})\Gamma(\bar{\tau})} + \bar{\sigma}\big[\Psi(\bar{\sigma}+\bar{\tau}) - \Psi(\bar{\sigma})\big](\big\langle\ln\sigma\big\rangle - \ln\bar{\sigma}) \\ &+ \bar{\tau}\big[\Psi(\bar{\sigma}+\bar{\tau}) - \Psi(\bar{\tau})\big](\big\langle\ln\tau\big\rangle - \ln\bar{\tau}) \\ &+ 0.5\bar{\sigma}^2\big[\Psi'(\bar{\sigma}+\bar{\tau}) - \Psi'(\bar{\sigma})\big]\big\langle(\ln\sigma - \ln\bar{\sigma})^2\big\rangle \\ &+ 0.5\bar{\tau}^2\big[\Psi'(\bar{\sigma}+\bar{\tau}) - \Psi'(\bar{\tau})\big]\big\langle(\ln\tau - \ln\bar{\tau})^2\big\rangle \\ &+ \bar{\sigma}\bar{\tau}\Psi'(\bar{\sigma}+\bar{\tau})(\big\langle\ln\sigma\big\rangle - \ln\bar{\sigma})(\big\langle\ln\tau\big\rangle - \ln\bar{\tau})\end{aligned} \quad (36)$$

$$\begin{aligned}\upsilon_{lk}^* = &\upsilon_{lk} + \sum_{d=1}^{D}\langle y_{dl}\rangle\bar{\mu}_{lk}\Big[\Psi\Big(\sum_{k=1}^{K}\bar{\mu}_{lk}\Big) - \Psi(\bar{\mu}_{lk}) \\ &+ \Psi\Big(\sum_{k=1}^{K}\Big)\sum_{a\neq k}^{K}(\big\langle\ln\mu_{la}\big\rangle - \ln\bar{\mu}_{la})\bar{\mu}_{la}\Big]\end{aligned} \quad (37)$$

$$\nu_{lk}^* = \nu_{lk} - \sum_{d=1}^{D}\langle y_{dl}\rangle\Big[\big\langle\ln\theta_{dk}\big\rangle - \Big\langle\ln\sum_{k=1}^{K}\theta_{dk}\Big\rangle\Big] \quad (38)$$

$$\begin{aligned}s_l^* = &s_l + \sum_{d=1}^{D}\langle y_{dl}\rangle\Bigg[\Psi(\bar{\sigma}_l + \bar{\tau}_l) - \Psi(\bar{\sigma}_l) \\ &+ \bar{\tau}_l\Psi'(\bar{\sigma}_l + \bar{\tau}_l)(\big\langle\ln\tau_l\big\rangle - \ln\bar{\tau}_l)\Bigg]\bar{\sigma}_l\end{aligned} \quad (39)$$

$$t_l^* = t_l - \sum_{d=1}^{D}\langle y_{dl}\rangle\Big\langle\ln\Big[\sum_{k=1}^{K}\theta_{dk}\Big]\Big\rangle \quad (40)$$

$$\begin{aligned}\Omega_l^* = &\Omega_{lk} + \sum_{d=1}^{D}\langle y_{dl}\rangle\Bigg[\Psi(\bar{\tau}_l + \bar{\sigma}_l) - \Psi(\bar{\tau}_l) \\ &+ \bar{\sigma}_l\Psi'(\bar{\tau}_l + \bar{\sigma}_l)(\big\langle\ln\sigma_l\big\rangle - \ln\bar{\sigma}_l)\Bigg]\bar{\tau}_l\end{aligned} \quad (41)$$

$$\Lambda_l^* = \Lambda_l - \sum_{d=1}^{D}\langle y_{dl}\rangle\Big\langle\ln\Big[1 - \sum_{k=1}^{K}\theta_{dk}\Big]\Big\rangle \quad (42)$$

$$f_{dk}^{*} = f_{dk} + \sum_{n=1}^{N_d} \langle z_{dnk} \rangle + \sum_{l=1}^{L} \langle y_{dl} \rangle \mu_{lk} \tag{43}$$

$$g_{d}^{*} = g_{d} + \sum_{n=1}^{N_d} \sum_{k=1}^{K} \langle z_{dnk} \rangle + \sum_{l=1}^{L} \langle y_{dl} \rangle \sigma_{l} \tag{44}$$

$$h_{d}^{*} = h_{d} + \sum_{l=1}^{L} \langle y_{dl} \rangle \tau_{l} \tag{45}$$

The expectations in these equations are defined with respect to BL distribution in (Maanicshah et al., 2022). Similar to Bi-LBLMA, we calculate equations 30 - 33 repeatedly until convergence to find the optimal solutions.

4 EXPERIMENTAL RESULTS

To evaluate the performance of our model, we build a system for anime recommendation based on a dataset in Kaggle containing information about anime[1] and another for recommending movies based on data from netflix prize data [2]. We compare our model with widely used LDA and examine how our models weigh up against unmodified latent generalized Dirichlet allocation (LGDA) and latent Beta-Liouville allocation (LBLA) models. The idea of our recommendation system is that we find the Euclidean distance between the document topic proportions ϕ_{dk} of the query document and the rest of the documents. We can then find the top N recommendations for that query. The following subsections detail our experiments for the two datasets.

4.1 Anime Recommendation

This dataset consisted of 2 files containing information about anime, reviews of users and user profile details. The anime file had around 16K anime details like, title, synopsis, genre, airing date, etc. The profiles file had details of users and the anime they have added as favourites. The reviews file has information on the reviews the user has written for different anime. All these data has been extracted from https://myanimelist.net. From the anime details file, the data that helps for content based recommendation is mainly the synopsis. However, the synopsis was not available for some of the anime within the data. Hence we used the myanimelist API to extract missing synopsis. There were cases in which some of the titles refer to a parent anime and the description of parent anime was taken in these cases. We ignore anime where the synopsis is too short. After applying these constraints we were left with around 1126 anime to use for our content based recommendation system. In the case of this dataset, there were a very few user profiles who had more than 20 anime in their favourites list which was not enough to evaluate our models. To understand the relevance of the topics that have been extracted by our model, we calculated UMass coherence score(Mimno et al., 2011) which takes into account, the probability of two words within a topic occurring in the corpus. It is given by,

$$score_{UMass}(k) = \sum_{i=2}^{M_k} \sum_{j=1}^{M_k - 1} \log \frac{p(w_i, w_j) + 1}{p(w_i)} \tag{46}$$

U_k in the above equation indicates the number of top words taken into consideration for the topic. In our case this value is 10. The equation basically calculates the relevancy of words within a topic by finding the ratio of the probability of two words w_i and w_j occurring together to the probability of word w_i for which the score is being calculated. Table 1 shows the coherence scores of topics derived from LDA, latent generalized Dirichlet allocation (LGDA), latent Beta-Liouville allocation (LBLA) and Bi-LGDMA and Bi-LBLMA for different values of L. It can be seen that using a GD and BL prior helps in obtaining better topics with a higher coherence score. Bi-LBLMA performs better than Bi-LGDMA according to our experiments, which is due to the fact that choosing the parameters for Bi-LGDMA is a little harder than Bi-LBLMA. We calculated the coherence scores for different values of K to find the correct number of topics for the model. The best results were observed when K was set to 5 as observed in Figure 1. Both Bi-LGDMA and Bi-LBLMA performed well when $L = 3$. In the case of Bi-LBLMA we see that the coherence is very close when $L = 3$ and $L = 4$. In these situations choosing the L as 3 or 4 will give similar recommendations. This being a quantitative assessment of the model, to qualitatively see how the model performs, Table 1 and 2 shows few of the top ten suggestions for a query anime for the two models.

'Bleach' is an anime based on travelling between worlds through portals in the action genre. The anime suggested by Bi-LGDMA aligns with this concept of inter-dimensional portals and magic. Similarly, the test query for Bi-LBLMA was an anime called 'Dragon Ball' which involves super-human fighting. It is interesting to see that our model identified the sequel to the original anime followed by a few other anime like 'Boku no Hero Academia' which also falls

[1] https://www.kaggle.com/datasets/marlesson/myanimelist-dataset-animes-profiles-reviews

[2] https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data

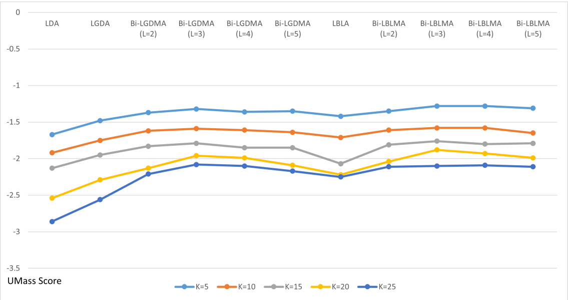

Figure 1: Coherence score for anime dataset for different values of K and L.

Table 1: Query results for Anime data with Bi-LGDMA.

S. No.	Bleach
1	Fullmetal Alchemist
2	Rosario to Vampire
3	World Trigger
4	FLCL
5	Tenjou Tenge

under the same category.

4.2 Netflix Movie Recommendation

The Netflix dataset is bigger compared to the anime datset. The dataset consists of details pertaining to ratings fo different users for around 17000 movies released before the year 2006. However, the problem with this dataset is that the synopsis of movies were not available. Hence, we scraped the data from wikipedia pages to get this details and then used it for content based recommendation. We selected the movies released after 2000 so that we are aware of them to test qualitatively. This gave us around 4000 movies with description. From the user details, we consider that a user likes a movie when they rate it as 4 or 5. We selected users who had liked at least 300 movies. This left us with 900 users as ground truth. These conditions are only to quantitatively access our models and can be ignored in realtime applications. When queried with a movie that an user likes, if one of the top N recommendations by our model is present in the list of movies liked by that user, then we consider it as a hit. By using this logic, we can calculate the accuracy of our model by calculating the ratio of total number of hits to total number of queries. We also calculate the coherence score of our topics as in the previous subsection which is graphed in Figure 2. We can see that both our models perform the best when $L = 2$ and $K = 5$. The performance improvement achieved by our models compared to the widely used LDA model proves the efficiency of our model to represent the topics better. In addition to these analysis Table 3 shows the accuracy of different models. Though both Bi-LGDMA and Bi-LBLA give comparatively better accuracy for our model, the improvement for Bi-LGDMA is not that much when compared to Bi-LBLA. Similar to the last experiment, we also check the quality of recommendations for two sample queries. This is shown in Table 4 and 5. We can see that Bi-LGDMA recommends a set of teenage and kids action movies like ‘Agent Cody Banks’ when queried with the movie ‘The Pacifier’ which is a kids action comedy. In the case of Bi-LGDMA ‘Resident Evil’ is a zombie movie where the virus causes the people to attack the non-infected people. The recommendations from our model found similar plot lines like ‘Dawn of the dead’, ‘Sasquatch’, etc which are movies based on virus outbreak, hunted by animals and so on.

Table 2: Query results for Anime data with Bi-LBLMA.

S. No.	Dragon Ball
1	Dragon Ball Z
2	Dragon Ball Super Movie: Broly
3	Boku no Hero Academia
4	Yu-Gi-Oh Duel Monsters
5	Fate/stay night

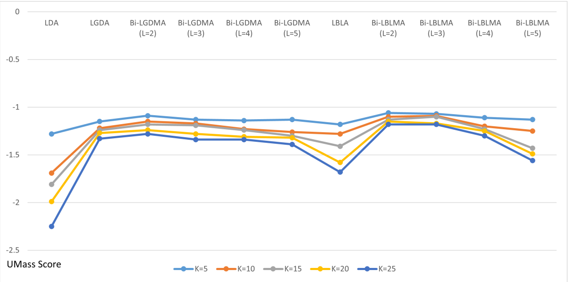

Figure 2: Coherence score for anime dataset for different values of K and L.

Table 3: Accuracy of recommendation at $N = 15$ for Netflix Data.

Model	Accuracy
LDA	85.59
LGDA	84.40
Bi-LGDMA	86.00
LBLA	86.50
Bi-LBLMA	87.36

Table 4: Query results for Netflix data.

S. No.	The Pacifier
1	Agent Cody Banks
2	Agent Cody Banks 2: Destination London
3	Lilo and Stitch 2
4	101 Dalmations II: Patch's London Adventure
5	Mean Creek

Table 5: Query results for Netflix data with Bi-LBLMA.

S. No.	Resident Evil
1	Dawn of the Dead
2	Sasquatch
3	Wrong Turn
4	Evil Remains
5	Dead Birds

5 CONCLUSION

We have introduced two novel models for topic modelling and applied it for recommendation tasks. The models are found to be effective when compared to widely used models such as LDA. From the example queries, we see that our models are able to deliver promising suggestions that the user might like. The improvement achieved by using GD and BL distributions is also clearly seen. Using biterms in conjunction with our models tend to improve the results considerably. Especially, the Bi-LBLMA model proves to be a good alternative to LDA based on the results from both the experiments.

REFERENCES

Attias, H. (1999). A variational baysian framework for graphical models. In Solla, S., Leen, T., and Müller, K., editors, *Advances in Neural Information Processing Systems*, volume 12, Cambridge, Masschusetts. MIT Press.

Bakhtiari, A. S. and Bouguila, N. (2014). Online learning for two novel latent topic models. In Linawati, Mahendra, M. S., Neuhold, E. J., Tjoa, A. M., and You, I., editors, *Information and Communication Technology*, pages 286–295, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

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

Bobadilla, J., Hernando, A., Ortega, F., and Bernal, J. (2011). A framework for collaborative filtering recommender systems. *Expert Systems with Applications*, 38(12):14609–14623.

Chien, J.-T., Lee, C.-H., and Tan, Z.-H. (2018). Latent dirichlet mixture model. *Neurocomputing*, 278:12–22. Recent Advances in Machine Learning for Non-Gaussian Data Processing.

Chong, W., Blei, D., and Li, F.-F. (2009). Simultaneous image classification and annotation. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1903–1910.

Fan, W., Bouguila, N., and Ziou, D. (2012). Variational learning for finite dirichlet mixture models and applications. *IEEE transactions on neural networks and learning systems*, 23(5):762–774.

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

Liu, Y., Du, F., Sun, J., and Jiang, Y. (2020). ilda: An interactive latent dirichlet allocation model to improve topic quality. *Journal of Information Science*, 46(1):23–40.

Maanicshah, K., Amayri, M., and Bouguila, N. (2022). Improving topic quality with interactive beta-liouville mixture allocation model. In *2022 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1143–1148.

Maanicshah, K., Amayri, M., and Bouguila, N. (2023). Interactive generalized dirichlet mixture allocation model. In *Structural, Syntactic, and Statistical Pattern Recognition: Joint IAPR International Workshops, S+SSPR 2022, Montreal, QC, Canada, August 26–27, 2022, Proceedings*, pages 33–42. Springer.

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., and McCallum, A. (2009). Polylingual topic models. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, page 880–889, USA. Association for Computational Linguistics.

Mimno, D., Wallach, H. M., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Proceedings of the Conference on Empirical Methods in Natural Language Processing; EMNLP '11, page 262–272, USA. Association for Computational Linguistics.

Nagori, R. and Aghila, G. (2011). Lda based integrated document recommendation model for e-learning systems. In *2011 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC)*, pages 230–233.

Opper, M. and Saad, D. (2001). *Advanced mean field methods: Theory and practice*. MIT press, Cambridge, Masschusetts.

Pazzani, M. J. and Billsus, D. (2007). *Content-Based Recommendation Systems*, pages 325–341. Springer Berlin Heidelberg, Berlin, Heidelberg.