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-occurrences 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 (Minno 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 (Bakhtiar and Bouguila, 2016; Bakhtiar 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 Dirichlet distribution. BL distribution helps to overcome this drawback and also sports a general covariance matrix. Based on these theo-
theoretical 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 as 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 \(w_{dny} = 1 \) if the word \(w_{dn} \) is the \(v^{th} \) word in the vocabulary and 0 otherwise. We also have another latent variable \(\zeta = \{z_{dn}\} \) 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 \(D \) different topics and 0 otherwise. The distribution of words for a topic \(k \) is given by a multinomial with parameters \(\tilde{\beta}_k \) which has a Dirichlet prior with parameter \(\tilde{\alpha}_k \). \(p(\tilde{\theta} | \Phi) \) is the distribution of the prior for the topic proportions for the documents and takes the form of GD or BL distributions with parameter \(\Phi \). In addition to this, we also have another latent variable \(Y = (y_1, y_2, ..., y_D) \) 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. \(Y \) in turn is sampled from a multinomial with parameters \(\pi = (\pi_1, \pi_2, ..., \pi_L) \) with the constraints, \(0 \leq \pi_l \leq 1 \) and \(\sum_{l=1}^{L} \pi_l \). According to these assumptions, for a corpus \(W \) containing \(D \) documents, we can write the marginal as,

\[
p(W | \pi, \Phi, \tilde{\beta}) = \prod_{d=1}^{D} \left[ \left( \sum_y \int \left( \sum_{\tilde{\theta}_d} p(\tilde{\theta}_d | y_d, \Phi) p(y_d | \pi) \right) \right) \prod_{n=1}^{N_d} \sum_{z_{dn}} \cdot p(w_{dn} | w_{d(n-1)} | z_{dn}, \tilde{\beta}) p(z_{dn} | \tilde{\theta}_d) \right]
\]

where \(\Phi, \tilde{\beta} \) 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 \((\sigma_1, \sigma_2, ..., \sigma_K, \tau_1, \tau_2, ..., \tau_K) \). The probability density function of the topic proportions can be written as,

\[
p(\tilde{\theta}_d | \sigma, \zeta) = \prod_{k=1}^{K} \frac{\Gamma(\tau_k + \sigma_k)}{\Gamma(\tau_k) \Gamma(\sigma_k)} \theta_{dk}^{\tau_k - 1} \left(1 - \sum_{j=1}^{K} \theta_{dj}\right)^{\gamma_{dk}}
\]

where \(\gamma_{dk} = \tau_k - \tau_{(k+1)} - \sigma_{(k+1)} \) for \(k = 1, 2, ..., K - 1 \) and \(\gamma_{dk} = \sigma_{k} - 1 \) 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,

\[
p(\tilde{\theta}_d | y_d, \tilde{\sigma}, \tilde{\tau}) = \prod_{l=1}^{L} \left[ \sum_{k=1}^{K} \frac{\Gamma(\tau_k + \sigma_k)}{\Gamma(\tau_k) \Gamma(\sigma_k)} \theta_{dk}^{\tau_k - 1} \left(1 - \sum_{j=1}^{K} \theta_{dj}\right)^{\gamma_{dk}} \right]^{\gamma_{wl}}
\]
The latent variable $Y$ is governed by a multinomial distribution with parameter $\pi$ which is the mixing weights for the mixture model. This is denoted by $p(y_i | \pi) = \prod_{j=1}^V \pi_{ij}$. 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,

$$p(w_{di(n-1)}, w_{dn} | z_{dn}, 2) = \prod_{k=1}^{K} \left( \prod_{i=1}^{V} \beta_{ik}^{y_{di(n-1)}+y_{dn}} \right)$$

(4)

The relation between the topic proportions and latent variable $z_{di}$ is given by the multinomial $p(z_{dn} | \theta_d) = \prod_{k=1}^{K} \phi_{dk}^{z_{dk}}$. In general, it is well known that introducing conjugate prior over the underdetermined 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 $\sigma$ as $p(\sigma_k) = G(\sigma_k | \nu_k, \lambda_k) = \frac{\nu_k^{\nu_k}}{\Gamma(\nu_k)} \sigma_k^{\nu_k-1} e^{-\nu_k / \sigma_k}$, where $G(\cdot)$ indicates a Gamma distribution. Similarly, following the same convention, the prior for $\tau$ is given by, $p(\tau_k) = G(\tau_k | s_k, t_k)$. 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 $\beta$ we can define,

$$p(\theta_k | \tilde{\alpha}_k) = \frac{\prod_{i=1}^{V} \lambda_{ik}^{\nu_i}}{\Gamma(\tilde{\alpha}_k)} \prod_{i=1}^{V} \beta_{ik}^{\nu_i-1}$$

(5)

We assume a GD distribution over $\theta_d$ given by the equation,

$$p(\theta_d | g_d, h_d) = \prod_{k=1}^{K} \frac{\Gamma(h_d + \theta_d) \theta_d^{h_d}}{\Gamma(\theta_d)} \left( 1 - \frac{1}{\sum_{k=1}^{K} \theta_d} \right)^{\zeta_{dk}}$$

(6)

where, $\zeta_{dk} = h_d - g_d + 1$ while $K - 1$ and $\zeta_{dk} = h_d - g_d - 1$ when $k = K$. This helps us in deriving the variational solutions. Thus, considering the parameters $\Theta = \{Z, \tilde{\beta}, \tilde{\alpha}, \tilde{\sigma}, \tilde{\tau}, \tilde{y}\}$, the joint posterior distribution can be written as,

$$p(W, \Theta) = p(W | Z, \tilde{\beta}) p(\tilde{\beta} | \tilde{\alpha}) p(\tilde{\alpha} | \tilde{\beta}, \tilde{\sigma}, \tilde{\tau}) p(\tilde{\tau} | \tilde{\sigma})$$

(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 $(\mu_1, \mu_2, ..., \mu_K, \sigma_1, \tau_1)$, we can write the prior as,

$$p(\theta_d, \tilde{\mu}, \tilde{\alpha}, \tilde{\tau}) = \prod_{k=1}^{K} \frac{\Gamma(\mu_k) \Gamma(\sigma_k + \tau_k)}{\Gamma(\sigma_k) \Gamma(\tau_k)} \theta_d^{\mu_k}$$

(8)

Assuming Gamma priors for the parameters quoting the same reasons for Bi-LGDMA, the priors are given by, $G(\mu_k | \nu_k, \lambda_k), G(\sigma_k | s_k, t_k)$ and $G(\tau_k | s_k, t_k)$ respectively. The variational distribution for the topic proportions in the case of Bi-LBLMA consequently takes the form,

$$p(\phi_{dk} | f_d, g_d, h_d) = \prod_{k=1}^{K} \frac{\Gamma(\sum_{l=1}^{K} f_{dl}) \Gamma(g_d + h_d)}{\Gamma(\phi_{dk}) \Gamma(h_d) \Gamma(f_{dk})} \phi_{dk}^{f_{dl}}$$

(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,

$$p(W, \theta) = p(W | Z, \tilde{\beta}) p(\tilde{\alpha} | \tilde{\beta}, \tilde{\sigma}, \tilde{\tau}) p(\tilde{\sigma} | \tilde{\beta}, \tilde{\tau}) p(\tilde{\tau} | \tilde{\beta}, \tilde{\sigma})$$

(10)

3 VARIATIONAL INFERENCE

Having defined the model, the next step is to estimate the parameters. In this work 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 $p(W | \Theta)$ 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 $p(W | \Theta)$ can be written as,

$$KL(Q || P) = - \int Q(\Theta) \ln \left( \frac{p(W | \Theta)}{Q(\Theta)} \right) d\Theta$$

(11)
We can simplify this equation as,
\[ KL(Q \parallel P) = \ln p(W) - L(Q) \]  
(12)
where, \( L(Q) = \int Q(\Theta) \ln \left( \frac{p(W)}{Q(\Theta)} \right) d\Theta \) is the lower bound. Theoretically, when \( KL(Q \parallel P) = 0 \) the distributions must be identical. Hence, maximizing the lower bound \( L(Q) \) 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,
\[ Q_j(\Theta_j) = \frac{\exp \langle \ln p(W, \Theta_j) \rangle_{\neq j}}{\int \exp \langle \ln p(W, \Theta_j) \rangle_{\neq j} d\Theta} \]
(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:
\[ Q(\gamma') = \prod_{d=1}^{D} \prod_{l=1}^{L} \phi_{dl}^{\gamma'_l}, Q(\zeta) = \prod_{d=1}^{D} \prod_{l=1}^{L} \prod_{k=1}^{K} \phi_{dk}^{\zeta_{lk}} \]
(14)
\[ Q(\beta) = \prod_{k=1}^{K} \prod_{v=1}^{V} \frac{\Gamma(\sum_{j=1}^{J} l_{jk} \beta_{kv})}{\Gamma(\beta_{kv})} \beta_{kv}^{l_{jk}-1} \]
(15)
\[ Q(\phi) = \prod_{d=1}^{D} \prod_{l=1}^{L} \prod_{k=1}^{K} \frac{\Gamma(g'_{dk} + h_{dk})}{\Gamma(g'_{dk}) \Gamma(h_{dk})} \phi_{dk}^{\gamma'_{lk} - 1} \left( 1 - \sum_{j=1}^{J} \phi_{dj} \right) \]
(16)
where,
\[ r_{dl} = \frac{\rho_{dl}}{\sum_{l=1}^{L} \rho_{dl}}, \phi_{dk} = \frac{\delta_{dk}}{\sum_{k=1}^{K} \delta_{dk}}, \pi_{l} = \frac{1}{D} \sum_{d=1}^{D} r_{dl} \]
(17)
\[ \rho_{dl} = \exp \left\{ \ln \pi_{l} + R + \sum_{k=1}^{K} (\sigma_{lk} - 1) (\ln \theta_{dk}) + \gamma_{dk} \left( 1 - \sum_{j=1}^{J} \theta_{dj} \right) \right\} \]
(18)
\[ \delta_{dk} = \exp \left\{ \left[ w_{d(n-1)v} + w_{dmv} \right] (\ln \beta_{kv}) + (\ln \theta_{dk}) \right\} \]
(19)
Here, \( R \) is the taylor series approximations of \( \langle \ln \frac{1}{\Gamma(\sigma + \tau)} \rangle \) and is given by,
\[ R = \ln \frac{\Gamma(\sigma + \tau)}{\Gamma(\sigma) \Gamma(\tau)} + \tau [\Psi(\sigma + \tau) - \Psi(\sigma)] - [\ln \tau - \ln \sigma] \]
(20)
In the above equations, \( \langle \cdot \rangle \) indicates the expectation of the variable, whose values are detailed in (Maaniscalah 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 \( Q(\bar{\theta}) \) and some definitions of related variables. The variational solutions are hence given by,

\[
Q(Y) = \prod_{d=1}^{D} \prod_{l=1}^{L} \prod_{i=1}^{N_d} \prod_{k=1}^{K} \phi_{dk}^{\gamma_{lk}} (30)
\]

\[
Q(\bar{\mu}) = G(\bar{\mu}^* | \bar{\mu}^*, \bar{s}^*), Q(\sigma_l) = G(\sigma_l | s_l, t_l) (31)
\]

\[
Q(\tau_l) = G(\tau_l | \Omega_l, \Lambda_l), Q(\bar{\beta}) = \prod_{k=1}^{K} \prod_{l=1}^{L} \prod_{i=1}^{N_d} \prod_{k=1}^{K} \phi_{dk}^{\Sigma_{i=1}^{V} \lambda_{ik}} \rho_{\Sigma_{i=1}^{V} \lambda_{ik}}^{-1} \times \left[ \sum_{k=1}^{K} \theta_{dk} \right] g_{\Sigma_{i=1}^{V} \lambda_{ik}} \left[ 1 - \sum_{k=1}^{K} \theta_{dk} \right] h_{\Sigma_{i=1}^{V} \lambda_{ik}}^{-1} (32)
\]

where,

\[
r_{dl} = \frac{\rho_{dl}}{\sum_{l=1}^{L} \rho_{dl}}, \phi_{dk} = \frac{\delta_{dk}}{\sum_{k=1}^{K} \delta_{dk}}, \pi_i = \frac{1}{D} \sum_{d=1}^{D} r_{dl} (33)
\]

\[
\rho_{dl} = \exp \left\{ \ln \frac{\Gamma(\Sigma_{i=1}^{V} \lambda_{ik})}{\prod_{i=1}^{V} \Gamma(\lambda_{ik})} + (\sigma_l - \sum_{k=1}^{K} \mu_{ik} - 1) \langle \ln \theta_{dk} \rangle + \left( \sigma_l - \sum_{k=1}^{K} \mu_{ik} \right) \left[ \ln \left( \sum_{k=1}^{K} \theta_{dk} \right) \right] + (\tau_l - 1) \left[ \ln \left( 1 - \sum_{k=1}^{K} \theta_{dk} \right) \right] \right\} (34)
\]

Due to intractability, we use Taylor series expansions for \( \langle \frac{\Gamma(\Sigma_{i=1}^{V} \lambda_{ik})}{\prod_{i=1}^{V} \Gamma(\lambda_{ik})} \rangle \) and \( \langle \ln \frac{\Gamma(\sigma_l + \tau_l)}{\Gamma(\sigma_l)} \rangle \) denoted by \( R \) and \( S \) respectively. The approximations are given as,

\[
\bar{\mu}_{ik} = \mu_{ik} + \sum_{d=1}^{D} \langle \gamma_{ld} \rangle \bar{\pi}_{ik} \left( \langle \ln \theta_{dk} \rangle - \langle \ln \left( \sum_{k=1}^{K} \theta_{dk} \right) \rangle \right) + \left( \sigma_l - \sum_{k=1}^{K} \mu_{ik} \right) \left[ \ln \left( \sum_{k=1}^{K} \theta_{dk} \right) \right] + (\tau_l - 1) \left[ \ln \left( 1 - \sum_{k=1}^{K} \theta_{dk} \right) \right] (35)
\]

\[
\bar{\beta}_{lk} = \beta_{lk} + \frac{1}{2} \sum_{d=1}^{D} \langle \gamma_{ld} \rangle \sigma_{lk} \left( \langle \ln \theta_{dk} \rangle - \langle \ln \left( \sum_{k=1}^{K} \theta_{dk} \right) \rangle \right) + \left( \sigma_l - \sum_{k=1}^{K} \mu_{ik} \right) \left[ \ln \left( \sum_{k=1}^{K} \theta_{dk} \right) \right] + (\tau_l - 1) \left[ \ln \left( 1 - \sum_{k=1}^{K} \theta_{dk} \right) \right] (36)
\]

\[
\bar{\lambda}_{ik} = \lambda_{ik} + \frac{1}{2} \sum_{d=1}^{D} \langle \gamma_{ld} \rangle \sigma_{lk} \left( \langle \ln \theta_{dk} \rangle - \langle \ln \left( \sum_{k=1}^{K} \theta_{dk} \right) \rangle \right) + \left( \sigma_l - \sum_{k=1}^{K} \mu_{ik} \right) \left[ \ln \left( \sum_{k=1}^{K} \theta_{dk} \right) \right] + (\tau_l - 1) \left[ \ln \left( 1 - \sum_{k=1}^{K} \theta_{dk} \right) \right] (37)
\]
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,

\[ \text{score}_{\text{UMass}}(k) = \sum_{i=2}^{M} \sum_{j=1}^{M} \log \frac{p(w_i, w_j)}{p(w_i)} + 1 \]  

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 \(\phi_{dk}\) 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

\[ f_{dk} = f_{dk} + \sum_{n=1}^{N_d} \langle z_{dnk} \rangle + \sum_{l=1}^{L} \langle y_{dl} \rangle \mu_{lk} \]  

\[ g_{d} = g_{d} + \sum_{n=1}^{N_d} K \langle z_{dnk} \rangle + \sum_{l=1}^{L} \langle y_{dl} \rangle \sigma_{l} \]  

\[ h_{d} = h_{d} + \sum_{l=1}^{L} \langle y_{dl} \rangle \tau_{l} \]  

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.

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

\[^2\]https://www.kaggle.com/datasets/netflix-inc/netflix-prize-data
4.2 Netflix Movie Recommendation

The Netflix dataset is bigger compared to the anime dataset. The dataset consists of details pertaining to ratings for 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 4: Query results for Netflix data.

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Agent Cody Banks</td>
</tr>
<tr>
<td>2</td>
<td>Agent Cody Banks 2: Destination London</td>
</tr>
<tr>
<td>3</td>
<td>Lilo and Stitch 2</td>
</tr>
<tr>
<td>4</td>
<td>101 Dalmations II: Patch’s London Adventure</td>
</tr>
<tr>
<td>5</td>
<td>Mean Creek</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>S. No.</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dawn of the Dead</td>
</tr>
<tr>
<td>2</td>
<td>Sasquatch</td>
</tr>
<tr>
<td>3</td>
<td>Wrong Turn</td>
</tr>
<tr>
<td>4</td>
<td>Evil Remains</td>
</tr>
<tr>
<td>5</td>
<td>Dead Birds</td>
</tr>
</tbody>
</table>

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


