Paired Feature Constraints for Latent Dirichlet Topic Models

Nagesh Bhattu Sristy* D.V.L.N.Somayajulu† and R.B.V.Subramanyam†
*PhD Student,Department of CSE
NIT Warangal
Email: nageshbhattu@gmail.com
†Department of CSE
NIT Warangal
Email: soma,rbvs66@nitw.ac.in

***Abstract*—Non Parametric Bayes models, so called family of Latent Dirichlet Allocation (LDA) Topic Models have found application in various aspects of pattern recognition like sentiment analysis, information retrieval, question answering etc. The topics induced by LDA are used for later tasks such as classification, regression(movie ratings), ranking and recommendation. Recently various approaches are suggested to improve the utility of topics induced by LDA using various side-information such as labeled examples and labeled features. Pair-Wise feature constraints such as *cannot-link* and *must-link*, represent weak-supervision and are prevalent in domains such as sentiment analysis. Though *must-link* constraints are relatively easier to incorporate by using dirichlet tree, the *cannot-link* constraints are harder to incorporate using the dirichlet forest. In this paper we proposed an approach to address this problem using posterior constraints. We introduced additional latent variables for capturing the constraints, and modified the gibbs sampling algorithm to incorporate these constraints. Our method of Posterior Regularization has enabled us to deal with both types of constraints seamlessly in the same optimization framework. We have demonstrated our approach on a product sentiment review data set which is typically used in text analysis.**

I. Introduction

Topic models are widely used in information retrieval and text mining. They provide low dimensional representation of data to capture the problem specific dependencies. Latent Dirichlet Allocation (LDA) is non-parametric bayesian inference method for topic models. LDA models the data as a multinomial over topics and each topic being a multinomial over features. As each document specific topic distribution is generated from a dirichlet, the overfitting issues of the earlier models such as PLSI are well addressed. The topics inferred by LDA are usually used for the sub tasks such as classification(sentiment-analysis), regression, ranking, etc. Blei and McAuliffe [1] proposed Supervised LDA(SLDA) to learn the topic models which are meant for classification. Instead of learning topic model and classifier separately, SLDA includes the classification label also into the generative process and it has reflected in better classification accuracy for SLDA compared to LDA.

Sentiment Analysis is a text classification problem where sentiment is often expressed in phrases expressing positive or negative polarity. Compared to normal text classification, sentiment often changes swiftly by the presence or absence of certain words. For example the presence of 'not' often changes the polarity in the opposite side. Fully supervised models like Support Vector Machines(SVM) have shown remarkable classification accuracy(87%). Availability of Labeled examples requires human labeller or domain expert (involving time and cost). Recent research in parametric methods focussed on learning a model using weak supervision in the form of feature constraints. These constraints are easy to express. For example it is easy express that the presence of word 'marvellous' will strongly bias a review to be 'positive'. Druck et al. [2] have shown a method to build a classifier from feature constraints, where a classifier was built with just 20 feature constraints. It is easier to label features compared to labelling complete examples. Pair-Wise Feature constraints are further weaker form of constraints where the objective is to restrict the model such that a pair of features should have similar sentiment or opposite sentiment. For example in the sentiment phrase 'battery lasts long but screen size is tiny', 'long' and 'tiny' should get opposite sentiment. Similarly in the sentence 'battery lasts long and screen size is large', 'long' and 'large' should get similar polarity. These are called cannot link and *must-link* constraints and they are studied in constrained clustering, which are well explained in the book by Basu et al. [3]. Andrzejewski et al. [4] has addressed it and proposed a solution based on dirichlet forest prior. The dirichlet forest prior encodes *must-link*s and cannot links in the form of weighted edges in a forest. So the features which must be linked together share a common dirichlet tree prior. On the otherhand, the cannot links are not easy to encode using a prior. Andrzejewski et al. [4] address this problem by forming clusters of *must-link* features. However, this solution represents all the node sets which don't violate the cannot link constraint(independent sets of a connected graph) to have the same dirichlet prior parameter. Looking at the cannot link constraint, this solution does not correspond to the model we are looking for. Our model should just model cannot link. The complexity of the algorithm is also effected, as there are $O(3^{|r|/3})$ number of clusters in the graph. Zhai et al. [5] also observed these issues when they tried to group product features based on cannot link constraints.

978-1-4799-3400-3/13/$31.00 ©2013 IEEE

There are broadly two ways of including the side information in the model, either by modifying the prior or by tweaking the posterior. We take the second approach to address the difficulty of modelling the cannot links through the prior. We introduce Latent variables each of which model the similarity or dissimilarity dependency among the features. The similarity or dissimilarity is measured by means of divergence between the feature's topic distributions. As this divergence should be minimum for *must-links*, we introduce pseudo observed variables and set them to 0. For *cannot-link* constraints we set them to 1, as divergence should be maximized.

In the next section we review the literature of the problem domain. In section 3 we will go through our approach for handling cannot links. In the subsequent section we will show our results on two datasets widely used in product review mining.

II. Related Work

Topic Models were introduced for modelling the domain specific dependencies using latent variables called topics, which gives lower dimensional representation of data which is inherently high dimensional such as text data or images. Earlier attempts such as Probabilistic Latent Semantic Indexing(PLSI) by Hofmann [6], do not model the document specific topic distributions well, so the topics learnt are suffered from the problems of overfitting. Blei et al. [7] addressed these issues in LDA by modelling document specific topic distributions to be generated from dirichlet prior. Conjugacy of Dirichlet with Multinomial distribution is useful in deriving the posterior. The latent variables are modelled as mutually independent given the data. This assumption simplifies the inference algorithms. But real world data need not confirm to these assumptions and topic coherence is an important issue that needs to be handled to infer meaningful topics. Lafferty and Blei [8] and Mimno and McCallum [9] and Kim and Sudderth [10] have proposed methods to handle the coherence among topics through different approaches. Blei and McAuliffe [1], Mimno and McCallum [9] and Lacoste-Julien et al. [11] are introduced to learn better topic models taking additional meta data also into modelling process. Blei and McAuliffe [1] uses the class labels of examples in a down stream model to learn a topic model where the class labels are incorporated in to the generative process. Including class labels into the model has improved the classification accuracy. Lacoste-Julien et al. [11] and Mimno and McCallum [9] are both down stream models which take metadata like author names for documents.

Topic Models are also used for Sentiment Analysis by Lin et al. [12] where sentiment and topic are learnt simultaneously. Lakkaraju et al. [13] have modelled coherence of sentiment words with in a span of text over different aspects using their variant topic models meant for modelling sequences.

The research literature on parametric models has recently proposed various ways of specifying constraints either through instance specific constraints or corpus wide constraints which are called Posterior Regularization (PR) Ganchev et al. [14] and Generelized Expectation (GE) Druck et al. [2] respectively. This way of constraining models keeping the inference tractable, is particularly useful when modifying the prior is difficult according to user's requirements. In the case of topic models the inference problem is not tractable which is the case with most of the problems dealt by PR or GE. Inducing feature constraints into topic models is studied in He [15]. They have explored inducing feature constraints into sentiment topic models by modifying the variational bayes procedure to take care of KL divergence feature constraints. In this paper we explore a different kind of prior knowledge in the form of cannot and must links. Balasubramanyan and Cohen [16] have also studied methods of constraining the topic models by restricting word-topic entropy to be less. This has resulted in topic models which encourage words to be associated to fewer topics, which is good in certain contexts.

Zhai et al. [5] have also studied how to use cannot link constraints in learning a topic model, meant for product feature grouping. But they use weighted average of topics for both cannot link and *must-link*. But it is not clear how such average works in a general context like ours. We deal with both of these constraints differently. We also found that it is easier to model *must-link* constraints using modified dirichlet prior and our regularization is useful to model the cannot links as it requires lesser number of constraints in posterior.

III. Approach

A. Generative Process in LDA

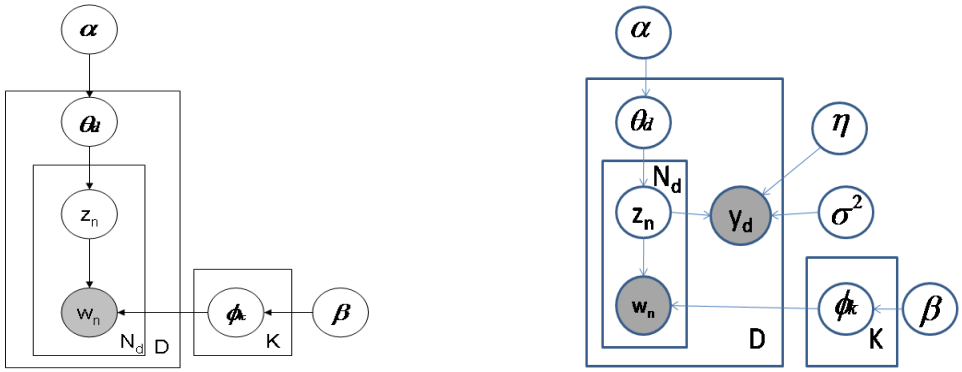
We build topic models for text data divided into documents $d \in \mathcal{D}$. Each document is a sequence of words w_n coming from a fixed vocabulary $\mathcal{V}$. LDA is a generative topic model where topics are modelled as latent variables $z_n \in \mathcal{T}$ for each w_{di}. Let there be $|\mathcal{T}|$ topics. Let there be N_d words in each document d. The generative process for LDA is given in 1a. The posterior distribution corresponding to the graphical model shown in figure is

$$p(\theta, \phi, z, w|\alpha, \beta) \propto \prod_{k=1}^{K} Dir(\phi_k|\beta) * \left(\prod_{d=1}^{\mathcal{D}} Dir(\theta_d|\alpha) * \left(\prod_{i=1}^{N_d} \theta_d^{z_n} * \phi_{z_n}^{w_n} \right) \right) \quad (1)$$

The generative process is given below

1) Let K be the number of topics in the model
2) Let $\{\phi_k\}_{k=1}^{K}$ be the topic specific word distributions which are samples from $\phi \sim Dir(\beta)$ over the vocabulary $\mathcal{V}$.
3) For each document $d \in \mathcal{D}$
   - Sample a topic distribution $\theta_d \sim Dir(\alpha)$
   - Let N_d be the document size generated from some known process
   - for each of the word w_n in document d, $n \in 1..N_d$
     - Sample a topic $z_n \sim Mult(\theta_d)$
     - Sample a word $w_n \sim Mult(\phi_{z_n}$

The above posterior is intractable due the coupling of topic and word-topic distributions. It is usually solved using approximate



(a) LDA (b) SLDA

Fig. 1. Graphical Models of LDA and Supervised LDA

inference techniques. The two predominant ways of approximate inference for the above posterior are variational inference and gibbs sampling. We use gibbs sampling as proposed by Griffiths and Steyvers [17]. The collapsed gibbs sampling algorithm will iterate through all the latent variables whose individual updates are given as

$$p(z_n = k|w_n, w^{-n}, z^{-n}, \alpha, \beta) \propto (N_{dk}^{-n} + \alpha) * \frac{N_{kw_n}^{-n} + \beta}{\sum_{w' \in \mathcal{V}} N_{kw'}^{-n} + |\mathcal{V}|\beta} \quad (2)$$

In the above equation the probability of $z_n = k$ given the current word and all the other assignments z^{-n} in all positions except this n and all the other words w^{-n} representing all the other word occurrences omitting the word occurrence at position n. N_{kw} represents the number of occurrences of word w in topic k. N_{dk} represents the number of topic assignments made to topic k in document d.

B. Modelling Side Information

LDA as explained above provides topics which are based on raw co-occurrences in the data. But it does not include any of the side information usually available such as author name, publishing year, ratings available for a review etc. SLDA was designed to model the classification and regression response variable which is typically sentiment class or ratings in our example. SLDA's generative process is shown in 1b. We can observe that the response variable is modelled to be dependent on the set of latent variables (one response for the whole document). This is called down stream model. Lacoste-Julien et al. [11] have modelled similar supervision in DiscLDA (2a). In this model the latent variables are modelled to be dependent on the response, a transformation matrix T and document specific topic distribution θ_d. The purpose of transformation matrix is to generate samples from class specific topics. The topics are divided into class dependent and class independent. This is called upstream model. The advantage of such a model is that, the transformation is learnt from data where as in SLDA, it is done either through linear regression (for real valued response) or softmax (for classification response). This allowed Mimno and McCallum [9] to learn Dirichlet Multinomial Regression(DMR), where the only difference is that the parameter vector α is learnt from meta data such as author names, venue, year. They also observe similar advantages of learning fully conditional model. These examples capture a case where the latent variables share some common source of properties typically observed in ancestor-children hierarchy. In the problem we have considered, the objective is to model dependencies among features which are words. The dependency here is that the words should have similar/dissimilar topic distributions. Topic distributions of words in the vocabulary are not directly modelled here and we want aggregation(divergence) over topic distributions of two or more words to minimum(0). So we link these pseudo observed variables and topic assignments using a common descendent C_f (Figure 2b) which represents the aggregation. Balasubramanyan and Cohen [16] have used similar regularization for controlling the entropy of word-topic distribution.

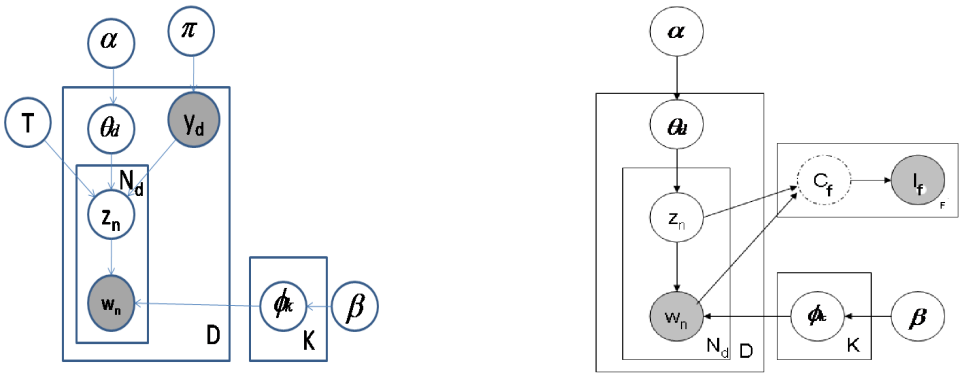
C. Paired Feature Constraints

In this section we will present pair-wise feature constraints in detail. As explained in the introduction a feature can be similar to multiple features and also dissimilar to some more features. When a feature is similar to some features the divergence among topic distributions of these features should be minimum. Let N_{kv} represent the number of times a word v appears in topic k. Using N_{kv} we can define the distribution of word v over topics as given below.

$$p(t = k|v) = \frac{N_{kv}}{\sum_{k'} N_{k'v}} \quad (3)$$

$p(t|v)$ defines a probability distribution over topics. The Jenson-Renyi Divergence for a set of probability distributions $p_1(x)$ to $p_K(x)$ defined for $x \in$ event space $\mathcal{X}$ as follows.

$$\begin{aligned} Let p_m(x) &= \frac{1}{K} \sum_{i=1}^{K} (p_i(x)) \\ D_{JS}(p_1, p_2, ..., p_K) &= \frac{1}{K} \sum_{i=1}^{K} H(p_i) - H(p_m) \\ H(p(x)) &= -\sum_{x \in \mathcal{X}} p(x) log(p(x)) \end{aligned} \quad (4)$$



(a) Disc LDA (b) Constrained LDA

Fig. 2. Graphical Models of DiscLDA and Pair Wise-Constrained LDA

We represent our domain knowledge as a set of factors. Each factor in this set either specifies similarity or dissimilarity. Let S be the set of all similarity factors. Each factor in S is a set of features among which we have to induce similarity. If f_s is one such factor in S, we use $D_{JS}(p(t|v_1), p(t|v_2)...p(t|v_{|f_s|}))$ to represent the factor divergence in the aggregate variable C_{f_s}. For each such factor f_s we also have one pseudo observed variable l_{f_s}, which is set to zero to minimize the aggregate topic divergence. We fit each of these terms to be gaussian distributed with mean l_{f_s} and a shared common variance σ^2. Similarly we have set of all factors D, each factor f_d in D is set of features among which we have to induce dissimilarity. l_{f_d} are the pseudo observed variables for the dissimilarity constraints. They are set to 1 as the JS Divergence can at most reach 1. The joint distribution considering this dependency is

$$p(z,w,\theta,\phi,l_v|\alpha,\beta,\delta) \propto \prod_{k=1}^{K} Dir(\phi_k|\beta)\left(\prod_{d=1}^{\mathcal{D}} Dir(\theta_d|\alpha)\left(\prod_{i=1}^{N_d} \theta_d^{z_i} * \phi_{z_i}^{w_i}\right)\right) * \prod_{f\in S\,UD} exp\left(\frac{t_{JS}(f)}{2*\sigma^2}\right) \tag{5}$$

$t_{JS}(f)$ is the term which aggregates the variance of a feature factor divergences from their expected means.

$$t_{JS}(f) = -\left(D_{JS}\left(p(t|v_1), p(t|v_2)..., p(t|v_{|f|})\right) - l_f\right)^2 \tag{6}$$

In the case of *must-link* constraints, it is quite intuitive to assume associative potential which is the Jenson-Renyi Divergence among those features. But associativity will not extend to dissimilarity. Andrzejewski et.al [4] tried to deal with both dependencies alike. So it resulted in complicating the algorithm, as the number of independent sets has increased exponentially. In the current set-up, dissimilarity constraints need not be associative and hence each constraint adds a single multiplicative term into the posterior. So the complexity of our approach is linear in the number of constraints irrespective of whether they are *must-link* or *cannot-link* constraints.

IV. Experiments

We have taken the datasets used by Blitzer [18]. This dataset consists of reviews about 4 types of data each of which has 2000 reviews. We have taken the raw data and used it to build LDA. We have used mallet, a publicly available source code[1] for topic modelling. We have taken publicly available sentiment lexicon[2]. We used this lexicon to prepare constraints. We have experimented with two kinds of constraints. Similarity (*must-link*) constraints are formed using words belonging to positive(similarly negative) class. We have taken factors of size 5 which vary between 100-300 factors per experiment. We have taken the total Jenson-Renyi Divergence of all the feature/topic distributions involved in the factors to evaluate our approach. As the sizes of factors are kept uniform average and sum have the same mode. Figure 3 shows the observed values of Divergence over multiple runs of the topic modelling obtained by varying the the number of topics. We have experimented with topics ranging from 5-25 taking multiples of 5. As can be seen in the results, the divergence of the topic model has been reduced significantly. This is seen across all the datasets and all combinations of number of topics, which is statistically significant. Figure 4 represents similar experiments taking factors from features belonging to negative class. In both figures 3 and 4, the objective is to reduce the divergence among similar features, so divergence value of constrained lda is observed to be quite low compared to lda. In contrast Figure 5 shows the case of cannot link constraints formed by taking pairs of features one from positive class and one from negative class. In this case the objective is increase the divergence, hence the observed values are having higher divergence than the lda's divergence values. In all our experiments we have fixed the value of σ to be 0.01.

[1] http://mallet.cs.umass.edu/

[2] http://www.cs.uic.edu/ liub/FBS/sentiment-analysis.html

TABLE I
DATASETS

NameOfDataset	Description	No of Classes	No of Features
Books	2000 instances	2 classes	473,856
DVD	2000 instances	2 classes	
Electronics	2000 instances	2 classes	
Kitchen	2000 instances	2 classes	

TABLE II
CONSTRAINTS

NameOfDataset	Similarity Constraints	Dissimilarity Constraints
Books	177(5) pos/238(5) neg	874
DVD	175(5) pos/279(5) neg	885
Electronics	97(5) pos/111(5)neg	488
Kitchen	95(5) pos/106(5)neg	438

(a) Books (b) DVD (c) Electronics (d) Kitchen

Fig. 3. Comparison of Observed Divergence of Factors due to positive features

(a) Books (b) DVD (c) Electronics (d) Kitchen

Fig. 4. Comparison of Observed Divergence of Factors due to negative features

(a) Books (b) DVD (c) Electronics (d) Kitchen

Fig. 5. Comparison of Observed Divergence of Factors due to cannot link features

A. Likelihood

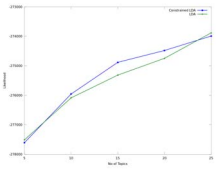
Evaluating topic models is a challenging task and perplexity or likelihood of held-out documents has been used as a measure for evaluating the efficacy of topic model being built. Our objective is to learn meaningful topics, though likelihood is not directly related to effectiveness of our technique, we show how our regularization effects the likelihood. The likelihood of the model using left-to-right algorithm as given in Hanna et.al [19] is as follows

$$p(d|\phi, \alpha k) = \prod_n P(w_n|d_{<n}, \alpha k, \phi) = \prod_n \sum_{z_{<=n}} p(w_n, z_{<=n}|d_{<n}, \alpha k, \phi)$$

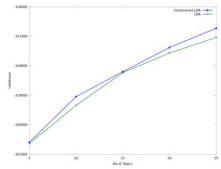
As we can see in Figure 6, the likelihood of the both constrained and unconstrained lda, do not differ much.

V. CONCLUSION & FUTURE WORK

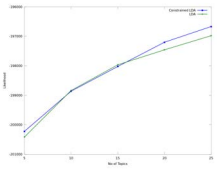
In this paper we have explored the posterior regularization of topic models using pair-wise constraints which include both similarity as well as dissimilarity. This provides unified



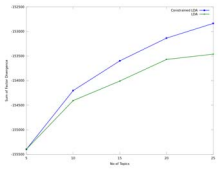
(a) Books



(b) DVD



(c) Electronics



(d) Kitchen

Fig. 6. Comparison of Likelihoods of Document

framework of regularization which otherwise would have been complicated with the use of appropriate prior. We want to explore other forms of weak supervision in the form feature constraints into this framework.

Acknowledgment

The authors would like to thank NIT Warangal for providing requisite facilities for this project.

References

[1] D. M. Blei and J. D. McAuliffe, "Supervised topic models," *arXiv preprint arXiv:1003.0783*, 2010.

[2] G. Druck, G. Mann, and A. McCallum, "Learning from labeled features using generalized expectation criteria," in *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '08. New York, NY, USA: ACM, 2008, pp. 595–602. [Online]. Available: http://doi.acm.org/10.1145/1390334.1390436

[3] S. Basu, I. Davidson, and K. Wagstaff, *Constrained clustering: Advances in algorithms, theory, and applications*. CRC Press, 2008.

[4] D. Andrzejewski, X. Zhu, and M. Craven, "Incorporating domain knowledge into topic modeling via dirichlet forest priors," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 25–32.

[5] Z. Zhai, B. Liu, H. Xu, and P. Jia, "Constrained lda for grouping product features in opinion mining," in *Advances in knowledge discovery and data mining*. Springer, 2011, pp. 448–459.

[6] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57.

[7] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[8] J. D. Lafferty and D. M. Blei, "Correlated topic models," in *Advances in neural information processing systems*, 2005, pp. 147–154.

[9] D. M. Mimno and A. McCallum, "Topic models conditioned on arbitrary features with dirichlet-multinomial regression," in *UAI*, 2008, pp. 411–418.

[10] D. I. Kim and E. B. Sudderth, "The doubly correlated nonparametric topic model," in *Advances in Neural Information Processing Systems*, 2011, pp. 1980–1988.

[11] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "Disclda: Discriminative learning for dimensionality reduction and classification," in *Advances in neural information processing systems*, 2008, pp. 897–904.

[12] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment-topic detection from text," *IEEE Trans. on Knowl. and Data Eng.*, vol. 24, no. 6, pp. 1134–1145, Jun. 2012. [Online]. Available: http://dx.doi.org/10.1109/TKDE.2011.48

[13] H. Lakkaraju, C. Bhattacharyya, I. Bhattacharya, and S. Merugu, "Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments," 2011.

[14] K. Ganchev, J. a. Graça, J. Gillenwater, and B. Taskar, "Posterior regularization for structured latent variable models," *J. Mach. Learn. Res.*, vol. 11, pp. 2001–2049, Aug. 2010. [Online]. Available: http://dl.acm.org/citation.cfm?id=1756006.1859918

[15] Y. He, "Incorporating sentiment prior knowledge for weakly supervised sentiment analysis," vol. 11, no. 2, pp. 4:1–4:19, Jun. 2012. [Online]. Available: http://doi.acm.org/10.1145/2184436.2184437

[16] R. Balasubramanyan and W. W. Cohen, "Regularization of latent variable models to obtain sparsity," 2013.

[17] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, pp. 5228–5235, 2004.

[18] J. Blitzer, "Domain adaptation of natural language processing systems," Ph.D. dissertation, University of Pennsylvania, 2008.

[19] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 1105–1112.