2015 International Conference on Computational Science and Computational Intelligence

BL-LDA: Bringing Bigram to Supervised Topic Model

Youngsun Park, Md. Hijbul Alam, Woo-Jong Ryu and SangKeun Lee
Department of Computer Science and Engineering
Korea University, Seoul, Korea
Email: {youngsun90, hijbul, skdirwj, yalphy}@korea.ac.kr

***Abstract*—With the increasing amount of data being published on the Web, it is difficult to analyze their content within a short time. Topic modeling techniques can summarize textual data that contains several topics. Both the label (such as category or tag) and word co-occurrence play a significant role in understanding textual data. However, many conventional topic modeling techniques are limited to the bag-of-words assumption. In this paper, we develop a probabilistic model called Bigram Labeled Latent Dirichlet Allocation (BL-LDA), to address the limitation of the bag-of-words assumption. The proposed BL-LDA incorporates the bigram into the Labeled LDA (L-LDA) technique. Extensive experiments on Yelp data show that the proposed scheme is better than the L-LDA in terms of accuracy.**

***Keywords*-Topic Modeling; Data Mining; Data Analysis; Text Classification**

I. Introduction

Recently, there has been a constantly increasing flow of data information, and we are expected to spend significant amount of time to process the large, unstructured collection of data in order to extract useful information from them. There is therefore a need to for an analytical method for large amounts of unstructured data. There have been extensive studies on the extraction of meaningful information from large amounts of data, and topic modeling is one of the popular approaches employed to analyze the content of unstructured text.

Topic modeling based on the probabilistic generative model is a technique used to identify hidden topics from within a collection of data, along with the clustering of words that are semantically relevant. Generative models describe how textual data can be modeled as a combination of the probability distributions of a topic. There are several representative models including the Latent Dirichlet Allocation (LDA), Labeled LDA (L-LDA). In most topic models, topics are represented as groups of words without the topic's name. By reading these topic terms, the content can be easily and rapidly understood.

LDA [1] has been widely used for many text mining applications, and it relies on the bag-of-words assumption, which ignores the word order. Unsupervised topic models such as LDA rely on word-occurrence (unigram in LDA) statistics in the textual data. Those models assume that the text is a combination of underlying topics. However, the number of topics in the text is not fixed. Finally, those models often discover topics that are not easily interpreted because the words have multiple meanings or new meanings when combined with other words. For example, the meaning of "Dr" is different from that of "Dr Pepper".

L-LDA [2] is a generative model for multi-labeled text that marries the multi-label supervision, which is common to modern textual data with the word-assignment ambiguity resolution of the LDA family of models. This model directly associates each label with one topic. Because L-LDA also relies on the bag-of-words assumption, this model considers a single word. For example, in the labeled text "Hawaiian pizza with Dr Pepper is the best combination. #Foods #junkfood" (the word after # is the label in the text), each unigram word is associated with both labels "Foods", "junkfood". The labels are related with the words "Dr" and "Pepper", not "Dr Pepper".

To alleviate the problems associated with models described above, in this paper, we propose the Bigram Labeled LDA (BL-LDA), which is a supervised generative model for multi-label text, and which extends L-LDA by applying the bigram concept. This model represents each text by a mixture of topics. Considering the word order, this model enables better interpretability of the topics by aligning a label to each topic. In addition, this model clarifies the meaning of the words. For example, assume that we are modeling the multi-labeled text, "Hawaiian pizza with Dr Pepper is the best combination. #Foods #junkfood " (the word after # is the label in the text). This model may treat the bigram word "Dr Pepper" as being related to its labels "Foods" and "junkfood".

To determine the performance of BL-LDA, we conduct extensive experiments on a real data (Yelp). Experimental results show that BL-LDA can classify the text into related topics, and that it is better than state-of-the-art in terms of accuracy.

In summary, the main contributions of the paper are as follows:

- We propose a supervised generative model by incorporating the bigram concept for multi-labeled corpora, i.e. Bigram Labeled LDA (BL-LDA).
- We conduct extensive experiments using a real data (Yelp) which is labeled as user-generated textual content. Further, we compare our model with L-LDA.
- The experimental results verify that the classification

978-1-4673-9795-7/15 $31.00 © 2015 IEEE
DOI 10.1109/CSCI.2015.146

result which is obtained by our model is better than of L-LDA in terms of accuracy.

The remainder of this paper is organized as follows. We first give a brief survey of related works in Section II. Section III describes in detail the generative process of BL-LDA. Experimental results are presented in Section IV, and we conclude this paper in Section V.

II. Related Works

Topic modeling identifies hidden semantic topics from observed textual data. Over the past decade, topic models have gained popularity in the management a textual data.

LDA [1] is an unsupervised algorithm that models each text as a mixture of topics, where the topics are unigram distributions over a given vocabulary. Even though a given text may be very short, and may consist of only a few words, the text would itself have several topics.

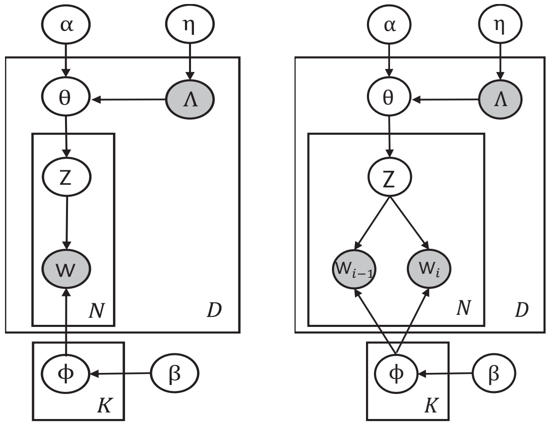
Unsupervised probabilistic topic models such as LDA can discover hidden topics within a text without the need for training. However, the discovered topics may at times be too general. Therefore, many supervised and semi-supervised topic models have been proposed to address this problem [3][4][5]. Unlike LDA, our proposed model is a supervised topic model. Given a set of labels that are themselves topics, it simplifies the interpretability of the discovered topics.

Several modifications to LDA have been proposed in order to incorporate supervision. These include supervised LDA and L-LDA. L-LDA [2] is a supervised generative model for multi-label text. This model extends LDA by defining one-to-one correspondence labels with its latent topics and observed labels to directly learn word-label correspondences. Similar to LDA, this model relies on the bag-of-words assumption, and it is therefore very dependent on word occurrence, while treating only unigram words without considering the word order. Many studies in the area of text classification and the prediction of labels for the text are based on supervised topic models [5][6][7]. Unlike L-LDA, in order to better understand textual data, our model considers both unigram and bigram words.

As many conventional topic modeling techniques rely on the bag-of-words assumption, each word has a topic distribution for each topic. To ensure that the text is understood, many extensive studies have been carried out [8][9][10][11]. Word units are considered as N-grams or word pairs with co-occurrence patterns instead of a single word.

Biterm Topic Model (BTM) [8] learns topics by directly modeling the generation of word co-occurrence patterns (i.e., biterms) in a corpus. This model makes word pairs (a biterm) that frequently occur together and that can be related to each other, while our model is designed based on both unigrams and bigrams by exploiting the corpus-level adjacent word.

Several methods have focused on simultaneously inferring phrases and topics by creating complex generative mechanisms [9][12]. The resultant models can directly output phrases and their latent topic assignment.



(a) L-LDA

(b) BL-LDA

Figure 1. Graphical representation of (a) L-LDA and (b) BL-LDA.

Table I
Summary of Notations.

Symbol	Meaning
K	number of topics
D	number of documents
N	number of words in a document
z	topic associated with the word in textual data
w	word token
θ	multinomial (Discrete) topic distribution
ϕ	binomial (Bernoulli) word distribution
Λ	presence of labels
η	Dirichlet prior of Λ
α	Dirichlet prior of θ
β	Dirichlet prior of ϕ

Other methods apply a post-processing step to unigram-based topic models. These methods assume that all of the words in a phrase will be assigned to a common topic.

III. Proposed Methodology

Conventional topic models are aimed at determining latent topics of text based on corpus-level word co-occurrence patterns. These models also rely on the bag-of-words assumption. By incorporating them with the bigram concept to L-LDA, we propose a supervised generative model for multi-label corpora. In this section, we introduce the generative process of BL-LDA and describe the graphical model.

A. Bigram Labeled LDA

BL-LDA is a supervised generative model that involves the incorporation of the bigram concept for multi-labeled text. The goal of this model is to correctly assign topics to textual data.

By considering the word order and modifying both the topic distribution θ and topic z, we solve the problem associated with the “Dr Pepper” example in Section II.

BL-LDA assumes the following generative process:

1) For each topic z :
    - a) Generate $\phi_{zw} \sim$ Dirichlet$(\phi_{zw}|\beta)$
2) For each document d :
    - a) For each topic z :
        - i) Generate $\Lambda^{(d)} \sim$ Bernoulli$(\Lambda^{(d)}|\Phi_k)$
    - b) Generate $\alpha^{(d)} = L^{(d)} \times \alpha$
    - c) Generate $\theta^{(d)} \sim$ Dirichlet$(\theta^{(d)}|\alpha^{(d)})$
    - d) For each word w_i :
        - i) Generate $z_i^{(d)} \sim$ Mult$(\theta^{(d)})$
        - ii) Generate unigram: $w_i \sim$ Mult$(\phi_{z_i(d)})$
        - iii) Generate bigram: $w_{i-1}w_i \sim$ Mult$(\phi_{z_i(d)})$

The graphical models of L-LDA and BL-LDA are shown in Figure 1. We observe that the graphical model of BL-LDA is similar to that of L-LDA. However, z_i of BL-LDA is a topic that is associated with the word w_i and the bigram word $w_{i-1}w_i$ in the text. BL-LDA considers not only unigram but also bigram words.

A word distribution ϕ is generated by a Dirichlet prior β for each topic. The existence of a set of labels $\Lambda^{(d)} = (l_1, \cdots, l_K)$ indicates the presence/absence of topics. We set the number of topics in BL-LDA to the number of labels K in the documents (one-to-one correspondence). In addition, the topic distribution θ is generated by both a Dirichlet topic prior α and the presence of labels Λ with a Dirichlet prior η. Then, the word-topic assignment z is generated from the topic distribution θ. Both the unigram word w_i and the adjacent bigram word $w_{i-1}w_i$ are generated from the word distribution ϕ with a Dirichlet prior β and topic z.

B. Inference

We apply collapsed Gibbs sampling [13] in order to estimate the hidden parameters, word distribution ϕ, topic distribution θ and the topic associated with the words z_i of the BL-LDA.

$$P(z_i = j|z_{-i}) \propto \frac{\{N_{w_{i-1},w_i,j}\}_{-i} + \beta_{w_{i-1},w_i}}{\{N_j\}_{-i} + \beta} \times \frac{\{N_{d_i,j}\}_{-i} + \alpha_j}{\{N_{d_i}\} + \alpha} \tag{1}$$

Equation (1) shows the sampling probability for a topic i in a document d. Moreover, $\{N_{w_{i-1},w_i,j}\}$ is the overall number of both bigram words w_{i-1},w_i in topic j. To determine its perword label assignments z, once the binomial word distribution ϕ is learned from the training data, we can perform inference on any new labeled test data using Gibbs sampling restricted to its tags. In addition, we can also compute θ by appropriately normalizing the topic assignments z. Because there is a one-to-one correspondence between the labels and topics, the model can take topical summaries for each label k in terms of the topic-specific distribution β_k Equation (1) is similar to equation for L-LDA, but in this paper, we consider both unigram and bigram words.

IV. Experiments

In this section, we first describe the dataset and the experimental results obtained. Given labeled data, we evaluate the text classification accuracy of BL-LDA for comparison with L-LDA[1].

A. Dataset

We collect multi-labeled text from Yelp[2] which publishes crowd-sourced reviews pertaining to local businesses. Each set of review from Yelp contains the business id, review text and categories. One set of review may have multiple categories. Thus, we use the categories as labels/topics of the review. The Yelp data has several unique characteristics. First, categories are very specific, e.g., hamburger, pizza and restaurant. Secondly, the review data depends greatly on its categories. Finally, words that represent topics well in both unigrams and bigrams occur frequently in both test and training data. In addition to comparing the predictive accuracy, it is valuable to look at the inferred topics.

The distribution of categories in the Yelp data is skewed toward a few categories. To deal with this problem, we consider the number of categories in the review data and randomly select 5, 10, 15, 20 categories (Table II) from among the top 50 most common categories. Table III lists the number of data, words, the average number of topics of data and the average number of words of data.

To reduce the amount of low-quality data, we first remove English stopwords and then perform stemming on the tokens in the data using the porter stemming algorithm[3] to address the various forms of words (e.g., ride, rides, riding). We randomly divide the Yelp data into training data for 80% and test data for 20%.

B. Text Classification

Given a training data consisting of textual data with multiple labels, we train the BL-LDA and L-LDA model using parameters α=0.1, β=0.01. Then, we predict the relevant labels for each review in the test data. Based on the word distribution for each topic, we consider the words that are ranked in top 100 and 1,000 to classify the test data which consist of 5, 10, 15 and 20 separate topics in the ground truth. The categories (topics) in the ground truth are listed in Table III. We compare three models, which are L-LDA (unigram), L-LDA (bigram) and BL-LDA (unigram and bigram), consider unigram only, bigram only and both unigram and bigram separately. To increase the reliability of

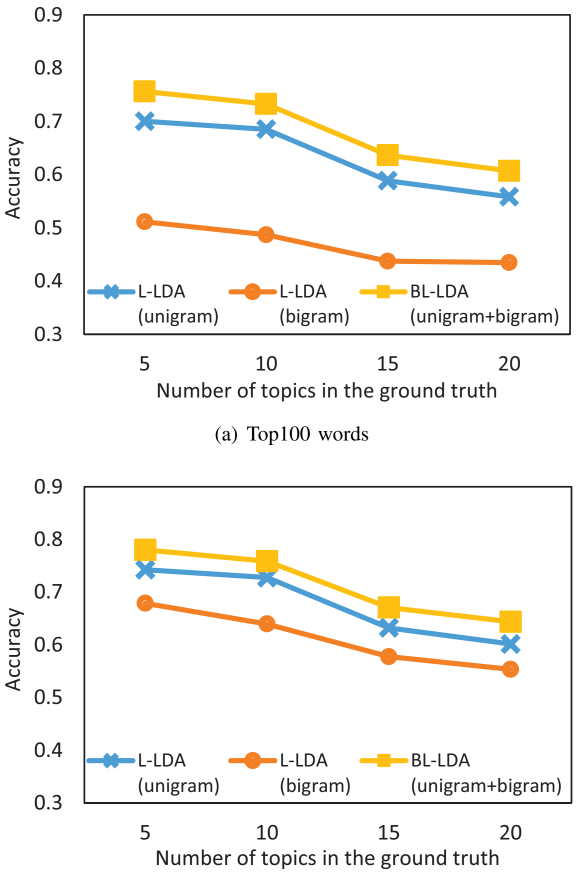
[1] https://github.com/myleott/LabeledLDA

[2] http://www.yelp.com

[3] http://tartarus.org/martin/PorterStemmer/java.txt

Table II
NUMBER OF CATEGORIES IN THE GROUND TRUTH AND ITS NAME.

Number of categories in the ground truth	Category Name
5	Health & Medical, Coffee & Tea, Italian, Chinese, Burgers
10	Health & Medical, Coffee & Tea, Italian, Chinese, Burgers, Hair Salons, Hotels, Nail Salons, Grocery, Home & Garden
15	Health & Medical, Coffee & Tea, Italian, Chinese, Burgers, Hair Salons, Hotels, Nail Salons, Grocery, Home & Garden, Pets, Breakfast & Brunch, Doctors, Fitness, Specialty Food
20	Health & Medical, Coffee & Tea, Italian, Chinese, Burgers, Hair Salons, Hotels, Nail Salons, Grocery, Home & Garden, Pets, Breakfast & Brunch, Doctors, Fitness, Specialty Food, Bakeries, Ice Cream & Frozen Yogurt, Pubs, Dentists, Desserts



(a) Top100 words

(b) Top1000 words

Figure 2. Comparison of classification performance with top100(a) and top1000(b) word.

Table III
SUMMARY OF DATASET.

Dataset	Number of topics in the ground truth			
	5	10	15	20
#data	1.3M	2.0M	2.6M	3.0M
#word	1.0M	1.7M	2.1M	2.4M
avg. #topic	2.00	2.01	2.06	2.11
avg. #word	38.98	41.53	41.22	40.85

our results, we perform each experiment twice and take the average the accuracy.

Figure 2 shows a comparison of the accuracy of the classification with regards to the different number of topics in the ground truth. Figure 2(a) shows experimental results considering the 100 top-ranked words for each topic, while Figure 2(b) shows the results when we consider the 1,000 top-ranked words for each topic. All of the test data have the ground truth with either one or more topics. From Table III, there is an average of more than two topics in the data. Thus, we calculate the accuracy as in Equation (2).

$$Accuracy = \frac{correctly\ classified\ data}{total\ test\ data} \qquad (2)$$

Unigram words have higher word distributions for each topic than bigram. When there are 20 topics in the ground truth, the accuracy of BL-LDA (A=1, B=10), L-LDA (unigram) and L-LDA (bigram) is 0.64, 0.60 and 0.56, respectively. The accuracy of BL-LDA is 4% and 8% higher than that of L-LDA (unigram) and L-LDA (bigram), respectively.

The accuracy of L-LDA (unigram) is higher than L-LDA (bigram). However, the number of distinct bigram words is higher than the number of distinct unigram words. It means that meaning of bigram is more clearer than of unigram. For example, there is a text "chicken pizza, pepperoni pizza, hawaiian pizza and cheese pizza". There are five distinct unigrams such as "chicken", "pizza", "pepperoni", "hawaiian" and "cheese", and seven distinct bigrams such

Table IV
AVERAGE NUMBER OF BIGRAM WORDS IN TOP-RANKED WORD DISTRIBUTION FOR EACH TOPIC.

	Top100				Top500				Top1000			
B (A=1)	5	10	15	20	5	10	15	20	5	10	15	20
1	1.8	2.1	1.7	2.0	33.7	32.6	32.8	33.5	132.4	132.3	128.8	130.5
3	14.1	13.9	13.5	14.0	140.8	142.8	138.9	138.1	387.0	389.5	379.7	381.1
5	28.0	27.2	27.1	27.1	221.9	222.4	216.4	216.3	524.5	523.3	519.1	520.7
10	57.6	56.6	55.1	54.4	331.4	330.9	324.0	323.3	706.2	706.0	697.8	697.9

Table V
MOST COMMONLY OCCURRING WORDS IN SOME OF THE TOPICS INFERRED FROM THE YELP DATA OBTAINED L-LDA AND BL-LDA.

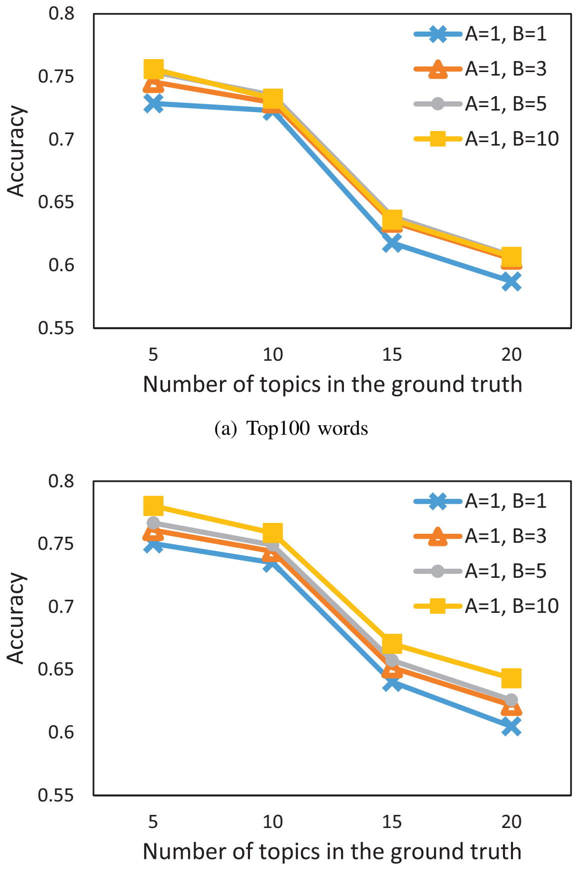
	Topic 1 (Italian)	Topic 2 (Nail Salons)	Topic 3 (Fitness)	Topic 4 (Dessert)	Topic 5 (Hotels)
L-LDA (unigram)	pizza food service time sauce	nail time pedicure salon gel	class gym time yoga love	chocolate cupcake food time flavor	hotel stay nice time check
L-LDA (bigram)	italian food happy hour food service italian restaurant pasta dish	nail salon mani pedi gel manicure nail tech customer service	yoga studio highly recommend hot yoga yoga class 24 hour	cheesecake factory ice cream grand lux red velvet peanut butter	stay hotel customer service 5 star time check
BL-LDA (unigram+bigram)	pizza food italian food happy hour food service	nail salon nail mani pedi gel manicure nail tech	class gym yoga studio highly recommend time	cheesecake factory ice cream grand lux chocolate red velvet	hotel stay stay hotel resort fee customer service

as “chicken pizza”, “pizza pepperoni”, “pepperoni pizza”, “pizza hawaiian”, “hawaiian pizza”, “pizza cheese” and “cheese pizza”. This may explain the low accuracy of the bigrams in the graph. Thus, combining unigram and bigram words with high word distributions can affect the classification accuracy of the test data. And Figure 2 shows that BL-LDA, which considers both unigram and bigram, is better than the other models in both cases, which consider the 100 top-ranked words and 1,000 top-ranked words for each topic. In Figure 2, all of the lines follow the same trend. Because we consider only one topic that returned with the highest word-distribution score for the test data, as the number of topics in the ground truth increases, the accuracy decrease. As shown in Figure 2 and Table IV, the bigram distribution is too low. If the test data contains bigram words that have a high probability of word distribution for each topic, we obtain a good prediction of its topic. However, the occurrence probability of bigrams is lower than that of unigrams. Thus, we perform experiments with different weights for the bigram word distribution. We test the models with the word distribution for topic k, WD^k_{word} and WD^k_{bigram}, with different weights A and B.

$$WD^k_{(words)} = A \times WD^k_{(unigram)} + B \times WD^k_{(bigram)} \quad (3)$$

Table IV represents the unigram and bigram word distribution of the top-ranked words for each topic. Then, Figure 3 shows the accuracy of the classification comparison with regards to the different weights for the bigram word distribution for each topic. The graphs, in Figure 3(a) and Figure 3(b), follow similar trends. The line for the parameters, A=1 and B=10, is better than other lines. As the number of top words increases (top-ranked 100 and 1000 words), the accuracy increases as well as the difference in the accuracy is increasing.

Table V presents the most-frequently occurring words in the topics extracted from the Yelp training data using two models, L-LDA and BL-LDA. The words in a topics inferred from the Yelp data obtained BL-LDA is better to represent the topics. For example, the meaning of a word “service” depends on the topics. On the other hand, the meaning of words in BL-LDA, such as “food service” in topic 1 and “customer service” in topics 1 and 2 are clearer than “service”.



(a) Top100 words

(b) Top1000 words

Figure 3. Comparison of classification performance with respect to different weights for bigram word distribution for each topic.

V. Conclusion

Topic modeling techniques can be used to summarize textual data having several topics. It is therefore useful to analyze the content of data within a short period of time. In this paper, we proposed Bigram Labeled LDA (BL-LDA), which is a supervised generative model, by incorporating the bigram concept for multi-labeled text. We performed experiments using a real Yelp data and we verified that the classification performance of the proposed model is better than that of the previous model L-LDA in terms of accuracy. Because in BL-LDA, both the label and word co-occurrence represent each topic, the accuracy of the classification performance of BL-LDA is better than that of L-LDA. We combined the unigram and bigram word distribution for each topic, and observed that the weight parameter for the bigram word distribution affects its classification performance.

However, there is still room for future improvement of our work. We will aim to conduct extensive experiments on other dataset and make comparisons with a BTM model that considers biterms, which comprise two terms that occur in a textual data, but which are not adjacent. Moreover, we will also aim to explore the usage of our model in various real-world applications such as content recommendation.

Acknowledgment

This work was supported by the Next-Generation Information Computing Development Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science and Technology (No. 2012M3C4A7033344).

References

[1] D. Blei, A. Ng, and M. Jodan, “Latent dirichlet allocation,” *JMLR*, vol. 3, pp. 993-1022, 2003.

[2] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, “Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora,” in *EMNLP*, 2009, pp. 248-256.

[3] N. Kawamae, “Supervised n-gram topic model,” in *WSDM*, 2014, pp. 473-482.

[4] D. M. Blei and J. D. McAuliffe, “Supervised topic models,” in *NIPS*, 2007, pp. 121-128.

[5] D. Quercia, H. Askham, and J. Crowcroft, “TweetLDA: supervised topic classification and link prediction in twitter,” in *WebSci*, 2012, pp. 247-250.

[6] J. L. Tang, S. Yu, and J. Ye, “Extracting shared subspace for multi-label classification,” in *SIGKDD*, 2008, pp. 381-389.

[7] Q. Mei, M. X. Shen, and C. Zhai, “Automatic labeling of multinomial topic models,” in *SIGKDD*, 2007, pp. 490-499.

[8] X. Cheng, X. Yan, Y. Lan, and J. Guo, “BTM: topic modeling over short texts,” *TKDE*, vol. 26, no. 12, pp. 2928-2941, 2014.

[9] X. Wang, A. McCallum, and X. Wei, “Topical n-grams: phrase and topic discover, with an application to information retrieval,” in *ICDM*, 2007, pp. 697-702.

[10] A. El-Kishky, Y. Song, C. Wang, C. R. Voss, and J. Han, “Scalable topical phrase mining from text corpora,” in *VLDB*, 2015, pp. 305-316.

[11] H. M. Wallach, “Topic modeling: beyond bag-of-words,” in *ICML*, 2006, pp. 977-984.

[12] R. V. Lindsey, W. P. Headden III, and M. J. Stipicevic, “A phrase-discovering topic model using hierarchical pitman-yor processes,” in *EMNLP*, 2012, pp. 214-222.

[13] T. Griffiths and M. Steyvers, “Finding scientific topics,” *PNAS*, vol. 101, no. suppl. 1, pp. 5228-5235, 2004