A Large Scale Multi-label Text Classification Method using Z-Label LDA

Cam-Van Thi Nguyen
vanntc@vnu.edu.vn
University of Engineering and Technology,
Vietnam National University, Hanoi

Mai-Vu Tran
vutranmai@gmail.com
University of Engineering and Technology,
Vietnam National University, Hanoi

ABSTRACT

Multi-label Learning (MLL) is a supervised learning model that has attracted much attention of the research community in recent years because of its wide variety applicability. In this paper, we built a multi-label classification model using Latent Dirichlet Allocation with Topic-in-set Knowlegde (z-Label LDA) on the Vietnamese data domain. z-Label LDA is a variant of LDA which is intended to provide additional supervised information as a hidden topic into the LDA called "z-label". We also have experimented on the dataset in the field of Education collected from Vietnamese online newspapers. Parallel, we applied the hidden topic model LDA to generate a prior-knowledge dataset comprising topics and typical keywords representing each topic. The supervised information also makes the topic assignment more consistent. With this approach, the effectiveness of the model has been demonstrated experimentally, this paper has obtained initial positive results.

KEYWORDS

Multi-label Learning, LDA with Topic-in-set Knowlegde, z-Label LDA

1 INTRODUCTION

Over the past decade, multi-label learning has garnered considerable attention from machine learning and related communities. Multi-label learning is extensively applied to a wide variety of issues such as automatic annotation for multimedia content, bioinformatics, web mining, rule mining, information retrieval, etc. There are many publications has been published to address multi-label classification problem. Therefore, there are several ways to classify multi-label classification methods.

In [8], authors categorize the approaches for multi-label classification problem into two categories: *problem transformation method* and *algorithm adaption method*. The problem transformation methods is the methods that transform the multi-label classification problem either into one or more single-label classification or regression problems, for both of which there exists a huge bibliography of learning algorithms. The algorithm adaptation methods is approaches that extend specific learning algorithms in order to handle multi-label data directly. According to [6], the popular algorithms for multi-label document classification can be divided into two categories: *discriminative approaches* and *generative modeling approaches*. The generative approach first learns a model with respect to words and labels and then constructs a discriminate function to predict testing documents via Bayesian rules. The advantages of the generative topic model are obvious:1) it would be easy to give complex latent structures responsible for a set of observations; 2) the relationship between the different factors could be easily exploited by introducing the latent variables [7].

An effective approach towards applying the generative model is to use model called Latent Dirichlet Allocation (LDA) [1]. LDA is an unsupervised topic model which is a generative probability model for discrete data sets based on the Dirichlet distribution. In the context of the text modeling context, the LDA perceives each document as being made up of a set of topics in which the continuous-valued mixture proportions are distributed as a random variable Dirichlet. Some proposed approaches for this generative model approach are Labeled LDA [9]; Flat-LDA, Prior-LDA, Dependency-LDA [10]; LDA with topic-in-set knowledge [4]. z-Label LDA is a modification of LDA algorithm. In z-Label LDA, words are assigned with 'z-labels'. The model predefined topic-in-set knowledge which implies predefined terms for certain topics was added to supervise the topic task for individual terms. This approach combines the pattern-discovery power of LDA with user-provided guidance that can encourage LDA to recover topics of corpus.

In this paper, we built a multi-label classification model using z-Label LDA for Vietnamese text in the field of Education. We chose z-label LDA to assign the topics for documents in our corpus because z-label LDA provides some control over the topic assignment so that the generated corpus is more balanced and better fit our purpose. The supervised information also make the topic assignment more consistent. Addition, according [11], if the topic-based representations of documents are to be used for document clustering or classification, providing z-labels for words can be seen as similar to semi-supervised learning with labeled feature. Therefore, our approach can be considered as a semi-supervised model.

We have the following contributions:

- Build a multi-label classification model which showed a relatively good performance for the multi-label classification task.
- Create a prior-knowledge dataset in the field of Education including topics and its corresponding keywords which can be useful for future researches on Vietnamese data domain.

2 RELATED WORKS

The approach used generative model has two advantages: (1) considering the prediction of labels at the word-level, rather than at the document-level; (2) modeling all the observed labels simultaneously, rather than handling each label independently. A common approach for modeling the corpus applying the generative model is known as LDA. Because LDA is an unsupervised topic model, hence many new approaches are proposed to exploit the quality of supervised learning such as Supervised LDA [13], Labeled LDA, etc.

Supervised LDA can be applied to labeled documents by adding each document d to a variable y_d for classification. Each y_d value is modeled by the Generalized Linear Model (GLM) in the vector of the average number of topic $\overline{z} = \frac{1}{N_d} \sum_{N_d}^{n=1} z_n$ for that documents. Therefore, this approach can predict labels by computing the topics assigned to a test document to obtain a $\overline{z}$ value. Moreover, training model in this way tends to create topics that can "explain" the value of y for the training set. In this way, indirect label information affects each document according to position discovered by the model. This helps LDA to adapt to supervised learning standards, but restricts the user's guidance in providing labels or values.

Labeled LDA proposed by [9] assumes that each document associated with a label is represented by a K-dimensional binary vector. Each label is associated with its own topic that can only be used in documents containing these labels. This allows them to learn topics that closely relate to their respective labels.

[10] have studied on multi-label document categorization, and then introduced three statistical topic models namely Flat-LDA, Prior-LDA, Dependency-LDA. Although under different definitions, Flat-LDA treats the a priori probabilities of labeling as important for predictions while the Prior-LDA expands Flat-LDA in order to calculates the difference in frequency of the observed labels in a corpus. The Prior-LDA is designed as a two-stage process for each document. First, it identifies the document-label Dirichlet prior through the label sampling on an entire distribution, as estimated by the label frequency observations. Then, it creates the words given the document-label prior. Dependency-LDA extends Prior-LDA, the focus of Dependency-LDA is on the dependencies of different labels. It assumes that there exist dozens of topics out of the label, in which each topic is a corpus-wire multinomial distribution across labels. Therefore, for each document, the process of document-label prior generation is that first samples a set of topics from the corresponding distribution. It samples labels for sampling topics, rather than sampling labels from an entire corpus distribution as Prior-LDA.

Like we mention above, our approach can be considered as a semi-supervised model. Therefore, it solved the problem of traditional supervised learning and unsupervised learning. Specifically, in many practical applications, it is very time-consuming to collect a large amount of labeled data, while the unlabeled data is very rich and easy to obtain. Supervised learning approach requires a large amount of labeled data to be effectively implemented, whereas unsupervised learning approach only focuses on unlabeled data. Therefore, the semi-supervised approach can solve the problem of both approaches. Seeded LDA was proposed by [12]. This approach could help the topic model to study a specific topic of user interest. By providing a set of seed words that are supposed to represent the topics in the dataset. This model uses this seed set to increase the probability of the word-topic and the document-topic distribution. Pham [5] proposed a semi-supervised multi-label classification algorithm called MASS which can exploit both unlabeled data and specific features to enhance the performance of classification model on the data set of hotel (for tourism) reviews. In the training process, MASS algorithm exploits the specific features per prominent class label chosen by a greedy approach as an extension of LIFT algorithm, and unlabeled data consumption mechanism from TESC. In classification, the 1-Nearest-Neighbor (1NN) is applied to select appropriate class labels for a new data instance. The role of labeled data in MASS characterizing the shape of the text clusters. The increments in size of labeled dataset also make some contribution to the performance of the model. authors show that the best result in each category seem to be stable with the different number of unlabeled texts.

3 METHODOLOGY

3.1 Review of Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is a generative probabilistic model widely used in topic modeling. LDA is an unsupervised learning model which is first proposed by [1] has been widely used on text topic modeling. The key idea behind the LDA model is to assume that the words in each document were generated by a mixture of topics, where a topic is represented as a multinomial probability distribution over words, for text data for example. [1] introduced the LDA model within a general Bayesian framework and developed a variational algorithm for learning the model from data. The completeness of the generative process for documents is achieved by considering Dirichlet priors on the document distributions over topics and on the topic distributions over words.

In this section, following the notations of [3], we briefly review LDA. The notations and its definitions is presented in the Table 1 below.

Table 1: LDA's parameters and variables

Notation	Definition
D	Number of documents
T	Number of topics
$\mathbf{w} = w_1 ... w_n$	A corpus of D documents
d_i	The document of word w_i
z_i	The hidden topic from generated from w_i
$\phi_j^{(w)} = p(w\|z = j)$	Word distribution for topic j
$\theta_j^{(d)} = p(z = j)$	Topic distribution for document d
α	Document-topic Dirichlet distributions
β	Topic-word Dirichlet distributions

LDA involves the following generative model:

$$\theta \sim Dirichlet(\alpha) \tag{1}$$

$$z_i | \theta^{(d_i)} \sim Multinomial(\theta^{(d_i)}) \tag{2}$$

$$\phi \sim Dirichlet(\beta) \tag{3}$$

$$w_i | z_i, \phi \sim Multinomial(\phi_{z_i}) \tag{4}$$

Estimating parameters of LDA exactly is intractable. In this paper, we use Gibbs Sampling [3] to inference the hidden topic $\mathbf{z}$. For simplicity, we assume Dirichlet parameters α and β are scalars. The full conditional equation used for sampling individual z_i values from the posterior is given by:

$$P(z_i = v|z_{-i}, \mathbf{w}, \alpha, \beta) \propto (\frac{n^{(d)}_{-i,v} + \alpha}{\sum_u^T (n^{(d)}_{-i,u} + \alpha)})(\frac{n^{(w_i)}_{-i,v} + \beta}{\sum_{w'}^W (n^{(w')}_{-i,v} + \beta)}) \tag{5}$$

where $n^{(d)}_{-i,v}$ is the number of times topic v is used in document d, and $n^{(w')}_{-i,v}$ is the number of ties word w_i is generated by topic v. The $-i$ notation signifies that the counts are taken omitting the value of z_i.

3.2 Latent Dirichlet Allocation with Topic-in-set Knowledge (z-LDA)

Topic-in-set knowledge [2] also called z-Label LDA is a modification of LDA algorithm. The main idea of z-label LDA is that some words are restricted to be generated by a subset of topic, called its "z-label". A z-label for observed word w_i consists of a set $C^{(i)}$ of possible values for the corresponding latent topic index z_i, and can be thought of as a (hard or soft) constraint. The probabilities of latent topic assignments **z** which violate these constraints are then penalized for a soft constraint, or set to zero in the case of a hard constraint [4]. The detailed modification is presented below:

Let

$$q_{iv} = (\frac{n^{(d)}_{-i,v} + \alpha}{\sum_u^T (n^{(d)}_{-i,u} + \alpha)})(\frac{n^{(w_i)}_{-i,v} + \beta}{\sum_{w'}^W (n^{(w')}_{-i,v} + \beta)}) \tag{6}$$

.

Let $C^{(i)}$ be the set of possible topics that can be assign to word t. We define an indicator function $\delta(v \in C^{(i)})$, which return 1 if $v \in C^{(i)}$ and return 0 otherwise. The Gibbs Sampling for a word t in formula (5) is modified to:

$$P(z_i = v|z_{-i}, \mathbf{w}, \alpha, \beta) \propto q_{iv}\delta(v \in C^{(i)}) \tag{7}$$

To restrict z_i to a single value(e.g., $z_i = 5$, this can be accomplished by setting $C^{(i)} = \{5\}$. Likewise, restricting z_i to a subset of values {1,2,3} by setting $C^{(i)} = \{1, 2, 3\}$. Finally, for unconstrained z_i, set $C^{(i)} = \{1, 2, 3, ..., T\}$, in which case our modified sampling (7) reduces to the standard Gibbs sampling (5).

According to [4], this formulation gives us a highly flexible method for inserting prior domain knowledge into the inference of latent topics, allowing us to set $C^{(i)}$ independently for every single word w_i in the corpus. This hard constraint model could also be relaxed. Let $0 \leq \eta \leq 1$ be the strength of our constraint, where η recovers the hard constraint (7) and $\eta = 0$ recovers unconstrained sampling (5). Then we can modify the Gibbs sampling equation as follows:

$$P(z_i = v|z_{-i}, \mathbf{w}, \alpha, \beta) \propto q_{iv}(\eta\delta(v \in C^{(i)}) + 1 - \eta) \tag{8}$$

4 PROPOSED APPROACH FOR MULTI-LABEL CLASSIFICATION

To better understand how our model works, we first describe the basic ideas behind user interests modeling by means of a diagram below. The model is described in Figure 1.

This model consists two main phases:

- Phase 1: Training model
- Phase 2: Inference and evaluation model

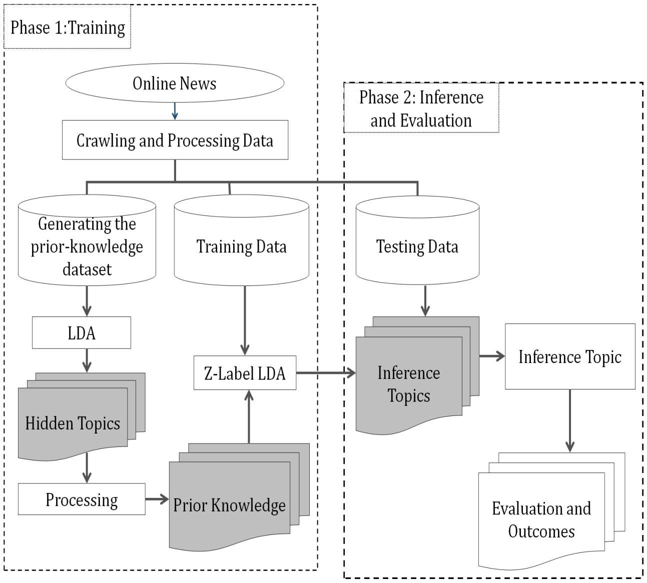

Figure 1: Multi-Label Classification model

Detailed step by step phases will be described in detail below.

Phase 1: Training model

Collecting and preprocessing data

We experiment on two sets of data in the area of Education and . With the Education field data set, we gathers articles from Vietnamese online newspapers. The data preprocessing consists of the following steps:

- Remove article which is not in Vietnamese.
- Tokenized the sentences in each post using Vietnamese Text Processing Toolkit[1] by Le Hong Phuong.
- Remove punctuations, symbols and words which do not appear in the dictionary.
- Remove stop words: base of the list of 600 Vietnamese stop words, we use statistic methods to extend this stop list: words with highest frequencies and do not contribute to the meaning of document are added to the list. Our final stop list consists of 1600 words.
- Remove too short articles.

Collected data after preprocessed will be divided into three different subsets: (1) Generating the prior-knowledge dataset, (2) Training dataset and (3) Evaluating dataset. Table 2 shows the number of collected dataset.

Table 2: Dataset

Dataset	Generating Data	Training data	Testing data
19820	6216	11650	1954

Generating the prior-knowledge dataset

[1] https://github.com/phuonglh/vn.vitk

In this step, we collected prior-knowledge for z-LDA algorithm. Specifically, we built the "seed list": a list of common topics in the field of Education. Each topic is provided with an explicit name and a set of "seed words", which are the most-related words to that topic. To generate a relatively complete, balanced and unbiased set of topics, the standard LDA was used. With our seed list, the training dataset was sampled to produce a dataset with relatively balanced topic distribution.

Table 3: An example of generated prior-knowledge dataset

Topic0	Topic1
register (đăng_ký) examination (thi_tuyển) selection (xét_tuyển)	orientation (định_hướng) job (công_việc) career (sự_nghiệp)

Table 3 gives an example of the results of a priori knowledge generated by applying the LDA algorithm. Based on the factual survey and generated data, this prior-knowledge dataset will be manually processing, keeping only topics with the most-related keywords and naming for the topic.

Training model using z-Label LDA

After the above two steps, an unlabeled data set and a prior-knowledge set will be collected to train the model using the z-Label LDA. The input of the LDA z-Label model includes the prior-knowledge set and the subject matter of this prior-knowledge set. We used the seed list to assign the z-Labels. Formally, if word w_i is a seed word of topic z, z-label set of w_i is $z : \delta(v \in C^{(t)}) = 1$ if $v = z$ and $\delta(v \in C^{(t)}) = 0$ otherwise.

We trained the z-LDA model on our corpus with above z-label assignments. Hyperparameter α and β are set to 0.5 and 0.1 respectively. This model is saved to infer the topic of new article later on.

After phase 1, we have an unlabeled corpus and the seed list as input for the z-LDA algorithm.

Phase 2: Inference and evaluation model

Inference

On the implementation side, it all depends on the Gibbs sampling step. The topics are randomly generated, then, Gibbs is run while tracking the document and the number of topics. When the sampling is converged, the probability distribution of the documents/topics and the distribution of topics/words can be simply calculated.

Evaluating

Evaluation methods based on labeling data are commonly used are Precision (P), Recall (R), and F_1.The model will calculate P, R, F measurements on each label and averaged for all labels.

5 EXPERIMENT AND RESULTS

5.1 Setting

LDA parameters setting: The LDA parameters will be set as in Table 4. After running LDA for the generated prior-knowledge dataset we will acquire 200 hidden topics and the keywords will be automatically generated for each topic. However, these 200 topics do not have a clear topic title for each topic, but only the keywords for each topic are generated by the LDA based on the word distribution in the article. Articles collected from Vietnamese newspapers are attached to tags. These tags are words that appear with high frequencies in the article and are most relevant to the content in the article. We takes all the tags of the collected data, and simultaneously records the actual survey and collects nearly 200 tags. Therefore, we select T = 200 and runs LDA to generate the first hidden topic dataset. We've found that there are many topics that have keywords that are meaningless and do not related to the rest of the same topic, which will be removed. Topics that include keywords that are not related to each other and do not share the meaning of the topic will also be discarded. In parallel, we name the topic and add new topics based on the hidden topic generated by the LDA.

Table 4: LDA parameters

Prior-knowledge	Loops	Topics	Stop words	α	β
6216	200	200	1631	0.5	0.1

Finally, we collected a prior-knowledge dataset consisting of 62 topics, covering the typical keywords for each of the topics in the field of Education.

z-Label LDA parameters setting: The LDA parameters will be set as in Table 5.

Table 5: z-Label LDA parameters

Training data	Loops	Topics	Stop words	α	β
11650	300	62	1631	0.5	0.1

5.2 Results

We created a dataset of 1954 articles in the Education and label them. The labels consists of 62 topics. The model will calculate P, R, F measurements on each label and averaged. Table 6 and Table 7 show experiment results in the Education data domain.

Table 6: The average classification result of all labels

P	R	F
73.16 %	41.56 %	53.00 %

In Table 7, the "ID" column illustrates the "name" of the topic in the seed list. As we described in the phase of Generating the prior-knowledge dataset, after using LDA algorithm and manually processing techniques, the collected seed list will contain topic with most-related keywords. But in this seed list, we have not named the topic yet and still keep the topic number as its name. The "No. Labeled" column presents the number of data that is labeled. P(%), R(%), and F(%) represent Precision, Recall, and F-1 score for classification respectively. Look at the classification results in Table 7, there are many labels that give classification result is

Table 7: Classification result for each label

ID	No. Labeled	P (%)	R (%)	F (%)
1	29	0.73	0.38	0.5
2	10	0.43	0.3	0.35
3	20	0.5	0.4	0.44
4	2	0	0	0
5	360	0.6	0.66	0.63
6	26	0.57	0.65	0.61
7	221	0.59	0.87	0.7
8	90	0.4	0.66	0.5
9	40	0.42	0.6	0.49
10	20	0.36	0.25	0.29
11	62	0.67	0.5	0.57
12	2	0	0	0
13	4	0	0	0
14	2	0	0	0
15	58	0.62	0.74	0.68
16	26	0.54	0.27	0.36
17	115	0.51	0.77	0.62
18	17	0.64	0.41	0.5
19	8	0.4	0.25	0.31
20	10	0	0	0
21	4	0	0	0
22	14	0.14	0.07	0.1
23	14	0.14	0.07	0.1
24	57	0.57	0.82	0.67
25	12	0.17	0.42	0.24
26	96	0.5	0.71	0.58
27	87	0.58	0.62	0.6
28	21	0.59	0.81	0.68
29	17	0.31	0.29	0.3
30	25	0.1	0.08	0.09
31	5	0.14	0.2	0.17
32	102	0.6	0.82	0.69
33	2	0	0	0
34	15	0.28	0.53	0.36
35	95	0.51	0.72	0.6
36	14	0.14	0.29	0.19
37	24	0.4	0.25	0.31
38	61	0.37	0.31	0.34
39	10	0.08	0.1	0.09
40	28	0.24	0.14	0.18
41	114	0.47	0.66	0.55
42	83	0.52	0.64	0.58
43	13	0.4	0.15	0.22
44	48	0.48	0.4	0.43
45	44	0.35	0.18	0.24
46	89	0.48	0.75	0.59
47	1	0	0	0
48	39	0.27	0.41	0.33
49	13	0	0	0
50	20	0.39	0.45	0.42
51	74	0.4	0.5	0.44
52	86	0.4	0.56	0.46
53	13	0.18	0.15	0.17
54	69	0.52	0.58	0.55
55	3	0	0	0
56	3	0	0	0
57	2	0	0	0
58	2	0	0	0
59	1	0	0	0
60	56	0.32	0.46	0.38
61	37	0.41	0.38	0.39
62	39	0.35	0.64	0.45

0%. This result can be explained by two main reasons: 1) The prior-knowledge dataset does not cover the dataset, so while there are some topics are unlabeled, there are some topics that are more assigned. 2) The creation of a golden corpus to evaluate the model is difficult because of the large number of labels so the evaluation dataset is unbalanced.

Due to the number of labels up to 62 labels, the classification result for each label is not good for all labels. However, some prominent labels still produce quite high results. Table 8 presents 5 topics for classification results ≤ 63% for F1 score.

Table 8: 5 topics have the highest classification results

ID	Topics	Articles	P (%)	R(%)	F(%)
7	Admissions (Tuyển sinh)	221	0.59	0.87	0.70
23	Inspection (Thanh tra giáo dục)	57	0.57	0.82	0.67
27	Educational regulation (Quy chế giáo dục)	21	0.59	0.81	0.68
15	Extracular activities (Hoạt động ngoại khóa)	58	0.62	0.74	0.68
5	Contests (Các cuộc thi)	360	0.60	0.66	0.63

6 CONCLUSION AND FUTURE WORKS

Z-Label LDA shows relatively good performance for many label categorization than traditional topic models, namely LDA. Our model has shown a high performance and potential when applying in practical dataset. Due to the large number of topics (62 topics), the number of evaluating data was not evenly distributed across topics, so there was only 53% for F1 score for all labels. However, the highlighted topics with the high distribution in the corpus still show good and satisfactory results, proving the effectiveness of the proposed model. In the future, we will continue to expand by researching and experimenting some other multi-label classification algorithms to compare with existing models. We also refining z-Label LDA model applied to multi-label classification. At the same time, we also used several measures to evaluate multi-label classification methods to demonstrate the effectiveness of the model. In addition, the prior-knowledge dataset needs to be finalized and more explored to improve the efficiency and accuracy of the classification method.

REFERENCES

[1] David M. Blei, Andrew Y. Ng and Michael I. Jordan. 2003. *Latent Dirichlet Allocation.* Journal of Machine Learning Research, 3:993-1022.

[2] D. Andrzejewski, and X. Zhu. 2009. *Latent Dirichlet Allocation with Topic-in-Set Knowledge.* NAACL HLT 2009 Workshop on SemiSupervised Learning for Natural Language Processing, 43-48.

[3] Griffiths, Thomas L., and Mark Steyvers. 2009. *Finding scientific topics.* Proceedings of the National academy of Sciences 101.suppl 1: 5228-5235.

[4] Andrzejewski, David Michael, Mark Craven, and Xiaojin Zhu. 2010. *Incorporating domain knowledge in latent topic models.* University of Wisconsin at Madison, Madison, WI.

[5] Pham, T. N., Nguyen, V. Q., Dinh, D. T., Nguyen, T. T., & Ha, Q. T. 2017. *MASS: a semi-supervised multi-label classification algorithm with specific features.* In Advanced topics in intelligent information and database systems (pp. 37-47). Springer International Publishing.

[6] Li, Ximing, Jihong Ouyang, and Xiaotang Zhou. 2014. *Supervised topic models for multi-label classification.* Neurocomputing 149, 811-819.

[7] Wang Hongning, Minlie Huang, and Xiaoyan Zhu. 2008. *A generative probabilistic model for multi-label classification.* Eighth IEEE International Conference on Data mining.

[8] Tsoumakas, Grigorios, and Ioannis Katakis. 2006. *Multi-label classification: An overview.* International Journal of Data Warehousing and Mining 3.3

[9] Ramage D., Hall D., Nallapati R., & Manning C. D. 2009. *Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora.* In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: olume 1 (pp. 248-256). Association for Computational Linguistics.

[10] Rubin T. N., Chambers A., Smyth P., & Steyvers M. 2012. *Statistical topic models for multi-label document classification.* Machine learning, 88(1), 157-208.

[11] Druck G., Mann G., & McCallum A. 2008. *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, pp. 595-602.

[12] Jagarlamudi J., Daumé III H., & Udupa R. 2012. *Incorporating lexical priors into topic models.* In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. 2012, pp. 204-213

[13] Mcauliffe, Jon D., and David M. Blei. 2008. *Supervised topic models.* Advances in neural information processing systems.