A Survey of Multi-Label Topic Models

Sophie Burkhardt, Stefan Kramer
Johannes Gutenberg University of Mainz, Institute of Computer Science
Staudingerweg 9
Mainz, Germany
{burkhardt,kramer}@informatik.uni-mainz.de

ABSTRACT

Every day, an enormous amount of text data is produced. Sources of text data include news, social media, emails, text messages, medical reports, scientific publications and fiction. To keep track of this data, there are categories, key words, tags or labels that are assigned to each text. Automatically predicting such labels is the task of multi-label text classification. Often however, we are interested in more than just the pure classification: rather, we would like to understand which parts of a text belong to the label, which words are important for the label or which labels occur together. Because of this, topic models may be used for multi-label classification as an interpretable model that is flexible and easily extensible. This survey demonstrates the manifold possibilities and flexibility of the topic model framework for the complex setting of multi-label text classification by categorizing different variants of models.

1. INTRODUCTION

Recently, a sub-field of multi-label classification has emerged which is called multi-label topic modeling. This field brings together unsupervised topic models based on latent Dirichlet allocation (LDA [5]) and multi-label classification, a supervised task where each instance in a dataset may be assigned one or multiple labels. LDA is a generative model for text corpora that yields highly interpretable models since each topic is associated with a probability distribution over words. While it is originally an unsupervised Bayesian model, it may also be used in the supervised or semi-supervised setting.

Multi-label topic models combine two main features: They are used to classify texts with one or several labels, this is the multi-label part, but at the same time, they also provide a semantic description of the different labels in the form of topics. Labels are grouped or divided into topics or topic hierarchies such that each topic is associated with a probability distribution, either over words or over labels or topics on a different hierarchy level. These topics, that are a byproduct of the classifier training, are useful in their own right and can provide a helpful addition to the pure classification output.

There are three main reasons why it is useful to combine multi-label classification with topic models.

- First, after training a topic model, each word in a text document is associated with a corresponding topic or at least a distribution over topics. This enables to understand *why* a document is classified in a certain way. The words that lead to assigning a specific label and relevant areas of the text may be identified.
- Second, independently from the classification performance at testing time, we can check what the model has learned after training by inspecting the topics. This way, certain words are identified as important for certain topics, and we may detect unwanted noise in the topics. For example, we might see that a topic contains stop words that are irrelevant to the overall theme of the topic and subsequently remove those words to improve generalization capabilities of the model. This makes such models explainable and interpretable.
- A third reason that the learned topics are useful is that they can be influenced from the start by changing the prior. We can choose the probability distribution (e.g. Gaussian or Dirichlet), change the parameters of the distribution to adapt the degree of sparseness of the topics or immediately fix certain parameters to, e.g., user inputs [47; 31] or prelearned values from earlier models to improve convergence with limited training data. Parameters could also be changed based on user interaction [30].

Overall, unsupervised learning is a powerful way to train general-purpose systems that are able to solve many different tasks [58]. This is achieved by learning a model of the data that can be transferred to fit different kinds of applications. Therefore, the reasoning is that a well-trained topic model can also be used as an efficient classifier while at the same time providing the user with a model of the data that is generalizable.

This survey aims to give an overview of the field by categorizing multi-label topic models according to different dimensions, hoping to make them more easily accessible to newcomers and point out possible connections to related fields. First, Section 2 proposes three different categories of multi-label topic models. The problem setting and essential aspects of the two sub-fields, topic modeling and multi-label classification, are introduced as well. Section 2.1 explains LDA and the different training methods, Gibbs sampling and variational Bayes, whereas multi-label classification is covered in Section 2.5. Potential applications of different methods are discussed in Section 3. Different variants of multi-label topic models are introduced in Section 4, and a relevant selection is explained in more detail. Section 5 lists some of the most commonly used datasets in multi-label topic modeling and Section 6 reports on relevant evaluation measures. Finally,

Table 1: This table provides an overview over topic models that are related to multi-label topic models in different ways.

	supervised	multi-label	online	dependencies	nonparametric
Multi-label topic models					
LabeledLDA [60]	yes	yes	(yes)	no	no
LF-LDA [83]	yes	yes	no	no	no
DependencyLDA [66]	yes	yes	no	yes	no
DFLDA [40]	yes	yes	no	yes	no
ML-PA-LDA-C [51]	yes	yes	no	no	no
Fast Dep.-LLDA [13]	yes	yes	yes	yes	no
Stacked HDP [11]	yes	yes	no	yes	yes
Hybrid HDP [10]	yes	yes	yes	no	yes
Correlated Labeling Model [76]	yes	yes	no	yes	no
HSLDA [55]	yes	yes	no	yes	yes
Single-label topic models					
Salakhutdinov *et al.* [67]	yes	no	no	yes	yes
Supervised LDA [44]	yes	no	no	no	no
Dirichlet-multinomial regression [46]	yes	no	no	no	no
SSHLDA [42]	yes	no	no	yes	no
DiscLDA [35]	yes	no	no	no	no
MedLDA [84]	yes	no	no	no	no
Other related models					
Author-topic model [65]	no	no	no	yes	no
Partially Labeled [62]	no	no	no	yes	no
PAM [39]	no	no	no	yes	no
Correlated TM [36]	no	no	no	yes	no
nPAM [38]	no	no	no	yes	yes
Coupled HDP [69]	no	no	no	yes	yes

Section 7 discusses future research directions and the influence on the broader field of machine learning. Section 8 concludes the survey.

2. DIMENSIONS OF MULTI-LABEL TOPIC MODELS

Multi-label topic models may be differentiated according to three different dimensions. First, topic models may be trained online, which means they can be updated with new data and are more scalable to large amounts of data. These topic models are usually based on the variational Bayes training method as opposed to sampling training methods. Second, topic models may be parametric or nonparametric, where nonparametric models allow to account for different prior topic or label frequencies. Nonparametric models are able to add new topics during training, which allows them to automatically adjust to the complexity in the training data and allows the possibility for suggesting new labels that are not yet present in the data. Third, multi-label topic models are differentiated according to the way they consider label dependencies, which is a crucial feature of multi-label classifiers. This section gives an introduction to LDA topic models and then covers these three aspects in more detail. An overview of models for each of these dimensions is given in Table 1. Here, also related models that are not supervised or not multi-label are included. Supervised topic models incorporate a target variable in some way, but are not necessarily multi-label. In multi-label topic models each document may exhibit multiple labels. Online topic models can be trained on streaming data. Dependencies between topics or labels are only modeled by some of the methods, whereas in nonparametric topic models, the number of topics is not fixed and they are in some way based on hierarchical Dirichlet processes.

General information on the relevant probability distributions and graphical models are for example to be found in the well-known books by Bishop [4] or Murphy [48]. This introduction to latent Dirichlet allocation is based on the papers by Blei *et al.* [5], who proposed a variational inference training method and Griffiths and Steyvers [26], who introduced a training method based on Markov chain Monte Carlo sampling. Multi-label classification is introduced in Section 2.5. There are already a number of surveys on general multi-label classification. Therefore, we will only cover basic aspects and refer the reader to existing surveys for more details [73; 81].

2.1 Topic Models

We start by describing how latent Dirichlet allocation (LDA) is used to model collections of text documents. LDA [5] is a generative model of document collections where each document is modeled as a mixture of latent topics (see Equation 1). LDA is built on the assumption that words as well as documents are exchangeable, which means that the order in which words or documents are viewed plays no role in the training process. With respect to words, this assumption is called the "bag-of-words" assumption, meaning that each document is viewed as a bag of words where the actual sequence of words is irrelevant.

The model is given as follows, where each topic $k \in 1, \ldots, K$ is represented by a multinomial distribution ϕ_k over words that is assumed to be drawn from a Dirichlet distribution with parameter β. Document d is generated by drawing a distribution over topics from a Dirichlet $\theta_d \sim Dirichlet(\alpha)$, and for the ith word token in the document, first drawing a topic indicator $z_{di} \sim \theta_d$ and finally drawing a word $w_{di} \sim \phi_{z_{di}}$.

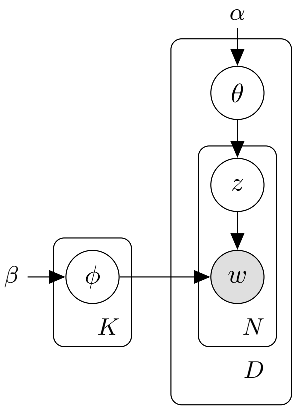


Figure 1: The graphical model of LDA.

$$
\begin{aligned}
w_{di}|z_{di}, \phi_{z_{di}} &\sim Multinomial(\phi_{z_{di}}) \\
\phi &\sim Dirichlet(\beta) \\
z_{di}|\theta_d &\sim Multinomial(\theta_d) \\
\theta &\sim Dirichlet(\alpha)
\end{aligned} \tag{1}
$$

The corresponding generative process is given as follows:

- For each topic $k \in 1, \ldots, K$
  - draw $\phi_k \sim Dirichlet(\beta)$
- For each document $d \in D$
  - draw $\theta_d \sim Dirichlet(\alpha)$
  - For each word token with indices $i = 1, \ldots, N_d$ in document d (N_d is the number of words in document d)
    - draw topic indicator $z_{di} \sim \theta_d$
    - draw word $w_{di} \sim \phi_{z_{di}}$

To learn a model over an observed document collection D, we compute the posterior distribution over the latent variables z, θ, and ϕ, which in general is intractable to compute directly.

$$
p(\phi, \theta, z|D, \alpha, \beta) = \\
\prod_{k=1}^{K} p(\phi_k|\beta) \prod_{d=1}^{D} p(\theta_d|\alpha) \prod_{i=1}^{N_d} p(z_{di}|\theta_d) p(w_{di}|\phi_{z_{di}})
$$

Therefore, it needs to be estimated. There are two main methods that are commonly used: Gibbs sampling and variational Bayes. We now describe each of the two methods.

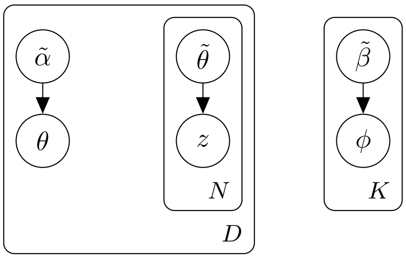


Figure 2: The graphical model of the variational distribution used to approximate the posterior of LDA.

Gibbs sampling is a special case of Markov chain Monte Carlo sampling (MCMC). Hereby, each variable is sampled conditioned on all other variables, which are fixed. Since the Dirichlet distribution is conjugate to the multinomial distribution, it is possible to integrate/collapse out the latent variables ϕ and θ from the joint distribution $p(w, z, \phi, \theta)$, where w and z are the word and topic variables for all tokens i and documents d:

$$
p(w, z) = p(w|z)p(z) = \\
\int\int \prod_k p(\phi_k) \prod_d p(\theta_d) \prod_i p(w_{di}|\phi_{z_{di}}) p(z_{di}|\theta_d) \, \mathrm{d}\phi \, \mathrm{d}\theta = \\
\int \prod_d \prod_i p(w_{di}|\phi_{z_{di}}) \prod_k p(\phi_k) \, \mathrm{d}\phi \\
\int \prod_d \prod_i p(z_{di}|\theta_d) \prod_d p(\theta_d) \, \mathrm{d}\theta,
$$

where the two integrals in the last expression can be performed separately. This enables efficient model training.

If we want to sample from the posterior for a specific $z_{di} = k$ given all remaining variables denoted by z_{-i} and all words w, the first part of the sampling equation is given by the first integral

$$
\prod_k \frac{\Gamma\left(\sum_{v \in V} \beta_v\right)}{\prod_{v \in V} \Gamma\left(\beta_v\right)} \frac{\Gamma\left(n_{wk} + \beta_w\right)}{\Gamma\left(\sum_{v \in V}\left(n_{vk} + \beta_v\right)\right)} \propto \frac{n_{-wk} + \beta_w}{\sum_v \left(n_{-vk} + \beta_v\right)}
$$

and the second part similarly by

$$
\frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \frac{\prod_k \Gamma(n_{dk} + \alpha_k)}{\Gamma(\sum_k (n_{dk} + \alpha_k))} \propto n_{-dk} + \alpha_k.
$$

Finally, the conditional probabilities for training an LDA topic model are given by [26]

$$
p(z_{di} = k|z_{-di}, w) \propto \frac{n_{wk} + \beta_w}{\sum_{w'} \left(n_{w'k} + \beta_{w'}\right)} \left(n_{dk} + \alpha_k\right), \tag{2}
$$

where n_{wk} and n_{dk} are the respective counts of topics k with words w or in documents d. α and β are hyperparameters. z_{-di} are all topic indicators except the one for token i in document d.

Intuitively, Equation 2 consists of two parts, where the first part describes the probability of a word in a certain topic. This part is responsible for words preferentially being assigned to topics where they already occur in, thus exploiting the clustering effect of the Dirichlet distribution. The second part is proportional to the probability of a topic in a certain document. Therefore, while the first part may be seen as ensuring the consistency with the global model and its topics, the second part ensures that each document minimizes the number of topics it exhibits at the local level, ensuring that a topic is more likely if other words in the same document have already been assigned to this topic.

In variational Bayesian inference a variational distribution (see Figure 2) is introduced to approximate the posterior by minimizing the Kullback-Leibler (KL) divergence between the variational distribution q and the true posterior.

$$\begin{aligned}\mathrm{KL}[q(z|D)||p(z|D)] &= \sum_z q(z|D) \log \frac{q(z|D)}{p(z|D)} \\ &= \mathbb{E}_{q(z|D)} \left[\log q(z|D) - \log p(z|D)\right]\end{aligned}$$

Usually, a fully factorized variational distribution is chosen:

$$q(\phi, \theta, z|\tilde{\beta}, \tilde{\alpha}, \tilde{\theta}) = \prod_d^D q(\theta_d|\tilde{\alpha}_d) \prod_i^{N_d} q(z_{di}|\tilde{\theta}_{di}) \prod_k^K q(\phi_k|\tilde{\beta}_k),$$

where $\tilde{\beta}$, $\tilde{\alpha}$ and $\tilde{\theta}$ denote the variational parameters.

The evidence lower bound (ELBO) that is to be maximized is given as follows:

$$\begin{aligned}&\log p(W|\alpha, \beta) \geq \\ &\mathcal{L}(\tilde{\beta}, \tilde{\alpha}, \tilde{\theta}) \triangleq \mathbb{E}_q[\log p(\phi, \theta, z, W)] + \mathcal{H}(q(\phi, \theta, z)) \\ &= \mathbb{E}_q[\log p(\theta|\alpha)] + \mathbb{E}_q[\log p(z|\theta)] + \\ &\mathbb{E}_q[\log p(w|z, \phi)] + \mathcal{H}(q(\phi, \theta, z)),\end{aligned}$$

where $\mathcal{H}$ denotes the entropy and in the first step the log-likelihood is lower bounded using Jensen's inequality. By calculating the gradient of the ELBO with respect to the variational parameters, the parameters are updated until convergence.

The local/document-level update equations for variational Bayes are [5; 72]:

$$\tilde{\alpha}_{dk} = \alpha + \sum_{i=1}^{N_d} \tilde{\theta}_{dki} \tag{3}$$

$$\begin{aligned}&\tilde{\theta}_{dki} \propto \\ &\exp\left(\Psi(\tilde{\beta}_{wk}) - \Psi(\sum_v \tilde{\beta}_{vk})\right) \exp\left(\Psi(\tilde{\alpha}_{dk}) - \Psi(\sum_{k'} \tilde{\alpha}_{dk'})\right),\end{aligned}$$

where for the expectation of the log-Dirichlet we have $\mathbb{E}[\log \theta|\alpha] = \Psi(\alpha) - \Psi(\sum_k \alpha_k)$ and Ψ is the digamma function.

However, Teh *et al.* [72] propose a collapsed version of variational Bayes where the parameters are marginalized. By using a Gaussian approximation and a Taylor expansion, that is not explained in detail here, they arrive at the update equation

$$\tilde{\theta}_{dki} \propto \frac{\tilde{\beta}_{wk} + \beta}{\sum_{w'} (\tilde{\beta}_{w'k} + \beta_{w'})} (\tilde{\alpha}_{dk} + \alpha), \tag{4}$$

where α and β are hyperparameters and N_d is the number of words in document *d*. This equation has a strong similarity with the sampling equation for collapsed Gibbs sampling. It is shown by Teh *et al.* [72] and Asuncion *et al.* [2] that this collapsed version called CVB0 [2; 23] has a better convergence than the uncollapsed one.

Based on the local variational parameters $\tilde{\theta}$, the global parameter $\tilde{\beta}$ is updated as follows:

$$\tilde{\beta}_{vk} = \beta + \sum_{d=1}^{|D|} \sum_{i=1}^{N_d} \tilde{\theta}_{dki} \mathbb{1}[w_{di} = v], \tag{5}$$

where $|D|$ is the number of documents. $\mathbb{1}[w_{di} = v]$ is one if word $w_{di} = v$ and zero otherwise.

Algorithm 1 Batch Variational Bayes

```
 1: while not converged do
 2:     for each document $d$ do
 3:         for each word token do
 4:             update local parameters (Equation 4)
 5:             normalize $\tilde{\theta}_{di}$ to sum to one
 6:         end for
 7:         update local parameters (Equation 3)
 8:     end for
 9:     update global parameters (Equation 5)
10: end while
```

For the batch variational Bayes algorithm, all local variational parameters $\tilde{\theta}_d$ for all documents d are computed and the global parameter is updated in one step. The algorithm (see Algorithm 1) may also be described as consisting of an expectation/E-step and a maximization/M-step:

- *E-Step*: For each document, the local variational parameters are optimized (lines 2–8).
- *M-Step*: The lower bound on the log-likelihood is maximized with respect to the global variational parameters (line 9).

2.1.1 *Gibbs Sampling vs. Variational Bayes*

Convergence of Gibbs sampling can be slow in comparison to variational methods, since updates only involve a sampled topic instead of the full distribution over topics as in variational Bayes. On the other hand, Gibbs sampling is unbiased, meaning it is guaranteed to learn the true posterior after an infinite number of iterations. How to determine if a Gibbs sampler has converged, however, is still an open problem. Another advantage of Gibbs samplers are the sparse updates. One word-topic assignment is updated at a time so the global counts can be efficiently updated for each document by only decrementing the counts of the previous word-topic assignments and incrementing counts for the new word-topic assignments. Variational Bayes updates are dense in comparison because the whole distribution over topics is kept for each word, making frequent updates inefficient. This is one reason why updates are usually done in minibatches. While variational Bayes converges generally faster than Gibbs sampling, the method is biased and not guaranteed to arrive at the true posterior.

The second main difference between variational Bayes and Gibbs sampling is that variational Bayes topic models may be trained online, one minibatch at a time. Gibbs sampling on the other hand is a batch method. Convergence of Gibbs sampling is only guaranteed if all data is kept available and all topic assignments keep being updated until convergence. While it is theoretically possible to train it online by only sampling topic assignments once, in practice this is only successful in restricted settings where the labels/topics for each document are already known and even then there is no guarantee for convergence. Particle filters [18] provide a way around this, but are generally not practical and efficient. This is why for streaming settings and models that have to be continuously updated, variational Bayesian methods are the preferred choice.

2.2 Online Topic Models

Algorithm 2 Online Variational Bayes

```
while not converged do
  draw minibatch $M$
  for each document $d \in M$ do
    for each word token do
      update local parameters (Equation 4)
      normalize $\tilde{\theta}_{di}$ to sum to one
    end for
    update local parameters (Equation 3)
  end for
  update global parameters (Equation 6)
end while
```

As a first dimension for differentiating multi-label topic models we consider whether the training algorithm is an online or a batch algorithm. In the batch method, all local parameters have to be kept in memory and are updated at once. For large datasets this can lead to large memory consumption and slow down training, especially at the beginning. Therefore, an online method based on minibatches is introduced by Hoffman *et al.* [28; 29] that converges faster and is efficient to train on large datasets. Teh *et al.* [72] and Asuncion *et al.* [2] improve this work by collapsing out the latent variables. Foulds *et al.* [23] combine the online part of Hoffman *et al.* and Cappe and Moulines [19], and the collapsing part of Asuncion *et al.*, resulting in an online stochastic collapsed variational Bayes (SCVB) with improved performance.

The online variational Bayes algorithm is summarized in Algorithm 2. It is similar to the batch algorithm except we now iterate over the documents minibatch by minibatch instead of over all documents at once and use a different update equation for the global parameters (line 10). Updating variational parameters $\tilde{\beta}$ for one minibatch M is done as follows, where the counts for one minibatch are scaled by $\frac{|D|}{|M|}$ to arrive at the expectation for the whole corpus and ρ_t is a parameter between zero and one:

$$\tilde{\beta}_{vk} = (1-\rho_t)\tilde{\beta}_{vk} + \rho_t \left(\beta + \frac{|D|}{|M|} \sum_{d \in M} \sum_{i=1}^{N_d} \tilde{\theta}_{dki} \mathbb{1}[w_{di} = v] \right). \tag{6}$$

Given appropriate updates and choice of hyperparameter ρ_t, the online algorithm is guaranteed to converge to the optimal variational solution. Since the global parameters are updated after each minibatch instead of each iteration over the whole dataset, the online algorithm usually converges faster than the batch algorithm, especially at the beginning of training.

2.3 Nonparametric Topic Models

According to the second dimension we differentiate parametric and nonparametric multi-label topic models. Nonparametric topic models are based on hierarchical Dirichlet processes (HDPs). In HDP topic models [71], the multinomial distribution θ from LDA is drawn from an HDP instead of a Dirichlet distribution:

$$\theta \sim DP(G_0, b_1), G_0 \sim DP(H, b_0).$$

Dirichlet processes (DPs) [71] are distributions over probability measures. If a distribution over topics is drawn from a DP, the number of topics is not fixed. This is why such models are called *nonparametric*.

Because the prior is hierarchical, there is a local topic distribution θ for each document and a global topic distribution G_0, which is shared among all documents. The advantage of this global topic distribution is that it allows topics of widely varying frequencies, whereas in standard LDA with a symmetric prior α, all topics are expected to have the same frequency. The asymmetric prior of HDP usually leads to a better representation and higher log-likelihood of the dataset [71].

Sampling methods for HDPs are mostly based on the Chinese restaurant process metaphor. Each word token is assumed to be a customer entering a restaurant, and sitting down at a certain table where a specific dish is served. Each table is associated with one dish, which corresponds to a topic in a topic model. The probability for a customer to sit down at a certain table is proportional to the number of customers already sitting at that table. This leads to a clustering effect where new customers are most likely to sit at a table that already has attracted a large number of customers. With a certain probability α (see Equation 7), the customer sits down at a new table. In this case, a topic is sampled from the base distribution. For an HDP topic model, each document corresponds to a restaurant. The topics in each document-restaurant are drawn from a global restaurant. Because all documents share the same base distribution that is discrete, represented by the global restaurant, the topics are shared by the local document restaurants. If a new table is added to a document restaurant, a pseudo customer enters the global restaurant (see Figure 3). If a new table is opened in the global restaurant, a new topic is added to the topic model.

$$P(z_n = k | z_1 \dots z_{n-1}) = \begin{cases} \frac{n_k}{n-1+\alpha} & , n_k > 0 \\ \frac{\alpha}{n-1+\alpha} & , new\ table \end{cases} \tag{7}$$

In terms of the statistics that need to be kept, in the basic version we need to store for each word not only the sampled topic, but also the table it is associated with. Also, we need to store the corresponding topic for each table.

The generative model for the two-level HDP topic model is given as follows:

$$\begin{aligned} \theta_0 | b_0, H &\sim DP(b_0, H), & \theta_d | b_1, \theta_0 &\sim DP(b_1, \theta_0) \\ z | \theta_d &\sim \theta_d, & w | z &\sim Mult(z) \end{aligned}$$

Each word w is assumed to be drawn from a multinomial distribution associated with a certain topic z. The topic indicator variable z is drawn from a document-specific distribution over topics θ_d, which is in turn drawn from a DP with base distribution θ_0. θ_0 is drawn from another DP with base distribution H. b are hyperparameters.

Three basic sampling methods were introduced by Teh *et al.* [71]. Two methods are directly based on the Chinese restaurant representation, whereas the third is the direct assignment sampler.

The currently most efficient sampling method for HDPs is by Chen *et al.* [20]. Here, an additional variable, the table indicator u, is introduced, which indicates up to which level a customer has a table contribution. In the case of a two-level HDP, $u = 2$ means the customer sits at an existing table, $u = 1$ means the customer opens a new table at the lowest level and sends a pseudo customer up to the next level, whereas $u = 0$ means that the customer opens a new table at the lowest level, sends a pseudo customer to the next level, which again opens a new table thereby adding a new topic to the topic model (see Fig. 3).

We now briefly explain how to arrive at the table indicator

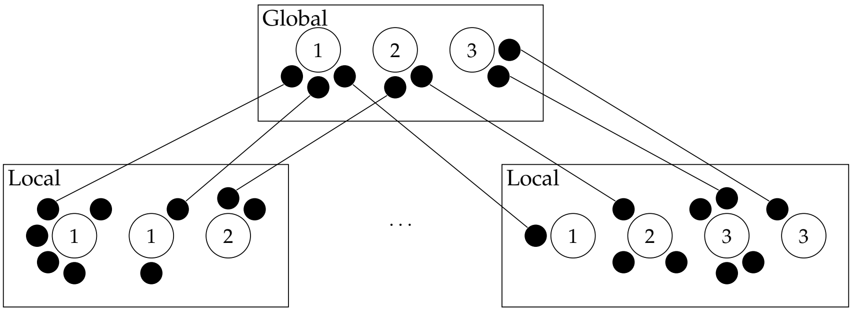


Figure 3: Illustration of a hierarchical Chinese restaurant process. For each table at a local restaurant one customer is sent to the global restaurant. The numbers represent different topics.

representation. The starting point is the direct assignment sampler by Teh *et al.* [71]. In this sampler each customer is directly assigned a topic and the number of tables per topic M_k is sampled separately. (The capital letters N, and M refer to global counts over all documents, whereas n and m refer to local document counts.)

The probability for the number of tables M_k for topic k given the number of customers per topic N_k is given by (see Buntine and Hutter [7], Lemma 8)

$$p(M_k|N_k, a, b) = \frac{(b|a)_{M_k}}{(b)_{N_k}} S^{N_k}_{M_k,a},$$

where $S^n_{m,a}$ is a generalized Stirling number defined by the recursion $S^{n+1}_{m,a} = S^n_{m-1} + (n - ma)S^n_{m,a}$, for $m \leq n$. It is zero otherwise and $S^0_{0,a} = S^1_{1,a} = 1$.

Note that in this work only the standard Dirichlet process is considered, which is the special case of the Poisson-Dirichlet process (PDP) for $a = 0$. The parameter $b > 0$ is usually estimated and n_m is the number of customers at table m. $(x)_N$ denotes the Pochhammer symbol $x \cdot (x+1) \cdot \ldots \cdot (x + N - 1) = \frac{\Gamma(x+N)}{\Gamma(x)}$ and $(x|y)_N$ denotes the Pochhammer symbol with increment y, $x \cdot (x+y) \cdot \ldots \cdot (x + (N-1)y)$, and $(x|0)_N = x^N$.

For $a = 0$, Equation 2.3 becomes (shown by Antoniak [1], compare Teh *et al.* [71], Equation 40):

$$p(M_k|N_k, b) = S^{N_k}_{M_k,0} b^{M_k} \frac{\Gamma(b)}{\Gamma(b + N_k)}.$$

Applying this result to the PDP with base distribution H, the joint probability of the samples z_i and the number of tables $M_1, M_2, \ldots, M_K$ for each topic $k = 1, \ldots, K$ is

$$p(z_1, z_2, \ldots, z_N, M_1, \ldots, M_K) = \frac{(b|a)_{M_k}}{(b)_{N_k}} \prod_{k=1}^{K} (H(k) S^{N_k}_{M_k,a}).$$

Now, this representation with the number of tables per topic M_k is converted to the representation with table indicators u_i that are assigned to each token and specify at which levels this token has a table contribution. E.g., if $u_1 = 0$, the first token contributes to the table count at all levels, whereas in the case of two levels and $u_1 = 2$ the token does not contribute to the table count, i.e. the customer sits at an existing table.

The joint posterior distribution of the hierarchical PDP given base distribution H_0 for the root node is now given by

$$p(z, u|H_0) = \prod_{j \geq 0} \left(\frac{(b|a)_{M_j}}{(b)_{N_j}} \prod_{k=1}^{K} S^{n_{jk}}_{m_{jk},a} \frac{m_{jk}!(n_{jk} - m_{jk})!}{n_{jk}!} \right),$$

where j is the index of the hierarchy level, N_j and M_j are the overall number of customers and tables for restaurant j and n_{jk} and m_{jk} are the respective counts for topic k.

To get the posterior distribution for a specific topic $z_i = k$ and table indicator $u_i = u$, application of the chain rule yields [20]

$$p(z_i = k, u_i = u|z_{-i}, u_{-i}, H) = \frac{p(z, u, H)}{p(z_{-i}, u_{-i}, H)} =$$

$$\prod_{j \in path} \frac{(b_j + a_j M_j)^{\delta_{M'_j \neq M_j}}}{(b_j + N_j)^{\delta_{N'_j \neq N_j}}} \left(\frac{S^{n'_{jk}}_{m'_{jk},a}}{S^{n_{jk}}_{m_{jk},a}} \right)^{\delta_{n'_{jk} \neq n_{jk} || m'_{jk} \neq m_{jk}}}$$

$$\frac{(m'_{jk})^{\delta_{m'_{jk} \neq m_{jk}}} (n'_{jk} - m'_{jk})^{\delta_{n'_{jk} - m'_{jk} \neq n_{jk} - m_{jk}}}}{(n'_{jk})^{\delta_{n'_{jk} \neq n_{jk}}}}.$$

Here the path consists of the restaurants in the hierarchy where the customer has a table contribution. The subscript $-i$ refers to all variables except the one with index i, N', M', n' and m' are the counts after adding the current token to the counts and N, M, n and m refer to counts before the addition of token i. The ratio of Pochhammer symbols $(x|y)_{N+1}/(x|y)_N$ reduces to $x + Ny$, whereas $(x)_{N+1}/(x)_N = \Gamma(x + N + 1)/\Gamma(x + N) = x + N$. S^n_m are generalized Stirling numbers of the first kind whose ratios can be efficiently precomputed and retrieved in $O(1)$.[1]

Finally, the sampling equations of the full 2-level HDP topic model for the joint sampling of topic and table indicator are as follows [20]. $rest$ is an abbreviation referring to all remaining variables.

If the topic is new for the root restaurant (table indicator is zero):

$$P(z_i = k_{new}, u_i = 0|rest) \propto \frac{b_0 b_1}{(M_. + b_0)(N_j + b_1)} P_{wk_{new}} \tag{8}$$

[1] See Buntine and Hutter [7] for an efficient way to compute ratios of these numbers. They can be precomputed once and subsequently retrieved in $O(1)$. Note that it may be necessary to store large values sparsely if the number of tokens in a restaurant becomes large.

If the topic is new for the base restaurant (e.g. a document), but not for the root restaurant (table indicator is one):

$$P(z_i = k, u_i = 1|rest) \propto \frac{b_1 M_k^2}{(M_k + 1)(M_. + b_0)(N_j + b_1)} P_{wk} \tag{9}$$

If the topic exists at the base restaurant and an already existing table is chosen (table indicator is two):

$$P(z_i = k, u_i = 2|rest) \propto \frac{S^{n^{jk}+1}_{m_{jk}}}{S^{n_{jk}}_{m_{jk}}} \frac{n_{jk} - m_{jk} + 1}{(n_{jk} + 1)(N_j + b_1)} P_{wk} \tag{10}$$

If the topic exists at the base restaurant and a new table is opened (table indicator is one):

$$P(z_i = k, u_i = 1|rest) \propto \tag{11}$$

$$\frac{b_1}{N_j + b_1} \frac{S^{n_{jk}+1}_{m_{jk}+1}}{S^{n_{jk}}_{m_{jk}}} \frac{m_{jk} + 1}{n_{jk} + 1} \frac{M_k^2}{(M_k + 1)(M_. + b_0)} P_{wk} \tag{12}$$

The prior term P_{wk} is calculated in the same way as for the standard LDA model:

$$P_{wk} = \frac{N_{wk} + \beta}{\sum_{w'} (N_{w'k} + \beta)} \tag{13}$$

In the above equations, b_0 is the hyperparameter for the root DP, b_1 is the hyperparameter for the lower level DP, M_k is the total number of tables for topic k, $M_.$ is the total number of tables, n_{jk} is the number of customers for topic k in restaurant j, N_{wk} is the total number of tokens for word w and topic k, and m_{jk} is the number of tables for topic k in restaurant j.

Summing up this section, nonparametric topic models and the most efficient sampling method were introduced. The main feature of nonparametric models is the ability to model topics or labels with different frequencies and to let the number of topics/labels adapt to the size of the dataset.

2.4 Dependency Topic Models

As a third dimension we consider whether or not label dependencies are modeled. This is a crucial feature of multi-label classifiers. For example, a text might have two labels, "Language" and "Programming". Maybe the corresponding text is about programming languages, meaning that there is some overlap between the two labels. This kind of dependency is probably not exhibited by labels such as "Dog" and "Matrices". A text about dogs is in all likelihood not about matrices, whereas a text about languages has a certain probability to also be about programming. Modeling dependencies therefore has the potential to improve the accuracy of multi-label classifiers.

In the streaming setting, large amounts of data have to be processed in a short time, which makes it more difficult to exploit such dependencies. While there is a large amount of training data available, it is not always the case that there is enough data for each label. Often there is a large number of rare labels that may benefit from additional dependency information. However, the classifier has to learn labels and their dependencies at the same time which can lead to errors that may affect performance.

2.5 Multi-label classification

Multi-label classification is the problem where each instance in a dataset is assigned one or several labels. It is to be distinguished from multi-class classification, which only assigns one out of multiple classes to each instance. Multi-label classification may also be seen as a special case of multi-label ranking, where each instance is associated with a ranking over the possible labels as well as a cut-off point which determines at which point to separate the negative from the positive labels [24].

Multi-label methods are commonly divided into algorithm-adaptation and transformation-based approaches [73]. The former directly adapt an algorithm to multi-label use, whereas the latter transform the problem into several single-label problems.

The most simple, yet efficient, transformation-based approach is the binary relevance method (BR [73; 6]). BR does not take dependencies between labels into account, but instead trains one classifier for each label separately. The predictions for the different labels are then combined into one multi-label prediction.

While BR is considered to be an efficient and scalable classifier, it still requires to learn one classifier per label, which can lead to very large models. Most multi-label algorithms in the literature are even more inefficient, many have a complexity that is quadratic in the number of labels (see e.g. Zhang and Zhou [82], Wicker *et al.* [77]), and therefore are not applicable in a large-scale setting. Recently, there have also been some approaches using deep learning and neural networks, but none of them scales well with very large label numbers [78; 25; 80; 49].

In light of these issues, a new line of work on so-called extreme mult-label classification has been developed in recent years. This work is concerned with datasets having several hundred thousand or even millions of labels and features. For example, FastXML (Fast eXtreme Multi Label) by Prabhu and Varma [56], an ensemble of decision trees, has prediction cost that is logarithmic in the number of labels.

Multi-label topic models can be considered to be more efficient than transformation-based approaches because they are algorithm-adaptation approaches that just train a single model for all labels. However, most multi-label topic models have not yet been applied in extreme multi-label classification settings although a lot of work has been done on developing inference algorithms for LDA that scale sub-linearly in the number of labels [37; 9; 12].

3. APPLICATIONS

Multi-label topic models have for example been applied in sociology [45] to answer questions such as "What terms do our categories reference?", "Have our categories changed over time?", or "Do certain groups have their own language and does it change over time?", among others. Ramage *et al.* [61] cluster web pages into semantic groups using a semi-supervised topic model called multi-multinomial LDA (MM-LDA) in which the labels are used as an additional input for the resulting clustering so that the topics are informed by the labels but do not correspond to them. They show that the inclusion of the labels improves the topic quality and the joint modeling of words and labels improves classification performance as compared to k-means.

Another line of work applies multi-label topic models on scientific writing such as papers or PhD theses. Papagiannopoulou *et al.* [53] employ multi-label LDA in a competition on large-scale indexing, where abstracts from biomedical scientific papers have to be tagged with their corresponding medical subject headings (MeSH). Ramage *et al.* [63] apply a multi-label topic on the PhD thesis abstracts for different universities that are associated with keywords in the Proquest UMI database. They then compare the topic vectors over time to find out which universities lean more towards the future and which universities are oriented more towards the past. One of the most well-known applications is the author-topic model [65] that assigns documents to authors and determines a distribution over topics for each author. This model is, however, not directly intended for multi-label classification. Johri *et al.* [33] use multi-label topic modeling to study the collaboration of scientists in computational linguistics with latent mixtures of authors.

The largest body of work consists of applications on Twitter data. Ramage *et al.* [59] use multi-label topic modeling to characterize a user's data stream on Twitter. Tweets are labeled into different categories according to substance, status, style, social or other. They aim to provide a way to recommend tweets according to different dimensions and characterize them by different criteria. Cohen and Ruths [22] use multi-label topic models to classify user's political orientation on Twitter. Quercia *et al.* [57] develop a model called Tweet-LDA that is used to assign labels to user profiles, to re-rank user feeds and to suggest new users to follow. Bhattacharya *et al.* [3] infer user interest on Twitter using hashtags on a scale of millions of users. Mukherjee and Liu [47] use a multi-label topic model to extract topics from text corpora with user guidance, meaning the users can specify seed words for the sentiment aspects of topics they wish to extract.

The above (probably incomplete) list shows that there is a diverse set of possible applications for multi-label topic models that is sure to keep growing.

4. DIFFERENT MULTI-LABEL TOPIC MODELS

This section gives an overview of the different kinds of multi-label topic models that belong to the different dimensions as well as some closely related non-multi-label topic models. The multi-label topic models are listed in the first section of Table 1. In Sections 4.2 and 4.3 Labeled LDA and Dependency-LDA are introduced in detail. Li *et al.* [40] introduce Frequency-LDA (FLDA) and Dependency-Frequency-LDA (DFLDA) that more or less correspond to Prior-LDA and Dependency-LDA by Rubin *et al.* with slight modifications in the training procedure that lead to improvements. Zhang *et al.* [83] introduce labeled LDA with function terms (LF-LDA), a topic model that extracts noisy function terms from textual data to improve the performance of multi-label classification. Padmanabhan *et al.* [51] propose Multi-Label Presence-Absence LDA with Crowd (ML-PA-LDA-C), a multi-label topic model that accounts for multiple noisy annotations from the crowd. Fast Dep.-LLDA, hybrid HDP and stacked HDP are introduced in Sections 4.4, 4.5 and 4.6.

The Correlated Labeling Model by Wang *et al.* [76] is introduced in Section 4.1. Hierarchically supervised latent Dirichlet allocation [55] (HSLDA) is a multi-label topic model that extends supervised LDA (sLDA [44]) to consider label dependencies. It is trained using Gibbs sampling and uses a non-parametric prior for the document-topic distributions trained by the direct-assignment sampler of Teh *et al.* [71]. Its main feature is the capability to model label hierarchies, i.e. labels that come from a predefined taxonomy.

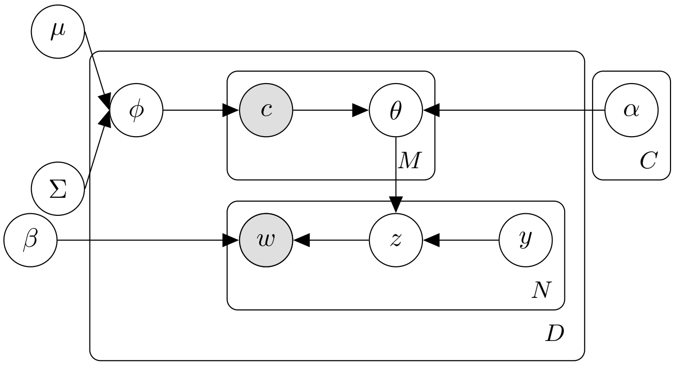


Figure 4: The graphical model of the Correlated Labeling Model (CoL).

Existing supervised models include Supervised LDA [44], Dirichlet-multinomial regression (DMR) [46], semi-supervised hierarchical topic model (SSHLDA), DiscLDA [35] and MedLDA [84]. However, these models are single-label classification or regression models and not usable in a multi-label setting.

There exist a number of methods that model dependencies between topics, but are (at least partially) unsupervised. Among these is the author-topic model [65] which assigns an author and a topic to each word, such that one document is modeled as a mixture of topics and each author is associated with a topic distribution. In the partially labeled topic model by Ramage *et al.* [62] each label is divided into several topics. Another method that models topic dependencies is the Pachinko allocation model (PAM, see Figure 5b) [39]. It assigns topics on two different hierarchy levels in such a way that each super-level topic is associated with a distribution over sub-level topics and each document has a distribution over both super- and sub-topics. A nonparametric version of this model is proposed by Li [38]. Nonparametric PAM (nPAM) is based on HDPs that model topic correlations. Another model based on nested DP called cHDP is proposed by Shimosaka *et al.* [69]. This model requires that each document is assigned to exactly one super-topic. The proposed learning procedure is based on variational Bayes. The generative process is defined as follows:

$$G_0 \sim DP(b_0, H), Q \sim DP(\alpha, DP(\beta, G_0)), G_d \sim Q$$

As the second equation shows, here one DP is nested into another DP as described by Rodriguez *et al.* [64]. Another model that allows topic sharing is proposed by Salakhutdinov *et al.* [67]. It is a supervised model, however, it does not allow multiple labels per document. Each document is assigned one label using a nested Chinese restaurant distribution. Then the whole document is sampled according to the document's label. The correlated topic model [36] is unsupervised and models correlations between topics using a logistic normal distribution. However, the model is complicated since the normal distribution is not conjugate to the multinomial distribution. The most important multi-label topic models and their relation to some of the mentioned related models are now discussed in more detail.

4.1 Correlated Labeling Model

Wang *et al.* [76] develop a model called CoL (Correlated Labeling Model). It models each label as a distribution over latent topics. A variational learning method is proposed and the results show that this model achieves a slightly better F-measure on the tested datasets than SVMs.

In this model, there are D documents, C classes and V words overall and one document consists of M classes and N words. ϕ is the document-specific distribution of classes, θ is the topic distribution for each class. μ and Σ are the mean and covariance of the Normal distribution. The graphical model is shown in Figure 4 and the generative process is defined as follows:

1. Sample $\phi \sim N(\mu, \Sigma)$
2. For each class/label c_m, $m \in \{1, 2, 3, \ldots, M\}$
   (a) Sample $c_m \sim Mult(\frac{\exp(\phi)}{1+\sum_i \exp(\phi_i)})$
   (b) Sample topic distribution $\theta_m \sim Dir(\alpha|c_m)$
3. For each word w_n, $n \in \{1, 2, 3, \ldots, N\}$
   (a) Sample class $y_n \sim Uniform(1, 2, 3, \ldots, M)$
   (b) Sample topic $z_n \sim Mult(\theta|y_n)$
   (c) Sample word $w_n \sim Mult(\beta_{z_n})$

They note that the model is especially good at predicting rare labels in unbalanced datasets. While this model has an efficient training procedure, the inference process is expensive for large numbers of labels and a heuristic has to be used.

4.2 Labeled LDA

Labeled LDA (LLDA) is introduced by Ramage *et al.* [60]. In this work, the collapsed Gibbs sampling topic model by Griffiths and Steyvers [26] is extended by introducing document labels Λ_d that are generated from a Bernoulli distribution for each topic k.

The model is defined in a slightly different way in Rubin *et al.* [66] although in practice the training procedure is the same. Here, the model is called Flat-LDA and does not include a generative procedure for the set of labels via Bernoulli variables. During training of both models, the Bernoulli variables do not play any role. In practice, both models correspond to LDA with a restriction of sampling only from the document labels during training. If each document is only assigned a single label, the model reduces to Naive Bayes [60].

Ramage *et al.* and Rubin *et al.* propose collapsed Gibbs sampling as a training algorithm, however, this is only one potential variant of the model. Since the idea of Flat-LDA is simply to replace unsupervised topics with labels, the same idea can be applied to topic models with other training methods as well:

1. Variational inference can be used as an alternative training algorithm (see, e.g., Papanikolaou *et al.* [54]). The disadvantage is that the algorithm is biased. Also it is more difficult to implement sparse updates. On the positive side, variational inference makes it possible to train the model online.
2. Nonparametric topic models are another alternative for supervised training (see Section 2.3). These hierarchical Dirichlet process topic models provide an asymmetric topic/label prior. This model may also be trained using different algorithms:
   (a) Variational Bayes
   (b) Gibbs sampling
   (c) Hybrid Variational-Gibbs (see Section 4.5)
3. More complex hierarchical topic models may be used. In particular,
   (a) the author-topic model [65]: Gibbs sampling is used for training.
   (b) Dependency-LDA (Section 4.3): Gibbs sampling is used for training.
   (c) Fast-Dependency-LDA (Section 4.4): This model can be trained with Gibbs sampling or variational inference.
   (d) Stacked HDP (Section 4.6): Gibbs sampling is used for training.

This summary shows that the simple idea of supervised topic models has many variants depending on the one hand on the exact model that is used (parametric or nonparametric, simple flat or hierarchical) and depending on the other hand on the training algorithm for the chosen model.

4.3 Dependency LLDA

Dependency-LDA (Dep.-LLDA, see Figure 5a) is a topic model for multi-label classification due to Rubin *et al.* [66]. The idea of Dep.-LLDA is to learn a model with two types of latent variables: the labels and the topics. The labels are associated with distributions over words, while the topics are associated with distributions over labels. The topics capture dependencies between the labels, since the frequent labels in one topic are labels that tend to co-occur in the training data. The notation for the following is summarized in Table 2.

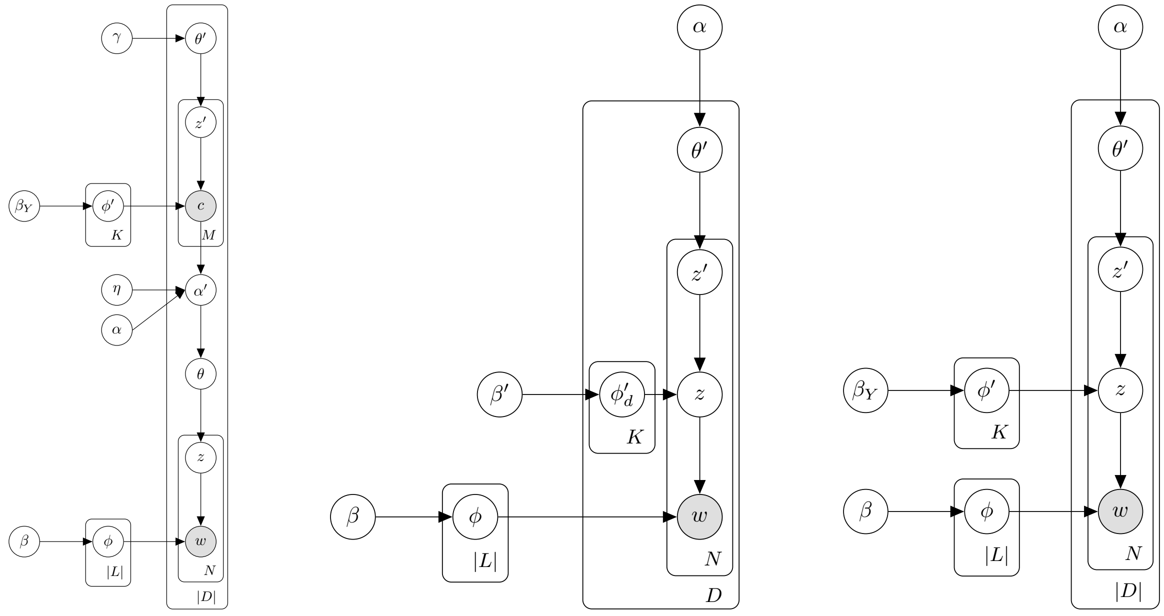
The generative process is given as follows:

1. For each topic $k \in 1, \ldots, K$ sample a distribution over labels $\phi'_k \sim Dirichlet(\beta_Y)$
2. For each label $y \in L$ sample a distribution over words $\phi_y \sim Dirichlet(\beta)$
3. For each document $d \in D$:
   (a) Sample a distribution over topics $\theta' \sim Dirichlet(\gamma)$
   (b) For each label token in d:
      i. Sample a topic $z' \sim Multinomial(\theta')$
      ii. Sample a label $c \sim Multinomial(\phi'_{z'})$
   (c) Sample a distribution $\theta \sim Dirichlet(\alpha')$
   (d) For each word token in d:
      i. Sample a label $z \sim Multinomial(\theta)$
      ii. Sample a word $w \sim Multinomial(\phi_z)$

The Gibbs sampling equations for the labels z and the topics z' are given by:

$$P(z = y|w, z_{-i}, z'_{-i}) \propto \frac{n_{-wy} + \beta}{n_{-\cdot y} + |W|\beta}(n_{-dy} + \alpha') \qquad (14)$$

$$P(z' = k|c = y, c_{-i}, z'_{-i}) \propto \frac{n_{-yk} + \beta_Y}{n_{-\cdot k} + |L|\beta_Y}(n_{-dk} + \gamma),$$



(a) Dep.-LLDA from Rubin *et al.* [66]

(b) PAM [39]

(c) Fast-Dep.-LDA

Figure 5: The graphical models of the original Dep.-LLDA by Rubin *et al.* [66], PAM [39], and Fast-Dep.-LDA.

where n_{-wy} is the number of times word w_i occurs with label y. $n_{-\cdot y}$ is the number of times label y occurs overall, n_{-dy} is the number of times label y occurs in the current document, n_{-yk} is the number of times label y occurs with topic k, $n_{-\cdot k}$ is the number of times topic k occurs overall and n_{-dk} is the number of times topic k occurs in document d. The subscript $-$ indicates that the current token is excluded from the count. The connection between the labels and the topics is made through the prior α'. To calculate α', Rubin *et al.* propose to make use of the label tokens c. According to these M_d label tokens, α' for document d is calculated as follows:

$$\alpha' = [\eta \frac{n_{d1}}{M_d} + \alpha, \eta \frac{n_{d2}}{M_d} + \alpha, ..., \eta \frac{n_{d|L|}}{M_d} + \alpha],$$

where n_{dy} is set to one during training, and to the number of times a particular label is sampled during testing, and η and α are parameters.

During testing however, instead of taking M samples and calculating α' as described above, a so-called "fast" inference method is used. This means the sampled z variables are used directly instead of c, and α' is calculated as follows:

$$\alpha' = \eta \hat{\theta}' \hat{\phi}' + \alpha,$$

where $\hat{\phi}$ and $\hat{\theta}$ are the current estimates of ϕ and θ. During training, since the labels of each document are given, ϕ and ϕ' are conditionally independent which allows separate training of both parts of the topic model. Finally, they apply a heuristic to scale α' according to the document length during testing. Overall, Dep.-LLDA is an effective and efficient method for multi-label classification.

4.4 Fast Dependency LLDA

Fast Dep.-LLDA [14] (see Figure 5c) is based on Dependency LLDA, but has a simpler model structure and thus can be trained online using variational Bayes. Fast-Dep.-LLDA and Dep.-LLDA have strong similarities. The main difference is the omission of θ and α' in Fast-Dep.-LLDA. Both models learn the label dependencies through the label-topic distributions ϕ'. Dep.-LLDA passes the dependency information down via the label-prior α' and the label distribution θ. Fast-Dep.-LLDA, however, takes the more direct approach of generating the labels from ϕ' directly instead of using the intermediary distribution θ (see the graphical models in Figures 5a and 5c). Thereby Fast-Dep.-LLDA avoids a couple of heuristics that are employed by Dep.-LLDA:

1. Dep.-LLDA employs a fast inference method that is empirically found to be faster and to lead to more accurate results.

2. The calculation of the parameter α' itself involves two parameters η and γ that are determined heuristically by the authors.

3. During evaluation the parameter α' is scaled according to the document length.

4. During evaluation, the label tokens c and in particular the number of labels are unknown. To circumvent this problem, the authors replace the label tokens c by the label indicator variables z during testing, thereby assuming that the number of labels is equal to the document length.

The full generative process of Fast-Dep.-LLDA is given in Table 3. Each document is only associated with one document-specific distribution θ' over the topics. In comparison, Dependency-LDA has two document-specific distributions, θ and θ', where θ is a label distribution. The label distribution θ is implicitly contained in Fast-Dep.-LLDA and can be obtained by multiplying the document-specific topic distributions θ' with the global topic-label distributions ϕ'.

Table 2: Notation for Dep.-LLDA Gibbs sampling models

V	words
K	number of topics
L	labels
D	documents
N_d	number of words in document d
i,j,y,k	indices over word tokens, documents, labels and topics resp.
z,c	label indicator variables
z'	topic indicator variables
α,β,β_Y,γ	hyperparameters (see generative processes)
ϕ, ϕ'	word-label distribution, label-topic distribution
θ,θ'	document-label, document-topic distribution
n_{-wy}	count for word w with label y excluding the current token
$n_{-\cdot y}$	count for label y excluding the current token
n_{-dy}	count for label y in document d excluding the current token
n_{-yk}	count for label y with topic k excluding the current token
$n_{-\cdot k}$	count for topic k excluding the current token
n_{-dk}	count for topic k in document d excluding the current token

Table 3: The generative process of Fast-Dep.-LLDA

For each topic $k \in 1, \ldots, K$
- sample a distribution over labels $\phi'_k \sim Dirichlet(\beta_Y)$
For each label $y \in L$
- sample a distribution over words $\phi_y \sim Dirichlet(\beta)$
For each document $d \in D$:
1. Sample a distribution over topics $\theta' \sim Dirichlet(\alpha)$
2. For each token in d:
2.1 Sample a topic $z' \sim Multinomial(\theta')$
2.2 Sample a label $z \sim Multinomial(\phi'_{z'})$
2.3 Sample a word $w \sim Multinomial(\phi_z)$

From the graphical model and the generative process, the joint distribution of Fast-Dep.-LLDA is given by

$$P(w, z, z') = P(w|z, \phi)P(z|z', \phi')P(z'|\theta').$$

To obtain a collapsed Gibbs sampler, ϕ, ϕ', and θ' have to be integrated out from the three conditional probabilities respectively. The integrals can be performed separately as in Griffiths and Steyvers [26], resulting in the following conditional distribution for the latent variables z and z':

$$P(z = y, z' = k|w, z_{-i}, z'_{-i}) \propto \\ \frac{n_{-wy} + \beta}{n_{-\cdot y} + |V|\beta} \frac{n_{-yk} + \beta_Y}{n_{-\cdot k} + |L|\beta_Y}(n_{-dk} + \alpha)$$

This sampling equation results in a blocked Gibbs sampler that samples two variables at a time instead of just one: each word is assigned a topic and a label. They propose the use of a basic Gibbs sampler that only samples one variable at a time instead. This may have the disadvantage of making successive samples more dependent [4], but the sampling complexity is reduced from $O(K \cdot |L|)$ to $O(K + |L|)$.
The corresponding sampling equations for the alternate sampling of labels and topics are given as follows. Given z', the equation for sampling z is

$$P(z = y|w, z' = k, z_{-i}, z'_{-i}) \propto \frac{n_{-wy} + \beta}{n_{-\cdot y} + |V|\beta}(n_{-yk} + \beta_Y). \tag{15}$$

The sampling equation for z' follows from $P(z'|z) = \frac{P(z,z')}{\sum_{z'} P(z,z')}$, where $P(z, z') = P(z|z', \phi')P(z'|\theta')$. The same steps as for sampling z apply, giving

$$P(z' = k|z = y, z_{-i}, z'_{-i}) \propto \frac{n_{-yk} + \beta_Y}{n_{-\cdot k} + |L|\beta_Y}(n_{-dk} + \alpha).$$

Instead of training the complete model at once, a greedy layer-wise training procedure is proposed. This leads to the following equation for sampling label assignments z during training of Fast-Dep.-LLDA:

$$P(z = y|w, z' = k, z_{-i}, z'_{-i}) \propto \frac{n_{-wy} + \beta}{n_{-\cdot y} + |V|\beta}.$$

The model is guaranteed to converge to the optimum given the chosen parameters. The greedy model may be viewed as letting $\sum \beta_Y \to \infty$ which means the Dirichlet becomes a uniform distribution in case of symmetric β_Y. Greedy training corresponds to choosing the most extreme parameter value for β_Y, which leads to the second term vanishing from Equation 15 completely. Empirically, it is the case that on all tested multi-label datasets the convergence is better using greedy training than non-greedy training.

4.4.1 Online Fast-Dep.-LLDA (SCVB-Dep.)

The online version of Fast-Dep.-LLDA is called SCVB-Dependency. For this, a method similar to the stochastic collapsed variational Bayes (SCVB) method by Foulds *et al.* [23] is developed. The fully factorized variational distribution of Fast-Dep.-LLDA is given by

$$q(z, z', \theta', \phi, \phi') = \prod_{ij} q(z_{ij}|\gamma_{ij}) \prod_{ij} q(z'_{ij}|\gamma'_{ij}) \prod_{j} q(\theta'_j|\tilde{\alpha_j})$$

for tokens i and documents j.
In the equation, an additional variational parameter γ' is introduced for the topic assignments z'. However, computing the updates for γ and γ' separately would lead to unnecessary

computational effort. Instead an intermediate value λ_{wyk} is computed, which corresponds to the expectation of a joint occurrence of word w, label y and topic k which can be expressed in terms of an expectation of the indicator function $\mathbb{1}$, which is one if these values occur together and otherwise zero: $\mathbb{E}[\mathbb{1}[w_i = w, y_i = y, k_i = k]]$, where i is the index of the token.

For each token (the ith word in the jth document) λ_{ijyk} is calculated for label y and topic k, where during training λ only has to be calculated for the labels of the document and should be set to zero for all other labels.

$$\lambda_{ijyk} :\propto \lambda^W_{ijy} \lambda^T_{ijyk}$$

$$\lambda^W_{ijy} :\propto \frac{N^{\phi}_{w_{ij},y} + \eta_w}{N^Z_y + \sum_w \eta_w}$$

$$\lambda^T_{ijyk} :\propto \frac{N^{\phi'}_{y_{ij},k} + \eta_y}{N^{Z'}_k + \sum_y \eta_y} (N^{\theta'}_{jk} + \alpha),$$

where N^Z is a vector storing the expected number of words for each label. N^{ϕ} is the expected number of tokens for words w and labels y in the whole corpus. Additionally, $N^{Z'}$ stores the expected number of tokens for each topic, $N^{\phi'}$ is the expected number of tokens for labels y and topics k, and $N^{\theta'}_j$ is the expected number of words per topic, only for document j.

Because greedy layer-wise training is used, the two parts of the model can be trained separately whereas during testing the full model has to be used. The first layer treats every word as an input token and updates the word-label distribution based on λ^W, whereas the second layer treats each label assignment as an input token and learns the label-topic distributions based on λ^T. Since the model is supposed to be trained online, it is not possible to wait for the greedy algorithm to learn the first layer before moving on to the second layer. Therefore, the input probabilities of the second layer are initialized by using the true labels. In this way, both layers can be trained simultaneously while not having to view any document more than once.

4.4.2 *Discussion*

Fast-Dep.-LLDA can be trained using a batch method based on Gibbs sampling or using an online method based on variational Bayes. The method was shown to perform especially well on rare labels, due to the modelling of the label dependencies and to be scalable to large datasets where it converges much faster than the batch methods.

4.5 Nonparametric topic model

One shortcoming of Fast-Dep.-LLDA (Section 4.4) is that the different frequencies of the topics and labels are not modeled, i.e. they are given a symmetric prior. This problem is addressed by the hierarchical Dirichlet process (HDP), which is used to train nonparametric topic models. HDP topic models are nonparametric in the sense that the number of topics is automatically determined from the data. However, their main advantage is the modeling of different topic frequencies, thus leading to better representations of the data. Therefore, the idea of labeled LDA can be extended to use HDPs instead of standard LDAs.

HDP can be made supervised in the same way as LDA: by assigning one topic to each label. Analogously to LLDA, the modification of HDP for multi-label classification is called Labeled HDP (LHDP) [8]. LHDP allows to take different label frequencies into account. Since the number of labels is fixed, a truncated HDP can be used.

As proposed by Li *et al.* [37], the sampling equations may be rewritten as $\frac{\beta}{\sum(N_{w'k}+\beta)} \cdot X + \frac{N_{wk}}{\sum(N_{w'k}+\beta)} \cdot X$, where X stands for the remaining part of the equation. The first part can be stored and sampled from in $O(1)$, since repeated samples from the same distribution are feasible in $O(1)$, adding a Metropolis-Hastings acceptance step to account for the difference with the updated counts. The second part only has to be computed for the topics that occur with word w. Therefore, the sampling complexity is reduced to amortized $O(K_w)$, where K_w is the number of topics that occur with word w.

Burkhardt and Kramer [12] employ the idea of Li *et al.*'s alias-sampling of storing a stale part of the probability distribution and sample from it in $O(1)$, correcting the difference with a Metropolis-Hastings acceptance step. However, in contrast to the original alias-sampling, the hierarchical structure of HDPs is exploited. Recall that the conditional probability for topic k is given by:

$$P(z = k|rest) = P(z = k, u = 0|rest) + \\ P(z = k, u = 1|rest) + P(z = k, u = 2|rest).$$

The last term is usually sparse since it is only non-zero for all topics that already have a table in the corresponding restaurant. The second part is dense, but changes rather slowly since the overall topic distribution changes much slower than the topic distribution within a document or label. Therefore, instead of dividing the distribution according to the language model term $\frac{N_{wk}+\beta}{\sum_{w'}(N_{w'k}+\beta)}$, it is divided according to the table indicator u, thus yielding a sampler that runs in $O(K_d)$ instead of $O(K_w)$ (in case of a standard two-level HDP).

The described method reduces the sampling complexity to $O(K_j)$, but, as can be inferred from Equations 9 and 11, q_{jw} depends on document j. This means the global topic distribution has to be saved separately for every document. The same is true for the alias-sampler by Li *et al.* [37], which puts a restriction on the size of the used datasets, since a topic distribution has to be saved for every single document. Therefore, Burkhardt and Kramer [12] propose a method that instead only uses a single global distribution.

The main idea is to assume for each topic that it does not exist in the document and save the resulting distribution q^e_w for an empty pseudo document e. This can be understood as replacing Equation 11 with Equation 9. In case a topic is sampled from this distribution that exists in the current document, it is discarded and a new one is drawn from the same distribution.

$$\tilde{p}_{jw}(k, u') := P(z = k, u = u'|rest)\mathbb{1}[n_{jk} > 0] \ ,$$

where $\mathbb{1}[n_{jk} > 0]$ is one if the number of tokens in document-restaurant j associated with topic k is at least one and zero otherwise. Accordingly, the normalization sum is

$$\tilde{P}_{jw} = \sum_k \sum_u \tilde{p}_{jw}(k, u).$$

An amount Δ_j needs to be subtracted from the normalization sum Q_w, which is different for each document j and accounts for the topics that are present in document j and would be rejected if drawn from distribution q. It is called the discard mass Δ and defined as

$$\Delta_j := \sum q_k^e \mathbb{1}[n_{jk} > 0].$$

Δ_j is computed in $O(K_j)$ time and therefore does not increase the overall computational complexity. The modified normalization sum is accordingly given by $\tilde{Q}_{jw} = Q_w - \Delta_j$, where $Q_w = \sum q_w^e$.

The difference to the true distribution needs to be corrected using Metropolis-Hastings (MH). The modified MH acceptance ratio is given by:

$$\pi = \frac{P(z=t, u=u_t|rest)}{P(z=s, u=u_s|rest)} \cdot \begin{cases} \tilde{P}_{jw}\tilde{p}_{jw}(s), & \text{if } n_{js} > 0 \\ Q_w q_w^e(s), & \text{otherwise} \end{cases} \cdot \begin{cases} \frac{1}{\tilde{P}_{jw}\tilde{p}_{jw}(t)}, & \text{if } n_{jt} > 0 \\ \frac{1}{Q_w q_w^e(t)}, & \text{otherwise} \end{cases}$$

Overall, there seems to be a slight advantage of the nonparametric method in large-scale experiments. Burkhardt [8] finds that nonparametric methods fare best on larger datasets where the number of labels is high. Given the right hyperparameters, the nonparametric method is able to perform well in the supervised setting, especially on frequent labels as compared to the parametric method, which performs better on rare labels.

4.6 Nonparametric dependency topic model

The previous section introduced LHDP, a nonparametric multi-label topic model, which can be trained on streaming data, but does not make use of label dependencies. In this section, stacked HDP (sHDP) is introduced [12], a model that extends Fast-Dep.-LLDA to use HDPs so that we have a model in which two HDPs are stacked on top of each other. In the literature there exist two models with a similar structure, albeit they are just employed in unsupervised settings. First, there is a variant of nested DPs, called coupled DP mixtures (cHDP), by Shimosaka *et al.* [69]. This model groups the documents into topics in addition to clustering them by labels (or rather sub-topics, since the model is unsupervised). cHDP is restricted in that each document belongs to exactly one topic. Second, there is a hierarchical topic model called nonparametric Pachinko allocation model (PAM), which associates a distribution over labels and topics with each document so that each document may belong to several labels and topics (see Fig. 6c). This, however, leads to a complex model with a three-level HDP and having to save document-specific distributions over topics as well as labels [38].

The third option is less complex than option two and does not have the restriction of option one. It is a combination of two two-level HDPs which are not nested as in option one, but rather stacked. This means that the word-tokens are clustered by labels and the labels are further clustered into different topics. Therefore, the model is called stacked HDP (sHDP). To make the model applicable in large-scale settings, the Alias sampling method introduced in the previous section is used. sHDP models a potentially infinite number of super-topics z', each of which is associated with a distribution over all sub-topics or labels. Thus the same sub-topic may appear in multiple super-topics. This allows the modeling of topic correlations. Additionally, sHDP is nonparametric, which allows the number of sub- and super-topics to be automatically determined from the data. Using Gibbs sampling each word-token is associated with a sub-topic and a super-topic that can be sampled independently and that only depend on the variables in their respective Markov-blanket.

The graphical model of sHDP is shown in Fig. 6a. The generative process is defined as follows:

- A distribution θ'_0 over super-topics is sampled from a DP
- A distribution ϕ'_0 over sub-topics is sampled from a DP
- For each super-topic k':
    - a distribution over sub-topics $\phi'_{k'}$ is sampled from a DP with base distribution ϕ'_0
- For each sub-topic k:
    - a distribution over words ϕ_k is drawn from a symmetric Dirichlet distribution
- For each document:
    - a distribution θ' over super-topics is sampled from a DP with prior θ_0
    - For each token in the document:
        - ∗ a super-topic z' is sampled from the document specific distribution over super-topics θ'
        - ∗ a sub-topic z is sampled from the distribution over sub-topics $\phi'_{z'}$ associated with super-topic z'
        - ∗ a word w is sampled from the word-topic distribution ϕ_z associated with sub-topic z

$$\begin{aligned} &\theta'_0|b_0, H \sim DP(b_0, H), && \theta'|b_1^{(0)}, \theta'_0 \sim DP(b_1^{(0)}, \theta'_0) \\ &z'|\theta' \sim Mult(\theta') && \\ &\phi'_0|b_0, H \sim DP(b_0, H), && \phi'|b_1^{(1)}, \phi'_0 \sim DP(b_1^{(1)}, \phi'_0) \\ &z|\phi'_{z'}, z' \sim Mult(\phi'_{z'}) && \\ &w|\phi_z, z \sim Mult(\phi_z), && \phi \sim Dirichlet(\beta) \end{aligned}$$

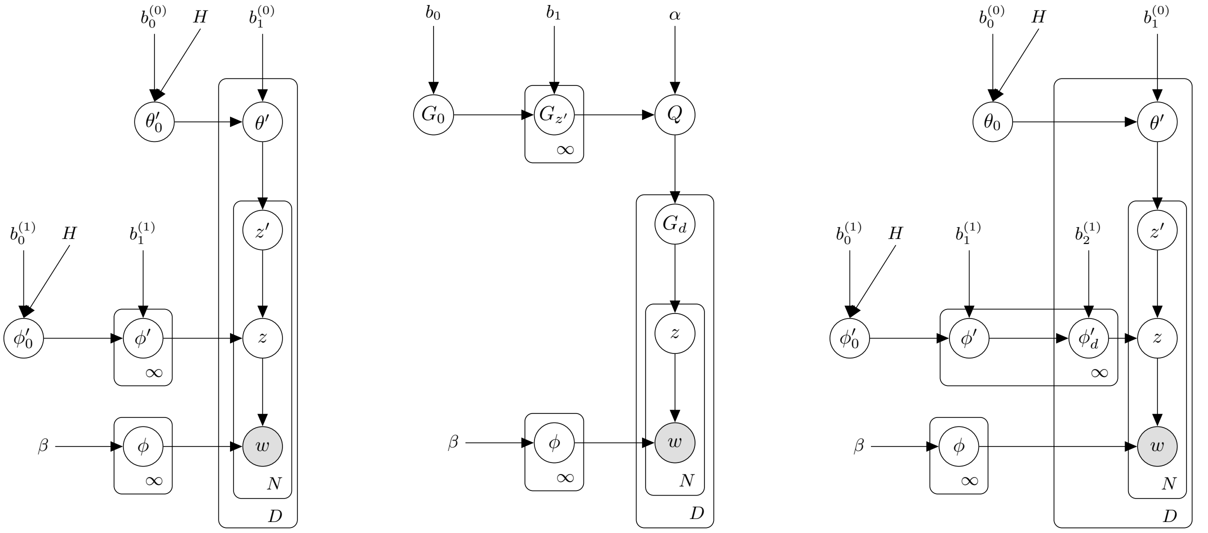
We can see from the above that the model corresponds to two two-level HDPs "stacked" on top of each other.

The sampling process is divided into two steps: First, z'_i is sampled conditioned on all z'_j with $i \neq j$, and z. Second, z_i is sampled conditioned on all z_j with $i \neq j$, z'_i, and w. In Equations 8 to 11 this is summarized as $rest$ for brevity. Since ϕ' is sampled from a DP and ϕ is sampled from a Dirichlet distribution, the equations for both steps are slightly different in one term, namely P_{wk}.

When no alias-sampling is used, the sampling equations for sampling the sub-topics k are equivalent to Equations 8 to 13. When sampling the super-topics Equations 8 to 11 are used, but P_{wk} is now given by:

$$P_{wk} = \sum_{u'} P'(z = w, u = u'|rest),$$

where w in this case corresponds to the sub-topic and k corresponds to the super-topic. P' is calculated using equations 8 to 11 disregarding the prior term given by 13. P_{wk} therefore



(a) Stacked HDP consists of two two-level HDPs, one for the topic distributions θ', and one for the label distributions ϕ'.

(b) The coupled HDP model allows for sub-topics to be shared among all super-topics. However, each document has to belong to exactly one super-topic.

(c) Wei Li's nonparametric PAM: In contrast to sHDP, here we have a 3-level HDP consisting of ϕ_0', ϕ', and ϕ_d'.

Figure 6: The graphical model of sHDP compared to two alternative models, the coupled HDP (cHDP) model by Shimosaka *et al.* [69] and the nonparametric PAM model by Wei Li [38]. sHDP is a simplified model with a more effective sampling procedure.

dataset	#labels	#documents
Reuters-21578[2][32]	90	12,902
bibtex [34]	159	7,395
delicious [75]	983	16,105
EUR Lex [41]	3,955	19,314
Ohsumed [68]	11,220	13,929
Amazon[3][43]	13330	1,493,021
BioASQ[4]	28,863	14,200,259

Table 4: A list of commonly used multi-label text datasets.

corresponds to the summed probability mass for sub-topic w given super-topic k.

The efficient sampling method introduced in the previous section is applicable in Stacked HDP at the sub-level as well as the super-level. At the sub-level the prior probability for the sub-topics is expected to change slowly relative to the probability of the sub-topics inside a given super-topic restaurant. If the actual probability estimates are used during training, the Gibbs sampler has a tendency to get stuck in local minima and less frequent labels are not sampled for many iterations. To alleviate this problem, a uniform document-label distribution is used during training similar to the inference procedure for Fast-Dep.-LLDA.

Burkhardt and Kramer [12] report a prediction performance for sHDP that is especially good on micro-averaged measures which indicate the performance on frequent labels. This shows that the model successfully models label frequencies and prefers frequent labels during prediction as well.

5. COMMON MULTI-LABEL TEXT DATASETS

An overview of some of the most common multi-label text datasets is given in Table 4. The `Reuters-21578` corpus consists of news stories that appeared on the Reuters newswire in 1987. The `bibtex` dataset is collected from the Bibsonomy system, which is a social bookmarking and publication-sharing system. Users store and organize bookmarks and BibTeX entries by assigning tags. `EUR Lex` is a dataset of legal documents concerning the European Union. It is hand annotated with almost 4,000 labels. The `Ohsumed` dataset[5] is a subset of MEDLINE medical abstracts that were collected in 1987 and that have 11,220 different human-assigned MeSH descriptors. The `Amazon` dataset consists of more than one million product reviews, annotated with corresponding product categories. The original dataset is available from `http://manikvarma.org/downloads/XC/XMLRepository.html` under the name AmazonCat-13K. This repository contains several more datasets of a similar nature. The `BioASQ` dataset contains article abstracts from the PubMed database. It is part of a yearly competition and updated every year. Currently it consists of over 14 million abstracts that are labeled with their corresponding MeSH categories.

6. PERFORMANCE EVALUATIONS

As multi-label classifiers, multi-label topic models are evaluated using standard multi-label classification measures. As suggested by e.g. Tsoumakas *et al.* [74], multi-label evaluation measures, e.g. the F-measure, can be computed as

[2] `http://trec.nist.gov/data/reuters/reuters.html`

[3] `http://manikvarma.org/downloads/XC/XMLRepository.html`

[4] `http://participants-area.bioasq.org/general_information/Task7a/`

[5] `http://trec.nist.gov/data/t9\_filtering.html`

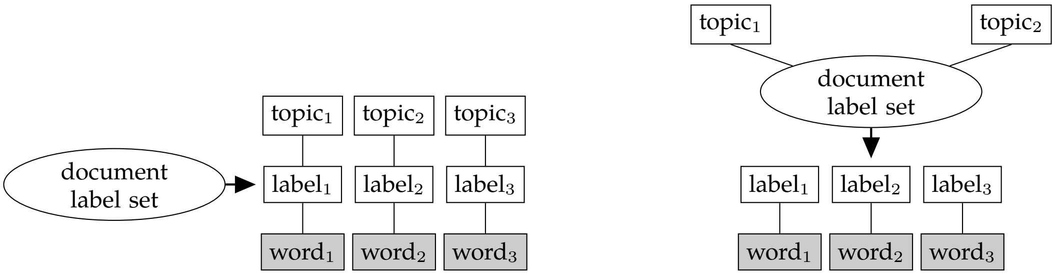


Figure 7: Illustration of the difference between Stacked HDP (left) and Dependency-LDA (right). The labels are drawn from the document label set in both cases. Stacked HDP samples one topic for each word/label token, whereas Dependency-LDA samples one topic for each label in the document label set. The white rectangles are sampled variables.

example-based or label-based measures. Label-based measures are further divided into micro- and macro-averaged measures. Additionally, rank-based measures such as area under the ROC curve are computed based on the ranking of labels instead of the binary predictions.

It is also possible to examine the topic coherence of the learned topics and the perplexity of the model, the most common measures in unsupervised topic modeling. This way, we might identify topics that do not have enough training data or where the training data is of low quality, independent of the classification performance. Theoretically it is possible that a label is predicted well by the model, but the topic coherence is low and does not correspond to what a human annotator would expect. This might be due to bad annotations that do not fit the given corpus well.

The per-word perplexity is calculated from the ELBO as

$$\exp\left(-\frac{1}{N_w}\sum_{d}^{D}\log p(d)\right),$$

where N_w is the number of words in the corpus. The topic coherence may be calculated following Srivastava and Sutton [70] using the normalized pointwise mutual information (NPMI), averaged over all pairs of words of all topics, where the NPMI is set to zero for the case that a word pair does not occur together. The NPMI for topic t is given as follows:

$$\text{NPMI}(t) = \sum_{j=2}^{N}\sum_{i=1}^{j-1}\frac{\log\frac{P(w_j,w_i)}{P(w_i)P(w_j)}}{-\log P(w_i,w_j)},$$

where N is the number of words in topic t, w_i is the ith word of topic t and $P(w_i, w_j)$ is the probability of words w_i and w_j occurring together in the test set, which is approximated by counting the number of documents where both words appear together and dividing the result by the total number of documents.

As an additional measure, Burkhardt and Kramer [15] introduce the topic redundancy measure, which corresponds to the average probability of each word to occur in one of the other topics of the same model. The redundancy for topic k is given as

$$R(k) = \frac{1}{K-1}\sum_{i=1}^{N}\sum_{j\neq k}P(w_{ik}, j),$$

where $P(w_{ik}, j)$ is one if the ith word of topic k, w_{ik}, occurs in topic j and otherwise zero, and $K-1$ is the number of topics excluding the current topic.

7. LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

Multi-label topic models have many advantages, but also important limitations. For example, given datasets with a limited amount of labels that are all sufficiently represented in the training data, they cannot outperform simple Binary Relevance classifiers using SVMs in terms of the multi-label classification performance. Thus, pure classification scenarios are not their purpose. They are applied when something beyond a suggested labeling is required such as semi-supervised learning, a semantic interpretation of the learned labels, a grouping of labels or explicit priors that are based on label frequencies or human input.

Future research directions depend on this modeling flexibility that may allow to apply it in dynamic contexts where changes in the training data and changes in modeling requirements are to be expected.

Possible future research directions include the following:

- In real-world applications, the label set is usually not static. New labels may be added over time, whereas others could become irrelevant. The capability of adding and removing new labels over time has been explored in few papers [79], but has not reached a level that allows use in real-world systems.

- Streaming data exhibits properties such as concept drift and recurring concepts. For example, a label might become less frequent during winter and more frequent in summer. Such scenarios are not handled properly by most available models.

- Another line of future work is to train topic models using active learning. In the case of text data streams it is often difficult to label all incoming new documents by hand. Active learning could help to actively select documents that differ from previously viewed documents or where the algorithm has the least confidence during labeling and automatically infers labels for the rest. Semi-supervised extensions are also related to this field and could help to train better models with less labeled training data [17; 16].

- A generalization of the HDP is given by the hierarchical Poisson-Dirichlet process (HPDP), sometimes also called hierarchical Pitman-Yor process: In this stochastic process, an additional parameter a is the so-called discount parameter. For $a = 0$ the process reduces to the normal HDP. So far, only the standard HDP was employed for multi-label classification, investigations on the effect of setting the parameter a to different values are still pending. This could potentially help to model label or topic frequencies that are even more skewed and follow a power law distribution.

- The importance of averaging over different samples and different estimators for multi-label topic models was investigated by Papanikolaou *et al.* [54]. Different methods of estimation and their effect on different evaluation measures remain open for further investigation.

- Recently, topic models are increasingly trained using neural networks [15], however, the research on multi-label classification using neural network topic models is still scarce [52]. This would enable the use of many recent advances in deep learning such as convolutions, recurrent networks and different prior distributions. For example, it does not increase the complexity of a neural network topic model when the assumption of a mixture model, that all documents are mixtures of topics, is dropped [70]. This can make the model more expressive by allowing each document to be represented by different combinations or products of topics. Additionally, neural networks can more easily be extended to use word vectors (i.e. pre-trained vector representations of words that capture semantic and syntactic attributes of the words) or other layer types that allow to take into account word order and syntax.

7.1 Related Areas

Instead of using topic models to perform the classification directly, they can also be used to train topic embeddings which may subsequently be used as features in classification [27]. They are also applicable in semi-supervised settings [17; 16] where unlabeled data improves the classification performance. In addition to text, multi-label topic models can also be applied on different data types not covered in this survey, e.g., images [50]. Another important research question is the visualization of topics [21], since the interpretability and usability of a system depends largely on the way the results are presented to users and analysts.

8. CONCLUSION

Multi-label topic models are applied to large-scale multi-label text classification problems where interpretability and flexibility of the model are important factors. This field, the relevant background and applications are summarized in this survey. The field is divided according to three different categories: Topic models are trained online or as a batch model, they are parametric or nonparametric and they may model label dependencies or not. For each of these categories, the most important work is discussed and compared. Commonly used datasets and evaluation measures are reviewed, limitations discussed and possible research directions are proposed. In conclusion, this survey gives an extensive overview of all aspects of this emerging field and thereby demonstrates the manifold possibilities and flexibility of the topic model framework for the complex setting of multi-label classification. Additionally, the number of open research questions shows that there will likely be a lot more work in this area in the near future.

9. REFERENCES

[1] C. E. Antoniak. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *Annals of Statistics*, 2(6):1152–1174, 11 1974.

[2] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 27–34, Arlington, Virginia, United States, 2009. AUAI Press.

[3] P. Bhattacharya, M. B. Zafar, N. Ganguly, S. Ghosh, and K. P. Gummadi. Inferring user interests in the twitter social network. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 357–360. ACM, 2014.

[4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[5] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, January 2003.

[6] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, March 2004.

[7] W. Buntine and M. Hutter. A Bayesian view of the Poisson-Dirichlet process. *arXiv preprint arXiv:1007.0296*, 2010.

[8] S. Burkhardt. *Online Multi-label Text Classification Using Topic Models*. PhD thesis, Johannes Gutenberg-Universität Mainz, 2018.

[9] S. Burkhardt and S. Kramer. Multi-label classification using stacked hierarchical dirichlet processes with reduced sampling complexity. In *ICBK 2017 - International Conference on Big Knowledge*, pages 1–8, Hefei, China, 2017. IEEE.

[10] S. Burkhardt and S. Kramer. Online sparse collapsed hybrid variational-gibbs algorithm for hierarchical dirichlet process topic models. In M. Ceci, J. Hollmén, L. Todorovski, C. Vens, and S. Džeroski, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 189–204, Cham, 2017. Springer International Publishing.

[11] S. Burkhardt and S. Kramer. Multi-label classification using stacked hierarchical dirichlet processes with reduced sampling complexity. *Knowledge and Information Systems*, pages 1–23, 2018.

[12] S. Burkhardt and S. Kramer. Multi-label classification using stacked hierarchical dirichlet processes with reduced sampling complexity. *Knowledge and Information Systems*, pages 1–23, 2018.

[13] S. Burkhardt and S. Kramer. Online multi-label dependency topic models for text classification. *Machine Learning*, 107(5):859–886, May 2018.

[14] S. Burkhardt and S. Kramer. Online multi-label dependency topic models for text classification. *Machine Learning*, 107(5):859–886, May 2018.

[15] S. Burkhardt and S. Kramer. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *Journal of Machine Learning Research*, 20(131):1–27, 2019.

[16] S. Burkhardt, J. Siekiera, J. Glodde, M. A. Andrade-Navarro, and S. Kramer. Towards identifying drug side effects from social media using active learning and crowd sourcing. In *Pacific Symposium of Biocomputing (PSB)*, page accepted, 2020.

[17] S. Burkhardt, J. Siekiera, and S. Kramer. Semi-supervised bayesian active learning for text classification. In *Bayesian Deep Learning Workshop at NeurIPS*, 2018.

[18] K. Canini, L. Shi, and T. Griffiths. Online inference of topics with latent dirichlet allocation. In D. van Dyk and M. Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 65–72, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

[19] O. Cappé and E. Moulines. On-line expectation-maximization algorithm for latent data models. *Journal of the Royal Statistical Society Series B*, 71(3):593–613, 2009.

[20] C. Chen, L. Du, and W. Buntine. Sampling table configurations for the hierarchical Poisson-Dirichlet process. In D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, editors, *Proc. of ECML-PKDD*, pages 296–311, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[21] J. Chuang, D. Ramage, C. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 443–452, New York, NY, USA, 2012. ACM.

[22] R. Cohen and D. Ruths. Classifying political orientation on twitter: It's not easy! In *Seventh International AAAI Conference on Weblogs and Social Media*, 2013.

[23] J. Foulds, L. Boyles, C. DuBois, P. Smyth, and M. Welling. Stochastic collapsed variational bayesian inference for latent dirichlet allocation. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, pages 446–454, New York, NY, USA, 2013. ACM.

[24] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, Nov 2008.

[25] H. Gouk, B. Pfahringer, and M. J. Cree. Learning distance metrics for multi-label classification. In *8th Asian Conference on Machine Learning*, volume 63, pages 318–333, 2016.

[26] T. L. Griffiths and M. Steyvers. Finding scientific topics. In *Proc. of the National Academy of Sciences of the United States of America*, volume 101, pages 5228–5235. National Acad Sciences, 2004.

[27] S. Gururangan, T. Dang, D. Card, and N. A. Smith. Variational pretraining for semi-supervised text classification. *arXiv preprint arXiv:1906.02242*, 2019.

[28] M. D. Hoffman, D. M. Blei, and F. R. Bach. Online learning for latent dirichlet allocation. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 856–864. Curran Associates, Inc., 2010.

[29] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14:1303–1347, May 2013.

[30] Y. Hu, J. Boyd-Graber, B. Satinoff, and A. Smith. Interactive topic modeling. *Machine Learning*, 95(3):423–469, Jun 2014.

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

[32] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning*, pages 137–142. Springer, 1998.

[33] N. Johri, D. Ramage, D. A. McFarland, and D. Jurafsky. A study of academic collaboration in computational linguistics with latent mixtures of authors. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, LaTeCH '11, pages 124–132, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[34] I. Katakis, G. Tsoumakas, and I. Vlahavas. Multilabel text classification for automated tag suggestion. In *ECML-PKDD discovery challenge*, volume 75, 2008.

[35] S. Lacoste-Julien, F. Sha, and M. I. Jordan. Disclda: Discriminative learning for dimensionality reduction and classification. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 897–904. Curran Associates, Inc., 2009.

[36] J. D. Lafferty and D. M. Blei. Correlated topic models. In *Advances in neural information processing systems*, pages 147–154, 2006.

[37] A. Q. Li, A. Ahmed, S. Ravi, and A. J. Smola. Reducing the sampling complexity of topic models. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 891–900, New York, NY, USA, 2014. ACM.

[38] W. Li. *Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations*. PhD thesis, University of Massachusetts Amherst, April 2007.

[39] W. Li and A. McCallum. Pachinko allocation: Dag-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 577–584, New York, NY, USA, 2006. ACM.

[40] X. Li, J. Ouyang, and X. Zhou. Supervised topic models for multi-label classification. *Neurocomput.*, 149(PB):811–819, Feb. 2015.

[41] E. Loza Mencía and J. Fürnkranz. Efficient multilabel classification algorithms for large-scale problems in the legal domain. In E. Francesconi, S. Montemagni, W. Peters, and D. Tiscornia, editors, *Semantic Processing of Legal Texts – Where the Language of Law Meets the Law of Language*, volume 6036 of *Lecture Notes in Artificial Intelligence*, pages 192–215. Springer-Verlag, 1 edition, May 2010.

[42] X.-L. Mao, Z.-Y. Ming, T.-S. Chua, S. Li, H. Yan, and X. Li. Sshlda: a semi-supervised hierarchical topic model. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*, pages 800–809. Association for Computational Linguistics, 2012.

[43] J. McAuley and J. Leskovec. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM conference on Recommender systems*, pages 165–172. ACM, 2013.

[44] J. D. Mcauliffe and D. M. Blei. Supervised topic models. In *Advances in neural information processing systems*, pages 121–128, 2008.

[45] D. A. McFarland, D. Ramage, J. Chuang, J. Heer, C. D. Manning, and D. Jurafsky. Differentiating language usage through topic models. *Poetics*, 41(6):607 – 625, 2013.

[46] D. Mimno and A. McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, UAI'08, pages 411–418, Arlington, Virginia, United States, 2008. AUAI Press.

[47] A. Mukherjee and B. Liu. Aspect extraction through semi-supervised modeling. In *Proceedings of the 50th annual meeting of the association for computational linguistics: Long papers-volume 1*, pages 339–348. Association for Computational Linguistics, 2012.

[48] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[49] J. Nam, J. Kim, E. Loza Mencía, I. Gurevych, and J. Fürnkranz. Large-scale multi-label text classification — revisiting neural networks. In T. Calders, F. Esposito, E. Hüllermeier, and R. Meo, editors, *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, pages 437–452, Berlin Heidelberg, 2014. Springer.

[50] C.-T. Nguyen, D.-C. Zhan, and Z.-H. Zhou. Multi-modal image annotation with multi-instance multi-label lda. In *Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[51] D. Padmanabhan, S. Bhat, S. Shevade, and Y. Narahari. Topic model based multi-label classification. In *2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 996–1003, Nov 2016.

[52] R. Panda, A. Pensia, N. Mehta, M. Zhou, and P. Rai. Deep topic models for multi-label learning. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of Machine Learning Research*, volume 89 of *Proceedings of Machine Learning Research*, pages 2849–2857. PMLR, 16–18 Apr 2019.

[53] E. Papagiannopoulou, Y. Papanikolaou, D. Dimitriadis, S. Lagopoulos, G. Tsoumakas, M. Laliotis, N. Markantonatos, and I. Vlahavas. Large-scale semantic indexing and question answering in biomedicine. In *Proceedings of the Fourth BioASQ workshop*, pages 50–54, Berlin, Germany, Aug. 2016. Association for Computational Linguistics.

[54] Y. Papanikolaou, J. R. Foulds, T. N. Rubin, and G. Tsoumakas. Dense distributions from sparse samples: Improved gibbs sampling parameter estimators for lda. *Journal of Machine Learning Research*, 18(62):1–58, 2017.

[55] A. J. Perotte, F. Wood, N. Elhadad, and N. Bartlett. Hierarchically supervised latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, pages 2609–2617, 2011.

[56] Y. Prabhu and M. Varma. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 263–272, New York, NY, USA, 2014. ACM.

[57] D. Quercia, H. Askham, and J. Crowcroft. Tweetlda: supervised topic classification and link prediction in twitter. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 247–250. ACM, 2012.

[58] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. 2019.

[59] D. Ramage, S. Dumais, and D. Liebling. Characterizing microblogs with topic models. In *Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[60] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning. 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: Volume 1 - Volume 1*, EMNLP '09, pages 248–256, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[61] D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina. Clustering the tagged web. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 54–63, New York, NY, USA, 2009. ACM.

[62] D. Ramage, C. D. Manning, and S. Dumais. Partially Labeled Topic Models for Interpretable Text Mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 457–465, New York, NY, USA, 2011. ACM.

[63] D. Ramage, C. D. Manning, and D. A. Mcfarland. Which universities lead and lag? toward university rankings based on scholarly output. In *In Proc. of NIPS Workshop on Computational Social Science and the Wisdom of the Crowds*, 2010.

[64] A. Rodríguez, D. B. Dunson, and A. E. Gelfand. The Nested Dirichlet Process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008.

[65] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[66] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, July 2012.

[67] R. Salakhutdinov, J. B. Tenenbaum, and A. Torralba. Learning with hierarchical-deep models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1958–1971, 2013.

[68] F. Sebastiani. Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1):1–47, Mar. 2002.

[69] M. Shimosaka, T. Tsukiji, S. Tominaga, and K. Tsubouchi. Coupled Hierarchical Dirichlet Process Mixtures for Simultaneous Clustering and Topic Modeling. In P. Frasconi, N. Landwehr, G. Manco, and J. Vreeken, editors, *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD), Lecture Notes in Computer Science, vol. 9852*, pages 230–246, Cham, 2016. Springer International Publishing.

[70] A. Srivastava and C. Sutton. Autoencoding variational inference for topic models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[71] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[72] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 1353–1360. MIT Press, 2007.

[73] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[74] G. Tsoumakas, I. Katakis, and I. Vlahavas. Mining multi-label data. In O. Maimon and L. Rokach, editors, *Data Mining and Knowledge Discovery Handbook*, pages 667–685. Springer US, 2010.

[75] G. Tsoumakas, I. Katakis, and I. P. Vlahavas. Effective and Efficient Multilabel Classification in Domains with Large Number of Labels. In *ECML/PKDD 2008 Workshop on Mining Multidimensional Data*, 2008.

[76] H. Wang, M. Huang, and X. Zhu. A generative probabilistic model for multi-label classification. In *Eighth IEEE International Conference on Data Mining*, pages 628–637. IEEE, Dec 2008.

[77] J. Wicker, B. Pfahringer, and S. Kramer. Multi-label classification using boolean matrix decomposition. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 179–186, New York, NY, USA, 2012. ACM.

[78] J. Wicker, A. Tyukin, and S. Kramer. *A Nonlinear Label Compression and Transformation Method for Multi-label Classification Using Autoencoders*, pages 328–340. Springer International Publishing, Cham, 2016.

[79] J. Zhang, Z. Ghahramani, and Y. Yang. A probabilistic model for online document clustering with application to novelty detection. In *Advances in neural information processing systems*, pages 1617–1624, 2005.

[80] L. Zhang, S. K. Shah, and I. A. Kakadiaris. Hierarchical multi-label classification using fully associative ensemble learning. *Pattern Recognition*, 70:89–103, 2017.

[81] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2013.

[82] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.

[83] Y. Zhang, J. Ma, Z. Wang, and B. Chen. Lf-lda: A topic model for multi-label classification. In *International Conference on Emerging Internetworking, Data & Web Technologies*, pages 618–628. Springer, 2017.

[84] J. Zhu, A. Ahmed, and E. P. Xing. Medlda: Maximum margin supervised topic models for regression and classification. In *Proceedings of the 26th International Conference on Machine Learning*, ICML '09, pages 1257–1264, New York, NY, USA, 2009. ACM.