Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora

Daniel Ramage, David Hall, Ramesh Nallapati and Christopher D. Manning
Computer Science Department
Stanford University
{dramage,dlwh,nmramesh,manning}@cs.stanford.edu

Abstract

A significant portion of the world's text is tagged by readers on social bookmarking websites. *Credit attribution* is an inherent problem in these corpora because most pages have multiple tags, but the tags do not always apply with equal specificity across the whole document. Solving the credit attribution problem requires associating each word in a document with the most appropriate tags and vice versa. This paper introduces *Labeled LDA*, a topic model that constrains Latent Dirichlet Allocation by defining a one-to-one correspondence between LDA's latent topics and user tags. This allows Labeled LDA to directly learn word-tag correspondences. We demonstrate Labeled LDA's improved expressiveness over traditional LDA with visualizations of a corpus of tagged web pages from *del.icio.us*. Labeled LDA outperforms SVMs by more than 3 to 1 when extracting tag-specific document snippets. As a multi-label text classifier, our model is competitive with a discriminative baseline on a variety of datasets.

1 Introduction

From news sources such as *Reuters* to modern community web portals like *del.icio.us*, a significant proportion of the world's textual data is labeled with multiple human-provided tags. These collections reflect the fact that documents are often about more than one thing—for example, a news story about a highway transportation bill might naturally be filed under both *transportation* and *politics*, with neither category acting as a clear subset of the other. Similarly, a single web page in *del.icio.us* might well be annotated with tags as diverse as *arts*, *physics*, *alaska*, and *beauty*.

However, not all tags apply with equal specificity across the whole document, opening up new opportunities for information retrieval and corpus analysis on tagged corpora. For instance, users who browse for documents with a particular tag might prefer to see summaries that focus on the portion of the document most relevant to the tag, a task we call *tag-specific snippet extraction*. And when a user browses to a particular document, a tag-augmented user interface might provide overview visualization cues highlighting which portions of the document are more or less relevant to the tag, helping the user quickly access the information they seek.

One simple approach to these challenges can be found in models that explicitly address the *credit attribution* problem by associating individual words in a document with their most appropriate labels. For instance, in our news story about the transportation bill, if the model knew that the word "highway" went with *transportation* and that the word "politicians" went with *politics*, more relevant passages could be extracted for either label. We seek an approach that can automatically learn the posterior distribution of each word in a document conditioned on the document's label set.

One promising approach to the credit attribution problem lies in the machinery of Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a recent model that has gained popularity among theoreticians and practitioners alike as a tool for automatic corpus summarization and visualization. LDA is a completely unsupervised algorithm that models each document as a mixture of topics. The model generates automatic summaries of topics in terms of a discrete probability distribution over words for each topic, and further infers per-document discrete distributions over topics. Most importantly, LDA makes the explicit assumption that each word is generated from one underlying topic.

Although LDA is expressive enough to model

multiple topics per document, it is not appropriate for multi-labeled corpora because, as an unsupervised model, it offers no obvious way of incorporating a supervised label set into its learning procedure. In particular, LDA often learns some topics that are hard to interpret, and the model provides no tools for tuning the generated topics to suit an end-use application, even when time and resources exist to provide some document labels.

Several modifications of LDA to incorporate supervision have been proposed in the literature. Two such models, Supervised LDA (Blei and McAuliffe, 2007) and DiscLDA (Lacoste-Julien et al., 2008) are inappropriate for multiply labeled corpora because they limit a document to being associated with only a single label. Supervised LDA posits that a label is generated from each document's empirical topic mixture distribution. DiscLDA associates a single categorical label variable with each document and associates a topic mixture with each label. A third model, MM-LDA (Ramage et al., 2009), is not constrained to one label per document because it models each document as a bag of words with a bag of labels, with topics for each observation drawn from a shared topic distribution. But, like the other models, MM-LDA's learned topics do not correspond directly with the label set. Consequently, these models fall short as a solution to the credit attribution problem. Because labels have meaning to the people that assigned them, a simple solution to the credit attribution problem is to assign a document's words to its labels rather than to a latent and possibly less interpretable semantic space.

This paper presents *Labeled LDA* (L-LDA), a generative model for multiply labeled corpora that marries the multi-label supervision common to modern text datasets with the word-assignment ambiguity resolution of the LDA family of models. In contrast to standard LDA and its existing supervised variants, our model associates each label with one topic in direct correspondence. In the following section, L-LDA is shown to be a natural extension of both LDA (by incorporating supervision) and Multinomial Naive Bayes (by incorporating a mixture model). We demonstrate that L-LDA can go a long way toward solving the credit attribution problem in multiply labeled documents with improved interpretability over LDA (Section 4). We show that L-LDA's credit attribution ability enables it to greatly outperform support vector machines on a tag-driven snippet extraction task on web pages from *del.icio.us* (Section 6). And despite its generative semantics, we show that Labeled LDA is competitive with a strong baseline discriminative classifier on two multi-label text classification tasks (Section 7).

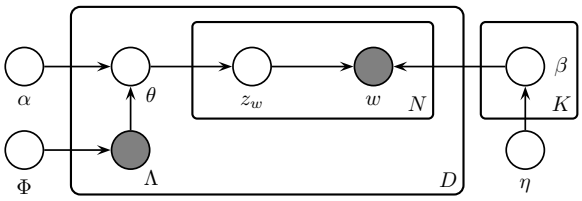

Figure 1: Graphical model of Labeled LDA: unlike standard LDA, both the label set $\mathbf{\Lambda}$ as well as the topic prior $\boldsymbol{\alpha}$ influence the topic mixture $\boldsymbol{\theta}$.

2 Labeled LDA

Labeled LDA is a probabilistic graphical model that describes a process for generating a labeled document collection. Like Latent Dirichlet Allocation, Labeled LDA models each document as a mixture of underlying topics and generates each word from one topic. Unlike LDA, L-LDA incorporates supervision by simply constraining the topic model to use only those topics that correspond to a document's (observed) label set. The model description that follows assumes the reader is familiar with the basic LDA model (Blei et al., 2003).

Let each document d be represented by a tuple consisting of a list of word indices $\mathbf{w}^{(d)} = (w_1, \ldots, w_{N_d})$ and a list of binary topic presence/absence indicators $\mathbf{\Lambda}^{(d)} = (l_1, \ldots, l_K)$ where each $w_i \in \{1, \ldots, V\}$ and each $l_k \in \{0, 1\}$. Here N_d is the document length, V is the vocabulary size and K the total number of unique labels in the corpus.

We set the number of topics in Labeled LDA to be the number of unique labels K in the corpus. The generative process for the algorithm is found in Table 1. Steps 1 and 2—drawing the multinomial topic distributions over vocabulary $\boldsymbol{\beta}_k$ for each topic k, from a Dirichlet prior $\boldsymbol{\eta}$—remain the same as for traditional LDA (see (Blei et al., 2003), page 4). The traditional LDA model then draws a multinomial mixture distribution $\boldsymbol{\theta}^{(d)}$ over all K topics, for each document d, from a Dirichlet prior $\boldsymbol{\alpha}$. However, we would like to restrict $\boldsymbol{\theta}^{(d)}$ to be defined only over the topics that correspond to

1 For each topic $k \in \{1, \ldots, K\}$:
2 Generate $\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,V})^T \sim \text{Dir}(\cdot|\boldsymbol{\eta})$
3 For each document d:
4 For each topic $k \in \{1, \ldots, K\}$
5 Generate $\Lambda_k^{(d)} \in \{0, 1\} \sim \text{Bernoulli}(\cdot|\Phi_k)$
6 Generate $\boldsymbol{\alpha}^{(d)} = L^{(d)} \times \boldsymbol{\alpha}$
7 Generate $\boldsymbol{\theta}^{(d)} = (\theta_{l_1}, \ldots, \theta_{l_{M_d}})^T \sim \text{Dir}(\cdot|\boldsymbol{\alpha}^{(d)})$
8 For each i in $\{1, \ldots, N_d\}$:
9 Generate $z_i \in \{\lambda_1^{(d)}, \ldots, \lambda_{M_d}^{(d)}\} \sim \text{Mult}(\cdot|\boldsymbol{\theta}^{(d)})$
10 Generate $w_i \in \{1, \ldots, V\} \sim \text{Mult}(\cdot|\boldsymbol{\beta}_{z_i})$

Table 1: Generative process for Labeled LDA: $\boldsymbol{\beta}_k$ is a vector consisting of the parameters of the multinomial distribution corresponding to the k^{th} topic, $\boldsymbol{\alpha}$ are the parameters of the Dirichlet topic prior and $\boldsymbol{\eta}$ are the parameters of the word prior, while Φ_k is the label prior for topic k. For the meaning of the projection matrix $L^{(d)}$, please refer to Eq 1.

its labels $\boldsymbol{\Lambda}^{(d)}$. Since the word-topic assignments z_i (see step 9 in Table 1) are drawn from this distribution, this restriction ensures that all the topic assignments are limited to the document's labels.

Towards this objective, we first generate the document's labels $\boldsymbol{\Lambda}^{(d)}$ using a Bernoulli coin toss for each topic k, with a labeling prior probability Φ_k, as shown in step 5. Next, we define the vector of document's labels to be $\boldsymbol{\lambda}^{(d)} = \{k|\Lambda_k^{(d)} = 1\}$. This allows us to define a document-specific label projection matrix $L^{(d)}$ of size $M_d \times K$ for each document d, where $M_d = |\boldsymbol{\lambda}^{(d)}|$, as follows: For each row $i \in \{1, \ldots, M_d\}$ and column $j \in \{1, \ldots, K\}$:

$$L_{ij}^{(d)} = \begin{cases} 1 & \text{if } \lambda_i^{(d)} = j \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

In other words, the i^{th} row of $L^{(d)}$ has an entry of 1 in column j if and only if the i^{th} document label $\lambda_i^{(d)}$ is equal to the topic j, and zero otherwise. As the name indicates, we use the $L^{(d)}$ matrix to project the parameter vector of the Dirichlet topic prior $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)^T$ to a lower dimensional vector $\boldsymbol{\alpha}^{(d)}$ as follows:

$$\boldsymbol{\alpha}^{(d)} = L^{(d)} \times \boldsymbol{\alpha} = (\alpha_{\lambda_1^{(d)}}, \ldots, \alpha_{\lambda_{M_d}^{(d)}})^T \quad (2)$$

Clearly, the dimensions of the projected vector correspond to the topics represented by the labels of the document. For example, suppose $K = 4$ and that a document d has labels given by $\boldsymbol{\Lambda}^{(d)} = \{0, 1, 1, 0\}$ which implies $\boldsymbol{\lambda}^{(d)} = \{2, 3\}$, then $L^{(d)}$ would be:

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Then, $\boldsymbol{\theta}^{(d)}$ is drawn from a Dirichlet distribution with parameters $\boldsymbol{\alpha}^{(d)} = L^{(d)} \times \boldsymbol{\alpha} = (\alpha_2, \alpha_3)^T$ (i.e., with the Dirichlet restricted to the topics 2 and 3).

This fulfills our requirement that the document's topics are restricted to its own labels. The projection step constitutes the deterministic step 6 in Table 1. The remaining part of the model from steps 7 through 10 are the same as for regular LDA.

The dependency of $\boldsymbol{\theta}$ on both $\boldsymbol{\alpha}$ and $\boldsymbol{\Lambda}$ is indicated by directed edges from $\boldsymbol{\Lambda}$ and $\boldsymbol{\alpha}$ to $\boldsymbol{\theta}$ in the plate notation in Figure 1. This is the only additional dependency we introduce in LDA's representation (please compare with Figure 1 in (Blei et al., 2003)).

2.1 Learning and inference

In most applications discussed in this paper, we will assume that the documents are multiply tagged with human labels, both at learning and inference time.

When the labels $\boldsymbol{\Lambda}^{(d)}$ of the document are observed, the labeling prior Φ is d-separated from the rest of the model given $\boldsymbol{\Lambda}^{(d)}$. Hence the model is same as traditional LDA, except the constraint that the topic prior $\boldsymbol{\alpha}^{(d)}$ is now restricted to the set of labeled topics $\boldsymbol{\lambda}^{(d)}$. Therefore, we can use collapsed Gibbs sampling (Griffiths and Steyvers, 2004) for training where the sampling probability for a topic for position i in a document d in Labeled LDA is given by:

$$P(z_i = j|\mathbf{z}_{-i}) \propto \frac{n_{-i,j}^{w_i} + \eta_{w_i}}{n_{-i,j}^{(\cdot)} + \boldsymbol{\eta}^T\mathbf{1}} \times \frac{n_{-i,j}^{(d)} + \alpha_j}{n_{-i,\cdot}^{(d)} + \boldsymbol{\alpha}^T\mathbf{1}} \quad (3)$$

where $n_{-i,j}^{w_i}$ is the count of word w_i in topic j, that does not include the current assignment z_i, a missing subscript or superscript (e.g. $n_{-i,j}^{(\cdot)}$)) indicates a summation over that dimension, and $\mathbf{1}$ is a vector of 1's of appropriate dimension.

Although the equation above looks exactly the same as that of LDA, we have an important distinction in that, the target topic j is restricted to belong to the set of labels, i.e., $j \in \boldsymbol{\lambda}^{(d)}$.

Once the topic multinomials $\boldsymbol{\beta}$ are learned from the training set, one can perform inference on any new labeled test document using Gibbs sampling

restricted to its tags, to determine its per-word label assignments $\mathbf{z}$. In addition, one can also compute its posterior distribution $\boldsymbol{\theta}$ over topics by appropriately normalizing the topic assignments $\mathbf{z}$.

It should now be apparent to the reader how the new model addresses some of the problems in multi-labeled corpora that we highlighted in Section 1. For example, since there is a one-to-one correspondence between the labels and topics, the model can display automatic topical summaries for each label k in terms of the topic-specific distribution $\boldsymbol{\beta}_k$. Similarly, since the model assigns a label z_i to each word w_i in the document d automatically, we can now extract portions of the document relevant to each label k (it would be all words $w_i \in \mathbf{w}^{(d)}$ such that $z_i = k$). In addition, we can use the topic distribution $\boldsymbol{\theta}^{(d)}$ to rank the user specified labels in the order of their relevance to the document, thereby also eliminating spurious ones if necessary.

Finally, we note that other less restrictive variants of the proposed L-LDA model are possible. For example, one could consider a version that allows topics that do not correspond to the label set of a given document with a small probability, or one that allows a common background topic in all documents. We did implement these variants in our preliminary experiments, but they did not yield better performance than L-LDA in the tasks we considered. Hence we do not report them in this paper.

2.2 Relationship to Naive Bayes

The derivation of the algorithm so far has focused on its relationship to LDA. However, Labeled LDA can also be seen as an extension of the event model of a traditional Multinomial Naive Bayes classifier (McCallum and Nigam, 1998) by the introduction of a mixture model. In this section, we develop the analogy as another way to understand L-LDA from a supervised perspective.

Consider the case where no document in the collection is assigned two or more labels. Now for a particular document d with label l_d, Labeled LDA draws each word's topic variable z_i from a multinomial constrained to the document's label set, i.e. $z_i = l_d$ for each word position i in the document. During learning, the Gibbs sampler will assign each z_i to l_d while incrementing $\beta_{l_d}(w_i)$, effectively counting the occurences of each word type in documents labeled with l_d. Thus in the singly labeled document case, the probability of each document under Labeled LDA is equal to the probability of the document under the Multinomial Naive Bayes event model trained on those same document instances. Unlike the Multinomial Naive Bayes classifier, Labeled LDA does not encode a decision boundary for unlabeled documents by comparing $P(\mathbf{w}^{(d)}|l_d)$ to $P(\mathbf{w}^{(d)}|\neg l_d)$, although we discuss using Labeled LDA for multi-label classification in Section 7.

Labeled LDA's similarity to Naive Bayes ends with the introduction of a second label to any document. In a traditional one-versus-rest Multinomial Naive Bayes model, a separate classifier for each label would be trained on all documents with that label, so each word can contribute a count of 1 to every observed label's word distribution. By contrast, Labeled LDA assumes that each document is a mixture of underlying topics, so the count mass of single word instance must instead be distributed over the document's observed labels.

3 Credit attribution within tagged documents

Social bookmarking websites contain millions of tags describing many of the web's most popular and useful pages. However, not all tags are uniformly appropriate at all places within a document. In the sections that follow, we examine mechanisms by which Labeled LDA's credit assignment mechanism can be utilized to help support browsing and summarizing tagged document collections.

To create a consistent dataset for experimenting with our model, we selected 20 tags of medium to high frequency from a collection of documents dataset crawled from *del.icio.us*, a popular social bookmarking website (Heymann et al., 2008). From that larger dataset, we selected uniformly at random four thousand documents that contained at least one of the 20 tags, and then filtered each document's tag set by removing tags not present in our tag set. After filtering, the resulting corpus averaged 781 non-stop words per document, with each document having 4 distinct tags on average. In contrast to many existing text datasets, our tagged corpus is highly multiply labeled: almost 90% of of the documents have more than one tag. (For comparison, less than one third of the news documents in the popular RCV1-v2 collection of newswire are multiply labeled). We will refer to

this collection of data as the del.icio.us tag dataset.

4 Topic Visualization

A first question we ask of Labeled LDA is how its topics compare with those learned by traditional LDA on the same collection of documents. We ran our implementations of Labeled LDA and LDA on the del.icio.us corpus described above. Both are based on the standard collapsed Gibbs sampler, with the constraints for Labeled LDA implemented as in Section 2.

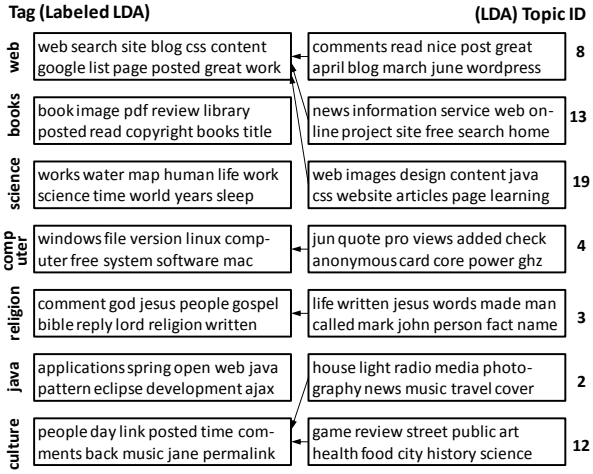

Figure 2: Comparison of some of the 20 topics learned on del.icio.us by Labeled LDA (left) and traditional LDA (right), with representative words for each topic shown in the boxes. Labeled LDA's topics are named by their associated tag. Arrows from right-to-left show the mapping of LDA topics to the closest Labeled LDA topic by cosine similarity. Tags not shown are: *design*, *education*, *english*, *grammar*, *history*, *internet*, *language*, *philosophy*, *politics*, *programming*, *reference*, *style*, *writing*.

Figure 2 shows the top words associated with 20 topics learned by Labeled LDA and 20 topics learned by unsupervised LDA on the del.icio.us document collection. Labeled LDA's topics are directly named with the tag that corresponds to each topic, an improvement over standard practice of inferring the topic name by inspection (Mei et al., 2007). The topics learned by the unsupervised variant were matched to a Labeled LDA topic highest cosine similarity.

The topics selected are representative: compared to Labeled LDA, unmodified LDA allocates many topics for describing the largest parts of the corpus and under-represents tags that are less uncommon: of the 20 topics learned, LDA learned multiple topics mapping to each of five tags (*web*, *culture*, and *computer*, *reference*, and *politics*, all of which were common in the dataset) and learned no topics that aligned with six tags (*books*, *english*, *science*, *history*, *grammar*, *java*, and *philosophy*, which were rarer).

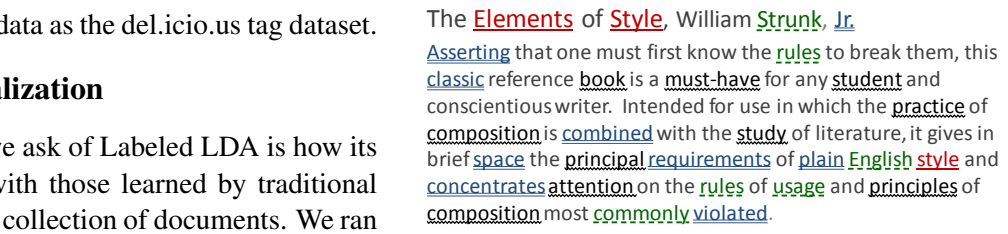

Figure 3: Example document with important words annotated with four of the page's tags as learned by Labeled LDA. Red (single underline) is *style*, green (dashed underline) *grammar*, blue (double underline) *reference*, and black (jagged underline) *education*.

5 Tagged document visualization

In addition to providing automatic summaries of the words best associated with each tag in the corpus, Labeled LDA's credit attribution mechanism can be used to augment the view of a single document with rich contextual information about the document's tags.

Figure 3 shows one web document from the collection, a page describing a guide to writing English prose. The 10 most common tags for that document are *writing, reference, english, grammar, style, language, books, book, strunk*, and *education*, the first eight of which were included in our set of 20 tags. In the figure, each word that has high posterior probability from one tag has been annotated with that tag. The red words come from the *style* tag, green from the *grammar* tag, blue from the *reference* tag, and black from the *education* tag. In this case, the model does very well at assigning individual words to the tags that, subjectively, seem to strongly imply the presence of that tag on this page. A more polished rendering could add subtle visual cues about which parts of a page are most appropriate for a particular set of tags.

books

L-LDA this classic reference book is a must-have for any student and conscientious writer. Intended for

SVM the rules of usage and principles of composition most commonly violated. Search: CONTENTS Bibliographic

language

L-LDA the beginning of a sentence must refer to the grammatical subject 8. Divide words at

SVM combined with the study of literature, it gives in brief space the principal requirements of

grammar

L-LDA requirements of plain English style and concentrates attention on the rules of usage and principles of

SVM them, this classic reference book is a must-have for any student and conscientious writer.

Figure 4: Representative snippets extracted by L-LDA and tag-specific SVMs for the web page shown in Figure 3.

6 Snippet Extraction

Another natural application of Labeled LDA's credit assignment mechanism is as a means of selecting snippets of a document that best describe its contents from the perspective of a particular tag. Consider again the document in Figure 3. Intuitively, if this document were shown to a user interested in the tag *grammar*, the most appropriate snippet of words might prefer to contain the phrase "rules of usage," whereas a user interested in the term *style* might prefer the title "Elements of Style."

To quantitatively evaluate Labeled LDA's performance at this task, we constructed a set of 29 recently tagged documents from del.icio.us that were labeled with two or more tags from the 20 tag subset, resulting in a total of 149 (document,tag) pairs. For each pair, we extracted a 15-word window with the highest tag-specific score from the document. Two systems were used to score each window: Labeled LDA and a collection of one-vs-rest SVMs trained for each tag in the system. L-LDA scored each window as the expected probability that the tag had generated each word. For SVMs, each window was taken as its own document and scored using the tag-specific SVM's un-thresholded scoring function, taking the window with the most positive score. While a complete solution to the tag-specific snippet extraction problem might be more informed by better linguistic features (such as phrase boundaries), this experimental setup suffices to evaluate both kinds of models for their ability to appropriately assign words to underlying labels.

Model	Best Snippet	Unanimous
L-LDA	**72 / 149**	24 / 51
SVM	21 / 149	2 / 51

Table 2: Human judgments of tag-specific snippet quality as extracted by L-LDA and SVM. The center column is the number of document-tag pairs for which a system's snippet was judged superior. The right column is the number of snippets for which all three annotators were in complete agreement (numerator) in the subset of document scored by all three annotators (denominator).

Figure 3 shows some example snippets output by our system for this document. Note that while SVMs did manage to select snippets that were vaguely on topic, Labeled LDA's outputs are generally of superior subjective quality. To quantify this intuition, three human annotators rated each pair of snippets. The outputs were randomly labeled as "System A" or "System B," and the annotators were asked to judge which system generated a better tag-specific document subset. The judges were also allowed to select neither system if there was no clear winner. The results are summarized in Table 2.

L-LDA was judged superior by a wide margin: of the 149 judgments, L-LDA's output was selected as preferable in 72 cases, whereas SVM's was selected in only 21. The difference between these scores was highly significant ($p < .001$) by the sign test. To quantify the reliability of the judgments, 51 of the 149 document-tag pairs were labeled by all three annotators. In this group, the judgments were in substantial agreement,[1] with Fleiss' Kappa at .63.

Further analysis of the triply-annotated subset yields further evidence of L-LDA's advantage over SVM's: 33 of the 51 were tag-page pairs where L-LDA's output was picked by at least one annotator as a better snippet (although L-LDA might not have been picked by the other annotators). And of those, 24 were unanimous in that

[1] Of the 15 judgments that were in contention, only two conflicted on *which* system was superior (L-LDA versus SVM); the remaining disagreements were about whether or not one of the systems was a clear winner.

all three judges selected L-LDA's output. By contrast, only 10 of the 51 were tag-page pairs where SVMs' output was picked by at least one annotator, and of those, only 2 were selected unanimously.

7 Multilabeled Text Classification

In the preceding section we demonstrated how Labeled LDA's credit attribution mechanism enabled effective modeling within documents. In this section, we consider whether L-LDA can be adapted as an effective multi-label classifier for documents as a whole. To answer that question, we applied a modified variant of L-LDA to a multi-label document classification problem: given a training set consisting of documents with multiple labels, predict the set of labels appropriate for each document in a test set.

Multi-label classification is a well researched problem. Many modern approaches incorporate label correlations (e.g., Kazawa et al. (2004), Ji et al. (2008)). Others, like our algorithm are based on mixture models (such as Ueda and Saito (2003)). However, we are aware of no methods that trade off label-specific word distributions with document-specific label distributions in quite the same way.

In Section 2, we discussed learning and inference when labels are observed. In the task of multilabel classification, labels are available at training time, so the learning part remains the same as discussed before. However, inferring the best set of labels for an unlabeled document at test time is more complex: it involves assessing all label assignments and returning the assignment that has the highest posterior probability. However, this is not straight-forward, since there are 2^K possible label assignments. To make matters worse, the support of $\boldsymbol{\alpha}(\Lambda^{(d)})$ is different for different label assignments. Although we are in the process of developing an efficient sampling algorithm for this inference, for the purposes of this paper we make the simplifying assumption that the model reduces to standard LDA at inference, where the document is free to sample from any of the K topics. This is a reasonable assumption because allowing the model to explore the whole topic space for each document is similar to exploring all possible label assignments. The document's most likely labels can then be inferred by suitably thresholding its posterior probability over topics.

As a baseline, we use a set of multiple one-vs-rest SVM classifiers which is a popular and extremely competitive baseline used by most previous papers (see (Kazawa et al., 2004; Ueda and Saito, 2003) for instance). We scored each model based on Micro-F1 and Macro-F1 as our evaluation measures (Lewis et al., 2004). While the former allows larger classes to dominate its results, the latter assigns an equal weight to all classes, providing us complementary information.

7.1 Yahoo

We ran experiments on a corpus from the Yahoo directory, modeling our experimental conditions on the ones described in (Ji et al., 2008).[2] We considered documents drawn from 8 top level categories in the Yahoo directory, where each document can be placed in any number of subcategories. The results were mixed, with SVMs ahead on one measure: Labeled LDA beat SVMs on five out of eight datasets on MacroF1, but didn't win on any datasets on MicroF1. Results are presented in Table 3.

Because only a processed form of the documents was released, the Yahoo dataset does not lend itself well to error analysis. However, only 33% of the documents in each top-level category were applied to more than one sub-category, so the credit assignment machinery of L-LDA was unused for the majority of documents. We therefore ran an artificial second set of experiments considering only those documents that had been given more than one label in the training data. On these documents, the results were again mixed, but Labeled LDA comes out ahead. For MacroF1, L-LDA beat SVMs on four datasets, SVMs beat L-LDA on one dataset, and three were a statistical tie.[3] On MicroF1, L-LDA did much better than on the larger subset, outperforming on four datasets with the other four a statistical tie.

It is worth noting that the Yahoo datasets are skewed by construction to contain many documents with highly overlapping content: because each collection is within the same super-class such as "Arts", "Business", etc., each sub-categories'

[2] We did not carefully tune per-class thresholds of each of the one vs. rest classifiers in each model, but instead tuned only one threshold for all classifiers in each model via cross-validation on the Arts subsets. As such, our numbers were on an average 3-4% less than those reported in (Ji et al., 2008), but the methods were comparably tuned.

[3] The difference between means of multiple runs were not significantly different by two-tailed paired t-test.

Dataset	%MacroF1		%MicroF1	
	L-LDA	SVM	L-LDA	SVM
Arts	30.70(1.62)	23.23 (0.67)	39.81(1.85)	48.42 (0.45)
Business	30.81(0.75)	22.82 (1.60)	67.00(1.29)	72.15 (0.62)
Computers	27.55(1.98)	18.29 (1.53)	48.95(0.76)	61.97 (0.54)
Education	33.78(1.70)	36.03 (1.30)	41.19(1.48)	59.45 (0.56)
Entertainment	39.42(1.38)	43.22 (0.49)	47.71(0.61)	62.89 (0.50)
Health	45.36(2.00)	47.86 (1.72)	58.13(0.43)	72.21 (0.26)
Recreation	37.63(1.00)	33.77 (1.17)	43.71(0.31)	59.15 (0.71)
Society	27.32(1.24)	23.89 (0.74)	42.98(0.28)	52.29 (0.67)

Table 3: Averaged performance across ten runs of multi-label text classification for predicting subsets of the named Yahoo directory categories. Numbers in parentheses are standard deviations across runs. L-LDA outperforms SVMs on 5 subsets with MacroF1, but on no subsets with MicroF1.

vocabularies will naturally overlap a great deal. L-LDA's credit attribution mechanism is most effective at partitioning semantically distinct words into their respective label vocabularies, so we expect that Labeled-LDA's performance as a text classifier would improve on collections with more semantically diverse labels.

7.2 Tagged Web Pages

We also applied our method to text classification on the del.icio.us dataset, where the documents are naturally multiply labeled (more than 89%) and where the tags are less inherently similar than in the Yahoo subcategories. Therefore we expect Labeled LDA to do better credit assignment on this subset and consequently to show improved performance as a classifier, and indeed this is the case.

We evaluated L-LDA and multiple one-vs-rest SVMs on 4000 documents with the 20 tag subset described in Section 3. L-LDA and multiple one-vs-rest SVMs were trained on the first 80% of documents and evaluated on the remaining 20%, with results averaged across 10 random permutations of the dataset. The results are shown in Table 4. We tuned the SVMs' shared cost parameter $C(= 10.0)$ and selected raw term frequency over tf-idf weighting based on 4-fold cross-validation on 3,000 documents drawn from an independent permutation of the data. For L-LDA, we tuned the shared parameters of threshold and proportionality constants in word and topic priors. L-LDA and SVM have very similar performance on MacroF1, while L-LDA substantially outperforms on MicroF1. In both cases, L-LDA's improvement is statistically significantly by a 2-tailed paired t-test at 95% confidence.

Model	%MacroF1	%MicroF1
L-LDA	**39.85** (.989)	**52.12** (.434)
SVM	39.00 (.423)	39.33 (.574)

Table 4: Mean performance across ten runs of multi-label text classification for predicting 20 tags on del.icio.us data. L-LDA outperforms SVMs significantly on both metrics by a 2-tailed, paired t-test at 95% confidence.

8 Discussion

One of the main advantages of L-LDA on multiply labeled documents comes from the model's document-specific topic mixture θ. By explicitly modeling the importance of each label in the document, Labeled LDA can effective perform some contextual word sense disambiguation, which suggests why L-LDA can outperform SVMs on the del.icio.us dataset.

As a concrete example, consider the excerpt of text from the *del.icio.us* dataset in Figure 5. The document itself has several tags, including *design* and *programming*. Initially, many of the likelihood probabilities $p(w|label)$ for the (content) words in this excerpt are higher for the label *programming* than *design*, including "content", "client", "CMS" and even "designed", while *design* has higher likelihoods for just "website" and "happy". However, after performing inference on this document using L-LDA, the inferred document probability for *design* ($p(design)$) is much higher than it is for *programming*. In fact, the higher probability for the tag more than makes up the difference in the likelihood for all the words except "CMS" (Content Management System), so

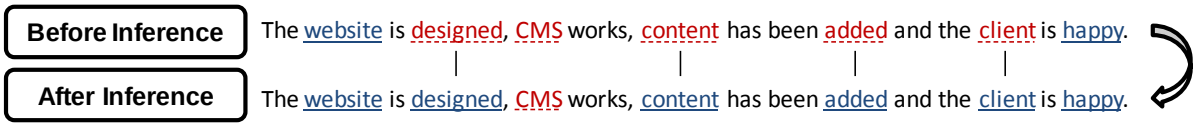

Figure 5: The effect of tag mixture proportions for credit assignment in a web document. Blue (single underline) words are generated from the *design* tag; red (dashed underline) from the *programming* tag. By themselves, most words used here have a higher probability in *programming* than in *design*. But because the document as a whole is more about *design* than *programming*(incorporating words not shown here), inferring the document's topic-mixture θ enables L-LDA to correctly re-assign most words.

that L-LDA correctly infers that most of the words in this passage have more to do with *design* than *programming*.

9 Conclusion

This paper has introduced Labeled LDA, a novel model of multi-labeled corpora that directly addresses the credit assignment problem. The new model improves upon LDA for labeled corpora by gracefully incorporating user supervision in the form of a one-to-one mapping between topics and labels. We demonstrate the model's effectiveness on tasks related to credit attribution within documents, including document visualizations and tag-specific snippet extraction. An approximation to Labeled LDA is also shown to be competitive with a strong baseline (multiple one vs-rest SVMs) for multi-label classification.

Because Labeled LDA is a graphical model in the LDA family, it enables a range of natural extensions for future investigation. For example, the current model does not capture correlations between labels, but such correlations might be introduced by composing Labeled LDA with newer state of the art topic models like the Correlated Topic Model (Blei and Lafferty, 2006) or the Pachinko Allocation Model (Li and McCallum, 2006). And with improved inference for unsupervised $\mathbf{\Lambda}$, Labeled LDA lends itself naturally to modeling semi-supervised corpora where labels are observed for only some documents.

Acknowledgments

This project was supported in part by the President of Stanford University through the IRiSS Initiatives Assessment project.

References

D. M. Blei and J. Lafferty. 2006. Correlated Topic Models. *NIPS*, 18:147.

D. Blei and J McAuliffe. 2007. Supervised Topic Models. In *NIPS*, volume 21.

D. M. Blei, A.Y. Ng, and M.I. Jordan. 2003. Latent Dirichlet allocation. *JMLR*.

T. L. Griffiths and M. Steyvers. 2004. Finding scientific topics. *PNAS*, 1:5228–35.

P. Heymann, G. Koutrika, and H. Garcia-Molina. 2008. Can social bookmarking improve web search. In *WSDM*.

S. Ji, L. Tang, S. Yu, and J. Ye. 2008. Extracting shared subspace for multi-label classification. In *KDD*, pages 381–389, New York, NY, USA. ACM.

H. Kazawa, H. Taira T. Izumitani, and E. Maeda. 2004. Maximal margin labeling for multi-topic text categorization. In *NIPS*.

S. Lacoste-Julien, F. Sha, and M. I. Jordan. 2008. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *NIPS*, volume 22.

D. D. Lewis, Y. Yang, T. G. Rose, G. Dietterich, F. Li, and F. Li. 2004. RCV1: A new benchmark collection for text categorization research. *JMLR*, 5:361–397.

Wei Li and Andrew McCallum. 2006. Pachinko allocation: Dag-structured mixture models of topic correlations. In *International conference on Machine learning*, pages 577–584.

A. McCallum and K. Nigam. 1998. A comparison of event models for naive bayes text classification. In *AAAI-98 workshop on learning for text categorization*, volume 7.

Q. Mei, X. Shen, and C Zhai. 2007. Automatic labeling of multinomial topic models. In *KDD*.

D. Ramage, P. Heymann, C. D. Manning, and H. Garcia-Molina. 2009. Clustering the tagged web. In *WSDM*.

N. Ueda and K. Saito. 2003. Parametric mixture models for multi-labeled text includes models that can be seen to fit within a dimensionality reduction framework. In *NIPS*.