
@inproceedings{Nguyen:Boyd-Graber:Resnik-2013,
  Url = {http://umiacs.umd.edu/~jbg/docs/2013_shlda.pdf},
  Title = {Lexical and Hierarchical Topic Regression},
  Author = {Viet-An Nguyen and Jordan Boyd-Graber and Philip Resnik},
  Booktitle = {Neural Information Processing Systems},
  Year = {2013},
  Location = {Lake Tahoe, Nevada},
}

Links:
- Supplement [https://umiacs.umd.edu/~jbg/docs/2013_shlda-supp.pdf]

Downloaded from [http://umiacs.umd.edu/~jbg/docs/2013_shlda.pdf]

Contact Jordan Boyd-Graber (jbg@boydgraber.org) for questions about this paper.
Abstract

Inspired by a two-level theory from political science that unifies agenda setting and ideological framing, we propose supervised hierarchical latent Dirichlet allocation (SHLDA), which jointly captures documents’ multi-level topic structure and their polar response variables. Our model extends the nested Chinese restaurant processes to discover tree-structured topic hierarchies and uses both per-topic hierarchical and per-word lexical regression parameters to model response variables. SHLDA improves prediction on political affiliation and sentiment tasks in addition to providing insight into how topics under discussion are framed.

1 Introduction: Agenda Setting and Framing in Hierarchical Models

How do liberal-leaning bloggers talk about immigration in the US? What do conservative politicians have to say about education? How do Fox News and MSNBC differ in their language about the gun debate? Such questions concern not only what, but how things are talked about.

In political communication, the question of “what” falls under the heading of agenda setting theory, which concerns the issues introduced into political discourse (e.g., by the mass media) and their influence over public priorities [1]. The question of “how” concerns framing: the way the presentation of an issue reflects or encourages a particular perspective or interpretation [2]. For example, the rise of the “innocence frame” in the death penalty debate, emphasizing the irreversible consequence of mistaken convictions, has led to a sharp decline in the use of capital punishment in the US [3].

In its concern with the subjects or issues under discussion in political discourse, agenda setting maps neatly to topic modeling [4] as a means of discovering and characterizing those issues [5]. Interestingly, one line of communication theory seeks to unify agenda setting and framing by viewing frames as a second-level kind of agenda [1]: just as agenda setting is about which objects of discussion are salient, framing is about the salience of attributes of those objects. The key is that what communications theorists consider an attribute in a discussion can itself be an object, as well. For example, “mistaken convictions” is one attribute of the death penalty discussion, but it can also be viewed as an object of discussion in its own right.

This two-level view leads naturally to the idea of using a hierarchical topic model to formalize both agendas and frames within a uniform setting. In this paper, we introduce a new model to do exactly that. The model is predictive: it represents the idea of alternative or competing perspectives via a continuous-valued response variable. Although inspired by the study of political discourse, associating texts with “perspectives” is more general and has been studied in sentiment analysis, discovery of regional variation, and value-sensitive design. We show experimentally that the model’s hierarchical structure improves prediction of perspective in both a political domain and on sentiment analysis tasks, and we argue that the topic hierarchies exposed by the model are indeed capturing structure in line with the theory that motivated the work.
Figure 1: SHLDA’s generative process and plate diagram. Words \( w \) are explained by topic hierarchy \( \phi \), and response variables \( y \) are explained by per-topic regression coefficients \( \eta \) and global lexical coefficients \( \tau \).

2 SHLDA: Combining Supervision and Hierarchical Topic Structure

Jointly capturing supervision and hierarchical topic structure falls under a class of models called supervised hierarchical latent Dirichlet allocation. These models take as input a set of \( D \) documents, each of which is associated with a response variable \( y_d \), and output a hierarchy of topics which is informed by \( y_d \); Zhang et al. [6] introduce the SHLDA family, focusing on a categorical response. In contrast, our novel model (which we call SHLDA for brevity), uses continuous responses. At its core, SHLDA’s document generative process resembles a combination of hierarchical latent Dirichlet allocation [7, HLD] and the hierarchical Dirichlet process [8, HDP]. HLD uses the nested Chinese restaurant process (nCRP(\( \gamma \))), combined with an appropriate base distribution, to induce an unbounded tree-structured hierarchy of topics: general topics at the top, specific at the bottom. A document is generated by traversing this tree, at each level creating a new child (hence a new path) with probability proportional to \( \gamma \) or otherwise respecting the “rich-get-richer” property of a CRP.

A drawback of HLD, however, is that each document is restricted to only a single path in the tree. Recent work relaxes this restriction through different priors: nested HDP [9], nested Chinese franchises [10] or recursive CRPs [11]. In this paper, we address this problem by allowing documents to have multiple paths through the tree by leveraging information at the sentence level using the two-level structure used in HDP. More specifically, in the HDP’s Chinese restaurant franchise metaphor, customers (i.e., tokens) are grouped by sitting at tables and each table takes a dish (i.e., topic) from the global menu. In our SHLDA, dishes are organized in a tree-structured global menu by using the nCRP as prior. Each path in the tree is a collection of \( L \) dishes (one for each level) and is called a combo. SHLDA groups sentences of a document by assigning them to tables and associates each table with a combo, and thus, models each document as a distribution over combos.\(^1\)

In SHLDA’s metaphor, customers come in a restaurant and sit at a table in groups, where each group is a sentence. A sentence \( w_{d,s} \) enters restaurant \( d \) and selects a table \( t \) (and its associated combo) with probability proportional to the number of sentences \( S_{d,t} \) at that table; or, it sits at a new table with probability proportional to \( \alpha \). After choosing the table (indexed by \( l_{d,s} \)), if the table is new, the group will select a combo of dishes (i.e., a path, indexed by \( c_{d,t} \)) from the tree menu. Once a combo is in place, each token in the sentence chooses a “level” (indexed by \( z_{d,s,n} \)) in the combo, which specifies the topic \( \phi_{c_{d,t}} \) producing the associated observation (Figure 2).

SHLDA also draws on supervised LDA [12, SLDA] associating each document \( d \) with an observable continuous response variable \( y_d \) that represents the author’s perspective toward a topic, e.g., positive vs. negative sentiment, conservative vs. liberal ideology, etc. This lets us infer a multi-level topic structure informed by how topics are “framed” with respect to positions along the \( y_d \) continuum.

\(^1\)We emphasize that, unlike in HDP where each table is assigned to a single dish, each table in our metaphor is associated with a combo—a collection of \( L \) dishes. We also use combo and path interchangeably.
Before describing each step in detail, let us define the following probabilities. For more thorough derivations, please see the supplement.

3 Posterior Inference and Optimization

Given documents with observed words $w = \{w_{d,s,n}\}$ and response variables $y = \{y_d\}$, the inference task is to find the posterior distribution over: the tree structure including topic $\phi_k$ and regression parameter $\eta_k$ for each node $k$, combo assignment $c_{d,t}$ for each table $t$ in document $d$, table assignment $t_{d,s}$ for each sentence $s$ in a document $d$, and level assignment $z_{d,s,n}$ for each token $w_{d,s,n}$. We approximate SHLDA’s posterior using stochastic EM, which alternates between a Gibbs sampling E-step and an optimization M-step. More specifically, in the E-step, we integrate out $\psi$, $\theta$ and $\phi$ to construct a Markov chain over $(t, c, z)$ and alternate sampling each of them from their conditional distributions. In the M-step, we optimize the regression parameters $\eta$ and $\tau$ using L-BFGS [13].

Before describing each step in detail, let us define the following probabilities. For more thorough derivations, please see the supplement.
• First, define \( v_{d,x} \) as a set of tokens (e.g., a token, a sentence or a set of sentences) in document \( d \). The conditional density of \( v_{d,x} \) being assigned to path \( c \) given all other assignments is

\[
f^{-d,x}_{c}(v_{d,x}) = \prod_{l=1}^{L} \frac{\Gamma(C^{-d,x}_{c,l} + V \beta)}{\Gamma(C^{-d,x}_{c,l} + \beta)} \prod_{v=1}^{V} \frac{\Gamma(C^{-d,x}_{c,v} + C_{d,v,t} + \beta)}{\Gamma(C^{-d,x}_{c,v} + \beta)}
\]

where superscript \(-d,x\) denotes the same count excluding assignments of \( v_{d,x} \); marginal counts are represented by \( \cdot \)'s. For a new path \( c^{new} \), if the node does not exist, \( C^{-d,x}_{c^{new},t,v} = 0 \) for all word types \( v \).

• Second, define the conditional density of the response variable \( y_d \) of document \( d \) given \( v_{d,x} \) being assigned to path \( c \) and all other assignments as \( g^{-d,x}_{c}(y_d) = \)

\[
N \left( \frac{1}{N_{d,c}} \sum_{w_{d,s,n} \in \{ w_{d,t,l} \}} \eta_{d,t,l} z_{d,s,n} + \sum_{i=1}^{l} \eta_{d,t,l} \cdot C_{d,t,l} + \sum_{s=1}^{S_{d,t,l}} \sum_{v} \eta_{d,t,l} \cdot \sum_{j=1}^{V_{d,t,l,v}} \alpha_{d,t,l,v} \right)^{\rho} \]

(2)

where \( N_{d,c} \) is the total number of tokens in document \( d \). For a new node at level \( l \) on a new path \( c^{new} \), we integrate over all possible values of \( \eta_{c^{new},t,l} \).

**Sampling t:** For each group \( w_{d,s} \), we need to sample a table \( t_{d,s} \). The conditional distribution of a table \( t \) given \( w_{d,s} \) and other assignments is proportional to the number of sentences sitting at \( t \) times the probability of \( w_{d,s} \) and \( y_d \) being observed under this assignment. This is

\[
P(t_{d,s} = t \mid \text{rest}) \propto P(w_{d,s}, y_d \mid t_{d,s} = t, w^{-d,s} \mid t_{d,s} = t, z, c, \eta)
\]

\[
\propto \begin{cases} 
S_{d,s} \cdot f_{d,s}(w_{d,s}) \cdot g_{d,s}(y_d), & \text{for existing table } t; \\
\alpha \cdot \sum_{c \in C} P(c_{d,t,new} = c) \cdot f^{-d,s}_{c}(w_{d,s}) \cdot g^{-d,s}_{c}(y_d), & \text{for new table } t_{d,new}.
\end{cases}
\]

(3)

For a new table \( t_{d,new} \), we need to sum over all possible paths \( c^{+} \) of the tree, including new ones. For example, the path \( c^{+} \) for the tree shown in Figure 2 consists of four existing paths (ending at one of the four leaf nodes) and three possible new paths (a new leaf off of one of the three internal nodes). The prior probability of path \( c \) is:

\[
P(c_{d,t,new} = c \mid c^{-d,s}) \propto \]

\[
\prod_{l=2}^{L} \left( \frac{M^{-d,s}_{c,l}}{M^{-d,s}_{c,l-1} + \gamma_{l-1}} \right) \quad \text{for an existing path } c;
\]

\[
\prod_{l=2}^{L} \left( \frac{M^{-d,s}_{c,new,l}}{M^{-d,s}_{c,new,l-1} + \gamma_{l-1}} \right) \quad \text{for a new path } c_{d,new} \text{ which consists of an existing path from the root to a node at level } l^{*} \text{ and a new node.}
\]

(4)

**Sampling z:** After assigning a sentence \( w_{d,s} \) to a table, we assign each token \( w_{d,s,n} \) to a level to choose a dish from the combo. The probability of assigning \( w_{d,s,n} \) to level \( l \) is

\[
P(z_{d,s,n} = l \mid \text{rest}) \propto P(z_{d,s,n} = l \mid z^{-d,s}_{d,s}) \cdot P(w_{d,s,n}, y_d \mid z_{d,s,n} = l, w^{-d,s,n}, z^{-d,s,n}, t, c, \eta)
\]

(5)

The first factor captures the probability that a customer in restaurant \( d \) is assigned to level \( l \), conditioned on the level assignments of all other customers in restaurant \( d \), and is equal to

\[
P(z_{d,s,n} = l \mid z^{-d,s}_{d,s}) = \frac{m \pi + N^{-d,s,n}_{d,c,l-1} \prod_{j=1}^{l-1} (1 - m) \pi + N^{-d,s,n}_{d,c,j},}{\pi + N^{-d,s,n}_{d,c,l-1} \prod_{j=1}^{l-1} (1 - m) \pi + N^{-d,s,n}_{d,c,j}}
\]

The second factor is the probability of observing \( w_{d,s,n} \) and \( y_d \) given that \( w_{d,s,n} \) is assigned to level \( l \): \( P(w_{d,s,n}, y_d \mid z_{d,s,n} = l, w^{-d,s,n}, z^{-d,s,n}, t, c, \eta) \)

\[
f_{c_{d,s,n}}^{c_{d,t,new}}(w_{d,t}) \cdot g_{c_{d,s,n}}^{c_{d,t,new}}(y_d)
\]

**Sampling e:** After assigning customers to tables and levels, we also sample path assignments for all tables. This is important since it can change the assignments of all customers sitting at a table, which leads to a well-mixed Markov chain and faster convergence. The probability of assigning table \( t \) in restaurant \( d \) to a path \( c \) is

\[
P(c_{d,t} = c \mid \text{rest}) \propto P(c_{d,t} = c \mid c^{-d,t}) \cdot P(w_{d,t}, y_d \mid c_{d,t} = c, w^{-d,t}, c^{-d,t}, t, z, \eta)
\]

(6)

where we slightly abuse the notation by using \( w_{d,t} \equiv \bigcup_{l \leq l_{d,s,n}} w_{d,s,n} \) to denote the set of customers in all the groups sitting at table \( t \) in restaurant \( d \). The first factor is the prior probability of a path given all tables’ path assignments \( c^{-d,t} \), excluding table \( t \) in restaurant \( d \) and is given in Equation 4.

The second factor in Equation 6 is the probability of observing \( w_{d,t} \) and \( y_d \) given the new path assignments, \( P(w_{d,t}, y_d \mid c_{d,t} = c, w^{-d,t}, c^{-d,t}, t, z, \eta) \)

\[
f_{c_{d,t}}^{c_{d,t,new}}(w_{d,t}) \cdot g_{c_{d,t}}^{c_{d,t,new}}(y_d)
\]
Optimizing $\eta$ and $\tau$: We optimize the regression parameters $\eta$ and $\tau$ via the likelihood,
\[
L(\eta, \tau) = -\frac{1}{2\rho} \sum_{d=1}^{D} (y_d - \eta^T z_d - \tau^T w_d)^2 - \frac{1}{2\sigma} \sum_{k=1}^{K^+} (\eta_k - \mu)^2 - \frac{1}{\omega} \sum_{v=1}^{V} |\tau_v|, \tag{7}
\]
where $K^+$ is the number of nodes in the tree. This maximization is performed using L-BFGS [13].

4 Data: Congress, Products, Films

We conduct our experiments using three datasets: Congressional floor debates, Amazon product reviews, and movie reviews. For all datasets, we remove stopwords, add bigrams to the vocabulary, and filter the vocabulary using tf-idf.

- **US Congressional floor debates:** We downloaded debates of the 109\textsuperscript{th} US Congress from GovTrack\footnote{http://www.govtrack.us/data/us/109/} and preprocessed them as in Thomas et al. [14]. To remove uninterestingly non-polarized debates, we ignore bills with less than 20\% “Yea” votes or less than 20\% “Nay” votes. Each document $d$ is a turn (a continuous utterance by a single speaker, i.e., speech segment [14]), and its response variable $y_d$ is the first dimension of the speaker’s DW-NOMINATE score [15], which captures the traditional left-right political distinction.\footnote{Scores were downloaded from http://voteview.com/dwnomin_joint_house_and_senate.htm} After processing, our corpus contains 5,201 turns in the House, 3,060 turns in the Senate, and 5,000 words in the vocabulary.

- **Amazon product reviews:** From a set of Amazon reviews of manufactured products such as computers, MP3 players, GPS devices, etc. [16], we focused on the 50 most frequently reviewed products. After filtering, this corpus contains 37,191 reviews with a vocabulary of 5,000 words. We use the rating associated with each review as the response variable $y_d$.\footnote{Data will be available after blind review.}

- **Movie reviews:** Our third corpus is a set of 5,006 reviews of movies [17], again using review ratings as the response variable $y_d$, although in this corpus the ratings are normalized to the range from 0 to 1. After preprocessing, the vocabulary contains 5,000 words.

5 Evaluating Prediction

SHLDA’s response variable predictions provide a formally rigorous way to assess whether it is an improvement over prior methods. We evaluate effectiveness in predicting values of the response variables for unseen documents in the three datasets. For comparison we consider these baselines:

- Multiple linear regression (MLR) models the response variable as a linear function of multiple features (or regressors). Here, we consider two types of features: topic-based features and lexically-based features. Topic-based MLR, denoted by MLR-LDA, uses the topic distributions learned by vanilla LDA as features [12], while lexically-based MLR, denoted by MLR-VOC, uses the frequencies of words in the vocabulary as features. MLR-LDA-VOC uses both features.

- Support vector regression (SVM) is a discriminative method [18] that uses LDA topic distributions (SVM-LDA), word frequencies (SVM-VOC), and both (SVM-LDA-VOC) as features.\footnote{The ratings can range from 1 to 5, but skew positive.}

- Supervised topic model (SLDA): we implemented SLDA using Gibbs sampling. The version of SLDA we use is slightly different from the original SLDA described in [12], in that we place a Gaussian prior $\mathcal{N}(0, 1)$ over the regression parameters to perform L2-norm regularization.\footnote{This performs better than unregularized SLDA in our experiments.}

For parametric models (LDA and SLDA), which require the number of topics $K$ to be specified beforehand, we use $K \in \{10, 30, 50\}$. We use symmetric Dirichlet priors in both LDA and SLDA, initialize
Although a formal coherence evaluation [21] remains a goal for future work, a qualitative look at
the topic hierarchy uncovered by the model suggests that it is indeed capturing agenda/framing
structure as discussed in Section 1. In Figure 3, a portion of the topic hierarchy induced from the
Congressional debate corpus, Nodes A and B illustrate agendas—issues introduced into political
discourse—associated with a particular ideology: Node A focuses on the hardships of the poorer
victims of hurricane Katrina and is associated with Democrats, and text associated with Node E
discusses a proposed constitutional amendment to ban flag burning and is associated with Republicans.
Nodes C and D, children of a neutral “tax” topic, reveal how parties frame taxes as gains in terms of
new social services (Democrats) and losses for job creators (Republicans).

### 6 Qualitative Analysis: Agendas and Framing/Perspective

Although a formal coherence evaluation [21] remains a goal for future work, a qualitative look at
the topic hierarchy uncovered by the model suggests that it is indeed capturing agenda/framing
structure as discussed in Section 1. In Figure 3, a portion of the topic hierarchy induced from the
Congressional debate corpus, Nodes A and B illustrate agendas—issues introduced into political
discourse—associated with a particular ideology: Node A focuses on the hardships of the poorer
victims of hurricane Katrina and is associated with Democrats, and text associated with Node E
discusses a proposed constitutional amendment to ban flag burning and is associated with Republicans.
Nodes C and D, children of a neutral “tax” topic, reveal how parties frame taxes as gains in terms of
new social services (Democrats) and losses for job creators (Republicans).
Figure 3: Topics discovered from Congressional floor debates. Many first-level topics are bipartisan (purple), while lower level topics are associated with specific ideologies (Democrats blue, Republicans red). For example, the “tax” topic (B) is bipartisan, but its Democratic-leaning child (D) focuses on social goals supported by taxes (“children”, “education”, “health care”), while its Republican-leaning child (C) focuses on business implications (“decrease tax”, “jobs”, “businesses”). The number below each topic denotes the magnitude of the learned regression parameter associated with that topic. Colors and the numbers beneath each topic show the regression parameter $\gamma$ associated with the topic.

Figure 4 shows the topic structure discovered by SHLDA in the review corpus. Nodes at higher levels are relatively neutral, with relatively small regression parameters. These nodes have general topics with an overall polarity. However, the bottom level clearly illustrates the positive/negative perspective. For example, Node A concerns washbasins for infants, and has two polarized children nodes: reviewers take a positive perspective when their children enjoy the product (Node B: “loves”, “splash”, “play”) but have negative reactions when it leaks (Node C: “leak(s/ed/ing)”).

Figure 4: Topics discovered from Amazon reviews. Higher topics are general, while lower topics are more specific. The polarity of the review is encoded in the color: red (negative) to blue (positive). Many of the first-level topics have no specific polarity and are associated with a broad class of products such as “routers” (Node D). However, the lowest topics in the hierarchy are often polarized; one child topic of “router” focuses on upgradable firmware such as “tomato” and “ddwrt” (Node E, positive) while another focuses on poor “tech support” (Node F, negative). The number below each topic is the regression parameter $\tau$ associated with that topic.

In addition to the per-topic regression parameters, SHLDA also associates each word with a lexical regression parameter $\gamma$. Table 3 shows the top ten words with highest and lowest $\gamma$. The results are unsurprising, although the lexical regression for the Congressional debates is less clear-cut than other

10All of the nodes at the bottom have slightly negative values for the regression parameters mainly due to the very skewed distribution of the review ratings in Amazon.
datasets. As we saw in Section 5, for similar datasets, SHLDA’s context-specific regression is more useful when global lexical weights do not readily differentiate documents.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Top 10 words with positive weights</th>
<th>Top 10 words with negative weights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Floor Debates</td>
<td>bringing, private_property, illegally, tax_relief, regulation, mandates, constitutional, committee_report, illegal_alien</td>
<td>bush_administration, strong_opposition, ranking, republicans, republican_leadership, secret, discriminate, majority, undermine</td>
</tr>
<tr>
<td>Amazon Reviews</td>
<td>highly_recommend, pleased, love, loves, perfect, easy, excellent, amazing, glad, happy</td>
<td>waste, returned, return, stopped, leak, junk, useless, returning, refund, terrible</td>
</tr>
<tr>
<td>Movie Reviews</td>
<td>hilarious, fast, schindler, excellent, motion_pictures, academy_award, perfect, journey, fortunately, ability</td>
<td>bad, unfortunately, supposed, waste, mess, worst, acceptable, awful, suppose, boring</td>
</tr>
</tbody>
</table>

Table 3: Top words based on the global lexical regression coefficient, $\tau$. For the floor debates, positive $\tau$’s are Republican-leaning while negative $\tau$’s are Democrat-leaning.

7 Related Work

SHLDA joins a family of LDA extensions that introduce hierarchical topics, supervision, or both. Owing to limited space, we focus here on related work that combines the two. Petinot et al. [22] propose hierarchical Labeled LDA (hLDA), which leverages an observed document ontology to learn topics in a tree structure; however, hLDA assumes that the underlying tree structure is known a priori. SSHLDA [23] generalizes hLDA by allowing the document hierarchy labels to be partially observed, with unobserved labels and topic tree structure then inferred from the data. Boyd-Graber and Resnik [24] used hierarchical distributions within topics to learn topics across languages. In addition to these “upstream” models [25], Perotte et al. [26] propose a “downstream” model called HSLDA, which jointly models documents’ hierarchy of labels and topics. HSLDA’s topic structure is flat, however, and the response variable is a hierarchy of labels associated with each document, unlike SHLDA’s continuous response variable. Finally, another body related body of work includes models that jointly capture topics and other facets such as ideologies/perspectives [27, 28] and sentiments/opinions [29], albeit with discrete rather than continuously valued responses.

Computational modeling of sentiment polarity is a voluminous field [30], and many computational political science models describe agendas [5] and ideology [31]. Looking at framing or bias at the sentence level, Greene and Resnik [32] investigate the role of syntactic structure in framing, Yano et al. [33] look at lexical indications of sentence-level bias, and Recasens et al. [34] develop linguistically informed sentence-level features for identifying bias-inducing words.

8 Conclusion

We have introduced SHLDA, a model that associates a continuously valued response variable with hierarchical topics to capture both the issues under discussion and alternative perspectives on those issues. The two-level structure improves predictive performance over existing models on multiple datasets, while also adding potentially insightful hierarchical structure to the topic analysis. Based on a preliminary qualitative analysis, the topic hierarchy exposed by the model plausibly captures the idea of agenda setting, which is related to the issues that get discussed, and framing, which is related to authors’ perspectives on those issues. We plan to analyze the topic structure produced by SHLDA with political science collaborators and more generally to study how SHLDA and related models can help analyze and discover useful insights from political discourse.

Acknowledgments

This research was supported in part by NSF under grant #1211153 (Resnik) and #1018625 (Boyd-Graber and Resnik). Any opinions, findings, conclusions, or recommendations expressed here are those of the authors and do not necessarily reflect the view of the sponsor.
References
