HDPsent: Incorporation of Latent Dirichlet Allocation for Aspect-Level Sentiment into Hierarchical Dirichlet Process-Based Topic Models

Ming Yang, William H. Hsu

Kansas State University
{yangming, bhsu}@ksu.edu

Abstract

We address the problem of combining topic modeling with sentiment analysis within a generative model. While the Hierarchical Dirichlet Process (HDP) has seen recent widespread use for topic modeling alone, most current hybrid models for concurrent inference of sentiments and topics are not based on HDP. In this paper, we present *HDPsent*, a new model which incorporates Latent Dirichlet Allocation (LDA)-based sentiment learning into an HDP topic modeling framework. This model preserves the benefits of nonparametric Bayesian models for topic learning, while simultaneously learning latent sentiment aspects. It automatically generates different word distributions for each single sentiment polarity within each topic that has been learned. We present results using existing corpora consisting of multi-aspect hotel and restaurant reviews, and discuss ramifications and applications of such a model for product reviews that are intrinsically hierarchical.

1. Introduction

1.1. Problem Statement: Sentiment-Topic Models

With the growing need for analyses of free text that extract both feature information and sentiment polarity, hybrid probabilistic models that support concurrent topic and sentiment analysis have also increased in relevance and significance. We seek to infer the topics of documents, but also want to infer the sentiment information for these topics. However, many models treat topic modeling and sentiment analysis as separate and independent processes, an approach that lacks the ability to isolate sentiment polarity from different topics.

For example, when we analyze product or service reviews, it is crucial that we have separate sentiment polarity information for each feature aspect, which helps us to differentiate opinion words for different aspects from one review text. This, in turn, extends our ability to perform feature-specific sentiment polarity analysis.

1.2. Objectives and Significance

We present a technique for simultaneously inferring sentiment and topic from free text, extending existing Hierarchical Dirichlet process (HDP) models presented in (Blei et al., 2006). This approach uses Gibbs sampling for inference, as do implementations of the Chinese restaurant franchise process (CRFP) presented in (Teh et al., 2006) for the generative HDP model. The purpose of this approach is to enable applications of aspect-level topic/sentiment inference such as sentiments about specific aspects in product reviews. Algorithms for hybrid inference, such as (Lin and He, 2009), (Mei et al., 2007), *etc.* exist, but they do not fully make use of the current state of the field in nonparametric Bayesian HDP models as a representational framework. Our model is the first to extend the existing Hierarchical Dirichlet Process (HDP) model by adding a sentiment label l along with a topic label k to each token in a document.

Here we assume that each token in a document not only carries latent topic information, but also represents the intended sentiment of the writer. Therefore, while HDP only assigns a topic label k to each word, we add a sentiment label l to each word, along with its topic label k. We assume that for each topic component existing in each document, there is a sentiment distribution for it. Thus, each word is sampled from a word distribution specifically for the combination of its topic and sentiment label. The number of sentiment polarity values is always small and well-defined in advance. In our model, we therefore fix the number of sentiment labels in advance, which follows the conventional approach in the area of sentiment analysis research. We set $L = \{positive, negative, neutral\}$, which denote positive words, negative words, and descriptive words separately. Because of the simplicity and non-hierarchical (flat) nature of this independent semantic component, we use a Latent Dirichlet Allocation (LDA) model (Blei et al., 2003) for latent sentiment label allocation, while using a nonparametric Bayesian HDP model.

There are several other advantages for our model. First and foremost is that it enables us to infer different word distributions for the same topic, with different sentiment polarities. Thus, from different word distributions for different sentiment polarities, we can isolate descriptive words, positive words, and negative words from the same topic. Another advantage is that our model makes it possible to infer sentiment distributions for each topic mentioned in the document. This will allow researchers and users to develop a deeper and more detailed sentiment analysis for not only the whole document, but also each different aspect in the document. This would potentially aid them in differentiating the distinct views of an author towards the topic aspects that are reflected within a document.

2. Related Work

Some significant work in the past decade has begun to combine topic modeling and sentiment analysis in a single model. In applications of the Topic Sentiment Mixture Model (TSM) developed by (Mei et al., 2007), a Probabilistic Latent Semantic Analysis (PLSA) model is used to represent the generative process. Furthermore, even it assigns topic label for each word (excluding background

words), that word itself is sampled from either general positive, negative model, or that specific topic model. This generative process generalizes sentiment polarity model and has limited ability to make different sentiment polarity word distributions for different topics. However, our intuition is that different topics might treat same words with different sentiment strength, or even different polarity. For example, the word "small" might be a positive word when it is describing the size of a MP3 player, but might be a negative word when it is describing the storage capacity of that MP3 player. One approach to handling this problem is word sense induction (Elshamy et al., 2010), which is beyond the scope of this work.

Our model is mainly inspired by and builds upon the Joint Topic/Sentiment Model of (Lin and He, 2009), which uses a Latent Dirichlet Allocation (LDA) model in topic modeling to incorporate sentiment analysis. In this model it is assumed that each word is labeled using both a topic label k and a sentiment label l, and that each word is sampled from a word distribution given both k and l. However, this inherits several basic limitations from LDA which the overall model incurs. It predefines and limits the number of topics K initially, which is impractical for large corpora. For example, for a large corpus with various service/product reviews (such as *Yelp* review data (Yelp, 2012)), it is hard for users to regulate the number of topics in advance. Furthermore, it is also inappropriate for users to predefine this parameter, since the number of total features would be extremely large but each review document would only occupy a few of them. The nonparametric Bayesian features of HDP can help us to alleviate this problem.

Other hybrid approaches include multi-grain topic models, cf. Titov and McDonald (Titov and McDonald, 2008), which have some flexibility with respect to local (aspect-level) topics, but are predominantly LDA-based and tied to fixed, preset numbers of topics. Yet another approach is constrained LDA cf. Zhai, Liu, Xu, Jia ((Zhai et al., 2011)), which uses clustering approaches to discovery synonymy (synonym sets) of words taken as feature terms. Both of these techniques are aimed at incorporating sentiment into LDA as a monolithic topic model and thus have limited ability to evolve a topic hierarchy (Teh et al., 2006), account for dynamic topic drift, and incorporate models of topics in relation to authors.

3. The *HDPsent* Topic-Sentiment Model

In this section we derive our extended hierarchical Dirichlet process (HDP) model, *HDPsent*, which augments the traditional nonparametric Bayesian HDP model for topic learning with LDA-based parameters for aspect-level sentiment.

3.1. Model Definition

We define $D = \{d_1, d_2, ..., d_m\}$ to be the corpus (document set) that we want to analyze, and $x_j = \{x_{j1}, x_{j2}, ...\}$ to be the word array in document d_j. As in traditional HDP-based topic models, ours treats each document as a bag of words, so that the positions of words in the same document are interchangeable. We then assume that each word x_{ji} is associated with a latent dyadic topic-sentiment combination label, denoted $< k, l >$, where k denotes a topic label and l a sentiment label from a predefined sentiment set L. In this paper we set L = {*positive, negative, neutral* }, denoting positive words, negative words, and descriptive words.

We extend the existing Hierarchical Dirichlet Process (HDP) model to accommodate sentiment label l for word x_{ji} as in Figure 1:

In this model, the global probability measure G_0 is drawn from a Dirichlet process with two generative hyperparameters: a base measure H and a concentration parameter γ. Each document j then generates its own probability measure G_j from a Dirichlet process with G_0 as its base measure and α_0 as a concentration parameter:

$$\begin{aligned} &G_0|\gamma, H \sim DP(\gamma, H) \\ &G_j|\alpha_0, G_0 \sim DP(\alpha_0, G_0) \quad \text{for each } j, \end{aligned} \tag{1}$$

Each observation x_{ji} in document j position i has two parameters, θ_{ji} and l_{ji}. θ_{ji} is independently and identically distributed (i.i.d.), drawn from G_j. Because each θ_{ji} is associated with an observation ψ_{jt}, which in turn has a corresponding factor k_{jt} sampled from G_0, we can denote $\theta_{ji} = \psi_{jt}$, $\psi_{jt} = \phi_k$ where $k_{jt} = k$. In our model, for each distinct k that emerges in document j, we assume that there is a particular sentiment distribution for k denoting the author's subjective attitude towards this topic. Therefore, we generate a Dirichlet distribution σ_{jk} over the sentiment label set L, which denotes the sentiment distribution for topic k in document j, with $Dir(\tau)$ as its conjugate prior. The sentiment label l_{ji} for observation x_{ji} is then drawn from this distribution, given its topic label k. This is given by:

$$\begin{aligned} &\sigma_{jk} \sim Dir(\tau) && \text{for each existing } k \text{ in each } j, \\ &\theta_{ji}|G_j \sim G_j && \text{for each } j \text{ and } i, \\ &l_{ji} \sim Mult(\sigma_{jk_{\theta_{ji}}}) && \text{for each } j \text{ and } i, \end{aligned} \tag{2}$$

We want to not only discover differences in word distributions between corresponding sentiment polarities in different topics, but also differentiate the word distributions for the same topic with different sentiment polarities. Therefore, we assume that each distinct $< k, l >$ combination should form a distinct word distribution. Here we use $F(k, l)$ to denote a Dirichlet distribution over the whole vocabulary for a specific $< k, l >$ combination, which uses $Dir(\beta)$ as its conjugate prior. Then each observation x_{ji} is drawn from this distribution with the latent $< \theta_{ji}, l_{ji} >$ generated by the generative model:

$$\begin{aligned} &F(k, l) \sim Dir(\beta) \\ &x_{ji}|\theta_{ji}, l_{ji} \sim F(k, l) \quad \text{for each } j \text{ and } i, \end{aligned} \tag{3}$$

3.2. Inference

In this section, we want to use the extended *Chinese restaurant franchise process* (CRFP) generative model that we described above to infer the Gibbs sampling schema for the *HDPsent* model.

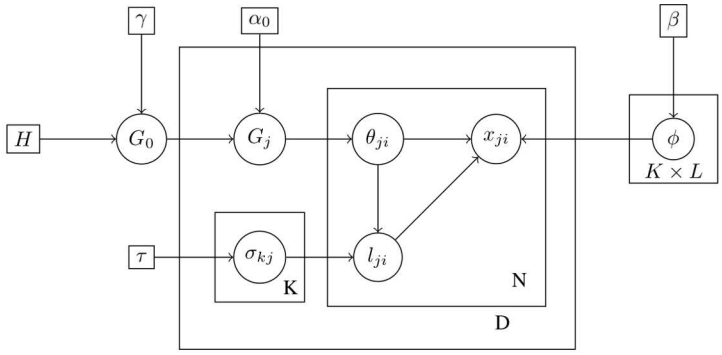

Figure 1: Plate model for HDPsent generative process

Here we define $\boldsymbol{\theta}^{-ji}$ and $\boldsymbol{l}^{-ji}$ as latent labels of all data items except observation x_{ji}:

$$\begin{aligned} \boldsymbol{\theta}^{-ji} &:= \{\theta_{j'i'} | j'i' \neq ji\} \\ \boldsymbol{l}^{-ji} &:= \{l_{j'i'} | j'i' \neq ji\} \end{aligned} \tag{4}$$

We assume in this model that each word is drawn from $F(< \theta_{ji}, l_{ji} >) = \phi_{kl}$, which is dependent on the combination of θ_{ji} and l_{ji}. We also assume that the latent sentiment label l_{ji} is drawn from a Dirichlet sentiment distribution for the specific topic parameter factor θ_{ji} in document d_j. Thus, we obtain the posterior conditional of $< \theta_{ji}, l_{ji} >$:

$$\begin{aligned} &p(\theta_{ji}, l_{ji} | x_{ji}, \boldsymbol{\theta}^{-ji}, \boldsymbol{l}^{-ji}) \\ &\propto p(x_{ji} | \theta_{ji}, l_{ji}) \cdot p(l_{ji} | \boldsymbol{l}^{-ji}, \boldsymbol{\theta}^{-ji}, \theta_{ji}) \cdot p(\theta_{ji} | \boldsymbol{\theta}^{-ji}) \end{aligned} \tag{5}$$

Here $p(\theta_{ji} | \boldsymbol{\theta}^{-ji})$ indicates the conditional distribution of topic factor θ_{ji} given all other data points.

We have supposed that the topic distribution for observations should follow an HDP model. To integrate out G_0 and G_j, the conditional distribution calculation for θ_{ji} in each G_j and ψ_{jt} for global G_0, should then be similar to that given in (Teh et al., 2006) equation (24) and (25), which can in turn be represented as follows:

$$\begin{aligned} &\theta_{ji} | \theta_{j1}, ..., \theta_{ji-1}, \alpha_0, G_0 \\ &\quad \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{i - 1 + \alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{i - 1 + \alpha_0} G_0 \end{aligned} \tag{6}$$

and

$$\begin{aligned} &\psi_{jt} | \psi_{11}, ..., \psi_{jt-1}, \gamma_0, H \\ &\quad \sim \sum_{k=1}^{K} \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} \delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot} + \gamma} H \end{aligned} \tag{7}$$

Now, we designate $\boldsymbol{\tau} = \{\tau_1, ..., \tau_{|L|}\}$ to represent the probability distribution of sentiment label l. Since the size of the sentiment label set L is predefined, this is a simple multinomial distribution across the document; therefore, we can simply choose a Dirichlet distribution as its conjugate prior:

$$\boldsymbol{\tau}_k \sim Dir(\boldsymbol{\sigma}) \tag{8}$$

We assume that the sentiment label for one word in a document is independent from that for other words in this document, given different topics. This also follows our intuition regarding how a document is written. A writer's personal sentiments about different topics may be quite different even within the same document.

Thus, the posterior sentiment distribution only takes into consideration the counts of sentiment labels for the same topic:

$$\begin{aligned} &p(\boldsymbol{\tau}_k | \boldsymbol{\sigma}, \boldsymbol{l}, \boldsymbol{k}) \\ &\quad \sim Dir(\sigma_1 + N_{dkl_1}, ..., \sigma_L + N_{dkl_{|L|}}) \end{aligned} \tag{9}$$

The conditional probability of sentiment label l for each data point x_{ji} can then be easily obtained by integrating τ out of equation (8), also eliminating x_{ji}:

$$\begin{aligned} &P(l_{x_{ji}} | \boldsymbol{l}^{-ji}, \boldsymbol{k}^{-ji}, \boldsymbol{\sigma}, k_{x_{ji}} = k) \\ &\quad = \int \tau_l Dir(\boldsymbol{\tau} | \sigma_1 + N^{-ji}_{dkl_1}, ..., \sigma_L + N^{-ji}_{dkl_{|L|}}) d\boldsymbol{\tau} \\ &\quad = \frac{\sigma_l + N^{-ji}_{dkl}}{\sum \sigma + N^{-ji}_{dk\cdot}} \end{aligned} \tag{10}$$

Finally, the data token x_{ji} is drawn from word distribution of $F(k, l)$. Here we assume that the conjugate prior is H, and that the conditional density depends on all data points in a topic k possessing the same sentiment label l, leaving x_{ji} out; here we can simply use ϕ_{kl} to denote this distribution. Then we can just directly use equation (30) from (Teh et al., 2006):

$$\begin{aligned} &p(x_{ji} | k, l) = f^{-xji}_{kl}(x_{ji}) = \\ &\frac{\int f(x_{ji} | \phi_{kl}) \prod_{\substack{j'i' \neq ji, \\ \theta_{j'i'} = k, \\ l_{j'i'} = l}} f(x_{j'i'} | \phi_{kl}) h(\phi_{kl}) d\phi_{kl}}{\int \prod_{\substack{j'i' \neq ji, \\ \theta_{j'i'} = k, \\ l_{j'i'} = l}} f(x_{j'i'} | \phi_{kl}) h(\phi_{kl}) d\phi_{kl}} \end{aligned} \tag{11}$$

Using all these components as derived above, we can now work out the posterior sampling schema for this extended Chinese restaurant franchise process (CRFP):

Sampling t

$$p(t_{ji}=t, l_{ji}=l|\boldsymbol{t}^{-ji}, \boldsymbol{l}^{-ji}k) \propto \begin{cases} n_{jt}^{-ji} \cdot p(l_{x_{ji}}|k, \boldsymbol{l}^{-ji}, \boldsymbol{k}^{-ji}) \cdot f_{k_{jt}l}^{-x_{ji}}(x_{ji}) \\ \text{if } t \text{ previously used,} \\ \alpha_0 \cdot p(x_{ji}|\boldsymbol{t}^{-ji}, \boldsymbol{l}^{-ji}, \boldsymbol{k}, t_{ji}=t^{new}) \\ \text{if } t \text{ is new.} \end{cases} \tag{12}$$

For the new table sampled, we can similarly derive the probability as:

$$p(k_{jt^{new}}=k, l_{ji}=l|\boldsymbol{t}, \boldsymbol{l}^{-ji}, \boldsymbol{k}^{-jt^{new}}) \propto \begin{cases} m_{\cdot k} \cdot p(l_{x_{ji}}|k, \boldsymbol{l}^{-ji}, \boldsymbol{k}^{-ji}) \cdot f_{kl}^{-x_{ji}}(x_{ji}) \\ \text{if } k \text{ previously used,} \\ \gamma \cdot p(l_{x_{ji}}|k^{new}, \boldsymbol{l}^{-ji}, \boldsymbol{k}^{-ji}) \cdot f_{k^{new}l}^{-x_{ji}}(x_{ji}) \\ \text{if } k \text{ is new.} \end{cases} \tag{13}$$

Sampling k

Sampling k for each table is a little different from the HDP process. This is because all the data points in a table share the same topic label k, but admit different sentiment labels l. Therefore, these data points may belong to different $F(k,l)$ components.

$$f_k^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) = \prod_{\substack{l \in L \\ \boldsymbol{x}_{jlt}=\{x_{ji}|x_{ji}\in t, l_{ji}=l\}}} p(l|k,d) f_{kl}^{-\boldsymbol{x}_{jlt}}(\boldsymbol{x}_{jlt}) \tag{14}$$

The probability of table t is assigned to each k follows:

$$p(k_{jt}=k|\boldsymbol{t}, \boldsymbol{l}^{-ji}, \boldsymbol{k}^{-jt}) \propto \begin{cases} m_{\cdot k} \cdot f_k^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) & \text{if } k \text{ previously used,} \\ \gamma \cdot f_{k^{new}}^{-\boldsymbol{x}_{jt}}(\boldsymbol{x}_{jt}) & \text{if } k \text{ is new.} \end{cases} \tag{15}$$

3.3. Model Prior

The traditional HDP model rarely introduces asymmetric priors for both documents and topics. Our *HDPsent* model, however, imports a sentiment layer into the traditional HDP model, which requires some structuring of asymmetric priors for sentiment modeling.

3.3.1. Sentiment Prior

In our model, the sentiment distribution is dependent only on the data for its corresponding topic. This does not cause problems in LDA models, but **does** cause problems in HDP models, because the HDP model spawns new topics at certain probabilities:

$$p(\boldsymbol{\tau}|\boldsymbol{\sigma}, \boldsymbol{l}^{-ji}, \boldsymbol{k}^{-ji}, k^{new}) \sim Dir(\sigma_1+0, ..., \sigma_{|L|}+0) = Dir(\boldsymbol{\sigma}) \tag{16}$$

Without any prior knowledge for sentiment labels for words assigned to a new or newly emerged topic, the sentiment label for this word is solely (or largely) dependent on its conjugate prior $Dir(\boldsymbol{\sigma})$.

Here we introduce different $\boldsymbol{\sigma}$ for different documents, each with its own conjugate prior. Using the LDA prior schema from (Wallach et al., 2009) for sentiment distributions, we use σ' as a concentration parameter for σ, and obtain:

$$\sigma_{dl} = \sum_l \sigma_l \cdot \frac{N_{d\cdot l} + \sigma'_l}{N_{d\cdot\cdot} + \sum_l \sigma'_l} \tag{17}$$

This allows equation (10) to be rewritten as:

$$P(l_{x_{ji}}|\boldsymbol{l}^{-ji}, \boldsymbol{k}^{-ji}, \boldsymbol{\sigma}, k_{x_{ji}}=k) = \begin{cases} \frac{\sigma_{dl}+N_{dkl}^{-ji}}{\sum_l \sigma_{dl}+N_{dk\cdot}^{-ji}} & \text{if } k \text{ previously used,} \\ \frac{\sigma_{dl}}{\sum_l \sigma_{dl}} & \text{if } k \text{ is new.} \end{cases} \tag{18}$$

3.3.2. Word Prior

Since our word distribution $F(k,l)$ has only the global conjugate prior $Dir(\beta)$, as shown in figure 1, any new $< k, l >$ combination has the same prior. In pure topic models, this is acceptable since we do not have prior knowledge for any word in the new topic at all. However, even though we do not have a prior preference for a word such as "good" in a new topic k^{new}, we shall have some prior preference for "good" in a new combination $< k^{new}, positive >$, versus a new combination $< k^{new}, negative >$.

This prior also helps us to adjust the probability for sampling word for sentiment labels. Without this prior, the sentiment assignment for words in the same topic can easily be reversed from their usual meaning, with positive words assigned to the predefined negative category, and negative ones to the positive category.

Using the same prior schema, and defining β' to be the concentration parameter for β, we directly obtain:

$$\beta_{lw} = \sum_w \beta_w \cdot \frac{N_{\cdot lw} + \beta'_w}{N_{\cdot l\cdot} + \sum_w \beta'_w} \tag{19}$$

Thus, the parameters in equation (11) can easily be integrated out, resulting in:

$$f_{kl}^{-x_{ji}}(x_{ji}) = \begin{cases} \frac{\beta_{lw}+N_{klw}^{-x_{ji}}}{\sum_w \beta_{lw}+N_{kl\cdot}^{-x_{ji}}} & \text{if } k \text{ previously used,} \\ \frac{\beta_{lw}}{\sum_w \beta_{lw}} = \frac{N_{\cdot lw}+\beta'_w}{N_{\cdot l\cdot}+\sum_w \beta'_{lw}} & \text{if } k \text{ is new.} \end{cases} \tag{20}$$

4. Experiment

4.1. TripAdvisor Review data set

The first experiment we performed as an application of the topic-sentiment model is on the *TripAdvisor* hotel review data set provided by Wang, Lu, and Zhai (Wang et al., 2010). This data set not only contains review text and overall rating values for a hotel in each review, but also contains separate rating values on eight different aspects: {*Business Service, Check in / front desk, Cleanliness, Value, Service, Location, Rooms, Sleep Quality*}. However, reviews that include aspect rating values for {*Business Service, Check in / front desk*} are too rare. Therefore, we ignored

Number of topics (sentiment polarity)	Δ^2_{aspect}	ρ_{aspect}	ρ_{review}	MAP	MAP@50
36(+)	0.792	0.350	0.627	0.691	0.854
36(-)	0.792	0.357	0.626	0.455	
137(+)	0.494	0.501	0.789	0.776	0.949
137(-)	0.427	0.518	0.816	0.730	
181(+)	0.388	0.555	0.836	0.808	0.951
181(-)	0.371	0.584	0.847	0.712	
LARA	1.190	0.180	0.425	0.657	0.703
SVR-A	1.012	-0.081	0.804	0.796	0.95
SVR-O	0.855	-0.007	0.579	0.714	0.79

Table 1: Evaluation measures for the *TripAdvisor* experiment compared to LARA and baseline models.

these two aspects, extracting only reviews with valid rating scores for the six aspects {*Cleanliness, Value, Service, Location, Rooms, Sleep Quality*}. We also filtered out short reviews to get an appropriately-sized data set that we use to learn.

We used the Stanford *CoreNLP* tool (Manning et al., 2014) to lemmatize the tokens in the review text. All stop words were also removed. We used the sentiment word list from MPQA Subjectivity Lexicon (Wilson et al., 2005) for sentiment prior initialization. Finally, we ran experiments on a data set consisting of 563 reviews.

We use similar prediction evaluation measures as introduced in (Wang et al., 2010) and (Wang and Ester,), such as:

1. Mean square error (MSE) on aspect rating prediction (Δ^2_{aspect})
2. Aspect correlation inside reviews (ρ_{aspect})
3. Aspect correlation across reviews (ρ_{review})
4. Mean Average Precision (MAP)

We use the counts of tokens labeled as positive or negative for each learned topic as feature vectors for each review, denoted $x^{(i)}_{\text{pos}}$ and $x^{(i)}_{\text{neg}}$. Next, we set the ground-truth rating value vector for six aspects, with the overall rating as the target learning value, denoted $y^{(i)} = < y_{\text{overall}}, y_{\text{cleanliness}}, y_{\text{value}}, y_{\text{service}}, y_{\text{location}}, y_{\text{rooms}}, y_{\text{sleep}} >$. We then set matrix θ_{pos} and θ_{neg} as for each $x^{(i)}_{\text{pos}}$, predicted $\hat{y}^{(i)}_{\text{pos}} = x^{(i)}_{\text{pos}} \cdot \theta_{\text{pos}}$, and for each $x^{(i)}_{\text{neg}}$, predicted $\hat{y}^{(i)}_{\text{neg}} = x^{(i)}_{\text{neg}} \cdot \theta_{\text{neg}}$. Finally, we use gradient descent to learn θ_{pos} and θ_{neg} by minimizing squared error.

MSE: We define mean squared error (MSE) as:

$$MSE = \frac{\sum_{i=1}^{D}\sum_{a=1}^{A}(\hat{y}_a^{(i)} - y_a^{(i)})^2}{D \times A} \tag{21}$$

which measures the overall rating prediction error.

ρ_{aspect}: measures the accuracy for relative ranking order of aspects being learned within review:

$$\rho_{\text{aspect}} = \frac{\sum_{i=1}^{D} \rho(\hat{y}^{(i)}, y^{(i)})}{D} \tag{22}$$

where $\rho(\hat{y}^{(i)}, y^{(i)})$ denotes the Pearson correlation coefficient between the predicted rating vector for review i and the corresponding ground-truth rating vector.

ρ_{review}: measures the accuracy for relative ranking order of reviews being learned for each aspect:

$$\rho_{\text{review}} = \frac{\sum_{a=1}^{A} \rho(\hat{y}_a, y_a)}{A} \tag{23}$$

where $\rho(\hat{y}_a, y_a)$ denotes the Pearson correlation coefficient between the predicted rating vector for aspect a across all reviews and the corresponding ground-truth rating vector.

Mean average precision (MAP): Because the ground-truth rating values are discrete numbers such as {*1.0, 2.0, 3.0, 4.0, 5.0*}, it is impractical to define the portion of top hotels as a fixed number, or a fixed percentage in our evaluation. Therefore, following (Wang et al., 2010) and (Wang and Ester, 2014), we define MAP in this experiment as the accuracy of ranking the top N hotels as top, where N is assigned dynamically as the total number of hotels in data set whose rating value is the highest value 5.0 as:

$$\begin{aligned} R_a &= \{i | y_a^{(i)} = 5.0\} \\ \hat{R}_a &= \{\text{top } |R_a| \text{ reviews predicted}\} \\ MAP &= \frac{\hat{R}_a \cap R_a}{|R_a|} \end{aligned} \tag{24}$$

We also estimate the percentage of top 50 reviews that we ranked, whose ground-truth review value is 5.0 for each aspect. We use MAP@50 to denote this value, also following the convention of (Wang et al., 2010) and (Wang and Ester, 2014).

We ran our models with different initial concentrate parameters of α_0, β and γ. Different parameters will generate different number of topics. Table 1 lists the resulting evaluation measures with different number of topics generated. We compared our results with the LARA and Support Vector Regression (SVR) models from (Wang et al., 2010).

We also use perplexity to test the convergence of this Markov chain and the performance of our model. The perplexity of our model is calculated as:

$$\begin{aligned} \text{perplexity}(\boldsymbol{w}_d|d) &= \exp\Big[-\frac{\sum_d \ln p(\boldsymbol{w}_d|d)}{\sum_d N_d}\Big] \\ p(\boldsymbol{w}_d|d) &= \prod_{x=1}^{N_d}[\sum_{k,l} p(w|k,l)p(k,l|d)] \end{aligned} \tag{25}$$

Figure 2 is a graph of the the perplexity of our model as a function of iterations:

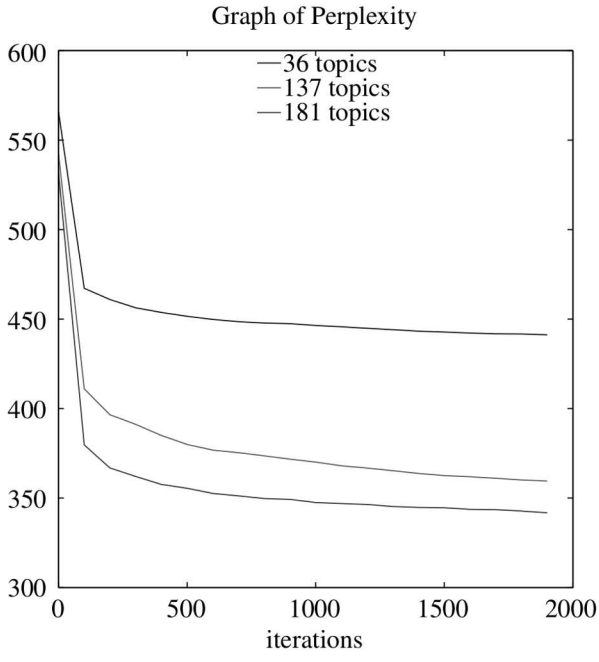

Figure 2: Graph of perplexity evolution

Topic 3		
Neutral	Positive	Negative
wedding	choose	flower
guest	great	didnt
day	marry	handle
estancia	top	yell
venue	special	dont
event	amazing	odd
reception	wonderful	stress
package	touch	bad
ceremony	love	scream

Topic 8		
Neutral	Positive	Negative
taste	fresh	side
flavor	nice	wasnt
dish	delicious	bland
sauce	tender	miss
bit	top	finish
food	enjoy	didnt
order	great	strong
sweet	love	lack
bite	touch	ill

Table 2: Table for two different topics from Yelp Reviews: Topic 3 (weddings) and 8 (restaurants)

4.2. Yelp Review data set

We performed an additional experiment using a subset of Yelp reviews [1]. The Yelp review data set itself contains reviews on restaurants, bars, beauty and spas, etc. - high variety among kinds of shops. The total number of topics in this review data set is hard to estimate, which is amenable to our nonparametric approach to developing topic and sentiment modeling algorithms.

We ran our *HDPsent* model in the same way as on the *TripAdvisor* data set on a data set of 582 reviews from Yelp. We generated 72 topics. In Table 2, we present a comparison of two different topics learned from this data set with top neutral, positive, and negative words. As an example, we can see that the most frequent neutral words about wedding ceremonies and restaurants are quite different. Also, even some generally positive words as "great", "love", "touch" occur in both topics, some words as "fresh", "delicious", "tender" only show up in restaurant-related topics, and "marry", "wonderful" only show up in wedding ceremony-related topics. Another interesting phenomena is that negation words such as "don't", "didn't", "miss", and "lack" show very often in both negative lists.

5. Conclusion

We have synthesized a Dirichlet process for aspect-level sentiment with the traditional HDP. Unlike other models, this permits the number of topics to be updated based on shared parameters of the generative topic model, rather than restricting them to a predefined, fixed set for a text document collection or to a predefined lexicon for these topics. Furthermore, it allows sentiments associated with these aspects to be inferred concurrently.

A key novel contribution of this topic model is the ability to automatically generate different topics with different word distributions for different sentiment polarities. We learn to assign weights from each topic to a set of aspects that we seek to infer using gradient descent learning. This permits empirical evaluation by calculating correlation with historical ground truth (on all reviews and ranked reviews) using the experimental test bed (*TripAdvisor*) we developed in Section 4.

This paper has focused on the design and development of an extended generative model, rather than on inference techniques for this model, for which we chose to use Gibbs sampling for ease of implementation (and parallelization). As with Gibbs sampling-based inference for traditional HDP, the main limitation of our system implementation is its lack of scalability. Our continuing work includes investigating and developing methods for approximation of this model by variational inference.

Broader applications of our inferential model thus include the discovery of new aspects not previously defined for a text corpus such as a collection of reviews. Additionally, the ability to track the evolution of aspect-level sentiments and topics over time is an important area of potential future work.

Our model requires a prior knowledge of sentiment words for initialization. However, this prior knowledge does not need to be very accurate. In learning process, it can automatically assign word tokens to different sentiment label in each topic, and is also robust to correct mistakes in prior knowledge.

[1] These reviews were retrieved from `https://www.yelp.com/academic_dataset`

6. References

Blei, David M, Michael I Jordan, et al., 2006. Variational inference for dirichlet process mixtures. *Bayesian analysis*, 1(1):121–143.

Blei, David M, Andrew Y Ng, and Michael I Jordan, 2003. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

Elshamy, Wesam, Doina Caragea, and William Hsu, 2010. Ksu kdd: Word sense induction by clustering in topic space. In *Proceedings of the 5th International Workshop on Semantic Evaluation, Uppsala, Sweden, July 2010*. Association for Computational Linguistics.

Lin, Chenghua and Yulan He, 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM.

Manning, Christopher D, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J Bethard, and David McClosky, 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Mei, Qiaozhu, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai, 2007. Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on World Wide Web*. ACM.

Teh, Yee Whye, Michael I Jordan, Matthew J Beal, and David M Blei, 2006. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476).

Titov, Ivan and Ryan McDonald, 2008. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web (WWW 2008)*. ACM Press.

Wallach, Hanna M, David Minmo, and Andrew McCallum, 2009. Rethinking lda: Why priors matter.

Wang, Hao and Martin Ester. A sentiment-aligned topic model for product aspect rating prediction.

Wang, Hao and Martin Ester, 2014. A sentiment-aligned topic model for product aspect rating prediction. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Wang, Hongning, Yue Lu, and Chengxiang Zhai, 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

Wilson, Theresa, Janyce Wiebe, and Paul Hoffmann, 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on human language technology and empirical methods in natural language processing*. Association for Computational Linguistics.

Yelp, 2012. Yelp's academic dataset. `https://www.yelp.com/academic_dataset`.

Zhai, Zhongwu, Bing Liu, Hua Xu, and Peifa Jia, 2011. Constrained lda for grouping product features in opinion mining. lecture notes in computer science. In *Lecture Notes in Computer Science*. Springer.