2012 IEEE Sixth International Conference on Semantic Computing

Hierarchical Approach to Sentiment Analysis

Noriaki Kawamae
Tokyo Denki University
5 Senju Asahi-cho, Adachi-ku, Tokyo 120-8551 Japan
Email: kawamae@gmail.com

***Abstract*—Sentiment analysis aims to extract the customer's attitude and feeling from his/her unstructured reviews by separating the subjective information from the other information. We propose a generative probabilistic topic model that detects both an aspect and corresponding sentiment, simultaneously, from review articles. Unlike existing sentiment analysis models, which generally consider rating prediction to be a side task, our proposal, the hierarchical approach to sentiment analysis, identifies both an item and its rating by dividing topics, traditionally treated as one entity, into aspect and sentiment topics. Since our model is aware of both objective and subjective information, it can discover fine-grained tightly coherent topics, and describe the generative process of each article in a unified manner. To handle the differences involved, HASA extends topic models by introducing both observed variables and a latent switch variable into each token, where topics are influenced not only by word co-occurrence but also item/rating information, and then classifying them as either aspect or sentiment topics. Experiments on review articles show that the proposed model is useful as a generative model to discover and distinguish aspects from sentiments.**

I. Introduction

Our goal is to distinguish subjective information from objective information in user-generated content such as online product review articles, web documents, and blogs. Text mining methods are needed to develop superior web services, since more and more people are expressing their opinions of product and services to the extent that the volume of such contents exceeds the analysis capabilities of any individual. Among these mining tasks, sentiment analysis, which captures customer's opinions expressed in text data, is becoming an important knowledge discovery approach given the number of web users in services such as Epinions[1] and Rotten Tomatoes[2]. These reviews are often associated with numerical ratings provided by authors for the items being discussed. This task aims to classify a text as exhibiting positive, negative or neutral sentiment [18], by extracting sentiments on each aspect of a product and the corresponding rating. Automatically discovering these sentiments from documents increases both the quality and the quantity of information available for marketing, decision making, and purchase recommendations, thereby decreasing both the time and the cost required for collecting this information in an electronic commerce [17].

Since sentiment polarities are dependent on the topics of each review, detecting both topics and their polarities simultaneously is critical for semantic analysis based on topic models. Topic models represent documents as topic distributions, where each topic is responsible for generating a word in each token. Such models provide useful descriptions for the generative process of various data and have been applied for information retrieval, social network analysis and collaborative filtering. Following this approach, more focus is being placed on systems that produce fine-grained sentiment analysis of these articles [3] to extract the desired information automatically. Although having the ratings and aspects together is helpful in decision making, most current works consider aspect identification to be the main task and treat rating prediction as, at most, a side task [13].

Motivated by the above observations, we propose a sentiment analysis model called Hierarchical Approach to Sentiment Analysis (HASA); it identifies both aspects and sentiments in a unified manner. To create HASA, we assume that each piece of subjective information can be presented as a mixture of rating and sentiment words, whereas each piece of objective information is presented as a mixture of a item and aspect words. To distinguish these differences, we extend switch Topic over Time (sTOT) [6] by using both a item and its rating instead of timestamps, and separating the trend class into an aspect topic and a sentiment class. In this model, the aspect topic has probabilistic distributions over words, items, and sentiment classes, and the sentiment class has probabilistic distributions over words and ratings. The switch variable is responsible for generating and handling various words by controlling these classes in each article. These newly introduced variables allow our approach to handle a wider variety of observed variables such words, items, and their ratings than previous models in each review article, and to explain the generative process of each article. Accordingly, HASA detects the objective information and the subjective information by using the aspect topic and the sentiment topic, respectively.

A key advantage of HASA is its ability to discover fine-grained tightly coherent topics by separating traditional topics into aspect and sentiment topics; it makes the following contributions.

(1) It discovers fine-grained tightly coherent topics, which distinguishes subjective information from objective information in online reviews by newly introducing item/rating variables, subdividing traditionally topics, and then considering the hierarchical structure of these newly identified topics.

(2)HASA models the joint distribution of words and time without using stop word lists or domain specific dictionaries, which are expensive to make and maintain.

[1]Epinions:www.epinions.com/

[2]Rotten Tomatoes:www.rottentomatoes.com/

978-0-7695-4859-3/12 $26.00 © 2012 IEEE
DOI 10.1109/ICSC.2012.62

Its performance supports automatic rating annotation and rating-based item retrieval. Since HASA is a unified probabilistic model and overcomes the limitation of conventional two-stage approaches by automatically learning the model parameters from a given data, this model removes noise words (such as typo or jargon etc.) as document-specific words, and groups synonym terms into same topics without any human supervision.

(3)HASA is easier to extend and adapt to a document collection with any other type of metadata (e.g., the number of comments, annotated tag, position information and the number of followers). Since items and ratings are one of the observed variables, HASA can replace this observed variable by other variables without loss of generality. For example, HASA can discover sentiment words in a given review data set by using the review rating instead of the timestamps. We demonstrate the efficacy of HASA through experiments and show that it can discover fine-grained tightly-coherent topics.

II. Previous works

Several topic models have been proposed for mining and summarizing sentiment in Weblogs and user reviews, as well as systems that subject user reviews to fine-grained sentiment analysis, where the main goal is to mine user opinions by identifying and extracting positive and negative opinions or analyzing and extracting topical contents. For example, Latent semantic association (LaSA) [19] models identify aspects from reviews. Joint Sentiment/Topic model (JST) [8] is completely unsupervised and detects sentiments and topics simultaneously from texts by adding an additional sentiment layer between the document and the topic layer. Structured PLSA [10] aims to generate an aggregated summary with aspect ratings inferred from overall ratings by modeling the dependency structure of phrases in short comments. Latent Aspect Rating Analysis (LARA) [15] aims at analyzing opinions expressed in each review at the level of topical aspects to discover each individual reviewer's latent rating on each aspect as well as the relative importance weight on different aspects when forming the overall judgment. Brody and Elhadad [1] propose to detect aspect-specific opinion words by taking account of the influence of aspect on sentiment polarity. MaxEnt-LDA [21] tries to jointly capture both aspect and opinion words within topic models; it allows non-adjective opinion words. Latent Evaluation Topic model (LET) [5] takes the collaborative filtering approach and focuses on the difference between writers' preferences to predict which item each writer will select and how he/she will evaluate it, based on like-minded writers' reviews. Multi-Aspect Topic Model (MAS) [14] extends MG-LDA to aggregate sentiment texts for each rating aspect extracted from MG-LDA. Yu. et al [20] proposes an aspect ranking algorithm to identify the important aspects; it is based on the assumption that the important aspects of a product should be the aspects that are frequently commented on by consumers and their opinions on the important aspects greatly influence their overall opinions of the product. Although these models discover co-occurrence information from the text, they do not explicitly separate aspect and opinion words.

TABLE I
Notation used in this paper

SYMBOL	DESCRIPTION
G	number of aspect topics
C	number of sentiment topics
D	number of documents
M	number of unique items
V	number of unique words
N_d	number of word tokens in document d
v_d	the rating associated with document d
m_d	the item associated with document d
g_i	the aspect associated with the ith token
c_i	the sentiment associated with the ith token
r_i	the switch associated with the ith token
w_i	the ith token
θ	the multinomial distribution of aspect topics ($\theta \vert \alpha \sim$ Dirichlet(α))
ψ_g	the aspect specific multinomial distribution of sentiment topics ($\psi \vert \gamma \sim$ Dirichlet(β))
ω_g	the aspect specific multinomial distribution of items ($\omega \vert \delta \sim$ Dirichlet(δ))
λ_c^1, λ_c^2	the sentiment c specific beta distribution of v
μ_d	the multinomial distribution of r associated with document d ($\mu_d \vert \epsilon \sim$ Dirichlet(ϵ))
$\phi_{c(g,d,b)}$	the multinomial distribution of words specific to sentiment s(aspect g, document d, background topic b) ($\phi_{c(g,d,b)} \vert \gamma \sim$ Dirichlet(γ))
α, β, γ	the fixed parameters of symmetric
δ, ϵ	Dirichlet priors

Interdependent LDA (ILDA) [13] model is an unsupervised model that learns a set of item aspects and corresponding ratings from a set of item reviews that have been preprocessed into a collection of opinion phrases. ILDA needs to investigate the correspondence between generated clusters and real aspects, item, and their ratings, because the relationship between these clusters and latent variables are not explicit [14] in ILDA. As many reviewing websites provide some additional information including a set of predefined aspects and their ratings, this prior knowledge can help in capturing these correspondences.

HASA aims to capture both aspect/sentiment phrases and item/rating from reviews using a set of referred items and their ratings, while many other models detect aspect or sentiment-specific words and predict ratings in separate steps. This model assumes that aspect can be detected by, first sampling both a given item and a topic, and then capturing sentiment words jointly with a given rating value and topics.

III. Hierarchical Approach to Sentiment Analysis

In this subsection, we describe HASA; it extends sTOT [6] to separate trends in explaining the generative process of review articles. Table I shows the notations used in this paper; Figure 1 shows the graphic models of HASA and the previous model, sTOT. In this paper, each article, d, is associated with a rating value, v, that is provided by the author, along with description **w**. This article is represented as the pair of variables(v, **w**), where each article addresses only one item.

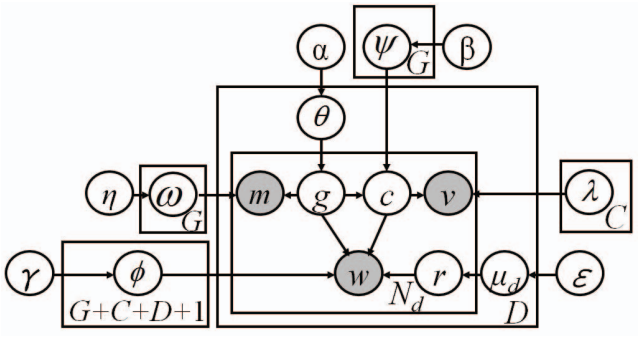

Fig. 1. Graphic Model of HASA: In this figure, shaded and unshaded variables indicate observed and latent variables, respectively. An arrow indicates a conditional dependency between variables and stacked panes indicate repeated sampling with the iteration number shown. In HASA, aspect topics are fully connected to sentiment topics, items, and words, and sentiment topics are fully connected to the rating value and words. This model defines the generative process in the framework of the hierarchical Bayesian model.

A. sTOT

Switch TOT (sTOT) focuses on differentiating trend-specific words from background topic words in each generative process of the given timestamped data. The background topic is the common topic over almost all documents regardless of their content and time, it generates non-temporal words, while the topic variable generates temporal words. For distinguishing temporal words from background topic words, sTOT defines r, it acts as a switch to handle these words in each token, and takes value r=0 if word w is generated via the background topic variable, or r=1 if word w is generated via the topic variable. Since the Beta distribution can take versatile shapes over a normalized time span covering all of the timestamped data, sTOT uses this to describe time t associated with the trend class in each document; sTOT uses it in each token, where all timestamps are normalized to lie within the range of 0 to 1. Therefore, the trend class enables sTOT to explore temporal co-occurrence patterns within a trend.

B. HASA

Hierarchical Approach to Sentiment Analysis (HASA) extends sTOT to split the trend class into aspect topic and sentiment topic. For this, the model introduces rating value v (instead of t), referred item m, and enforces r to handle more status types. One new approach is to add the document-specific word distribution into generating words. Therefore, this switching variable allows HASA to distinguish the role of words in each document, and better detect aspect and sentiment words than the existing schemes.

The procedure of generating a word in document boils down to three stages. First, one chooses aspect topic g according to distribution θ. Following that, one chooses sentiment topic c according to aspect-specific distribution ψ_g, where ψ_g is chosen conditioned on aspect topic g. It is worth noting that this characteristic differentiates HASA from conventional topic models in drawing topic distributions. HASA separates conventional topics into aspect topic and sentiment topic layer to distinguish between item-specific and rating-specific topic distributions; previous models assign words into only the topic layer without considering these differences. This feature allows HASA to discover large numbers of fine-grained, tightly coherent topics in review articles, because these two different classes capture the correlations among topics, and make the shape of word distributions more clearly different from each other. Finally, one draws a switch from the document-specific distribution μ_d. We extend r as a switch for handling more kinds of words as follows, to distinguish the differences in word tokens. If r=0, HASA selects the background topic as responsible for generating a word. If r=1, HASA selects the document-specific topic as responsible for generating a word. If r=2, HASA selects the aspect topic as responsible for generating an aspect word associated with each item. If r=3, HASA selects the sentiment topic as responsible for generating a word associated with each rating score.

C. Inference and Learning

Since HASA is a generalization of sTOT, we can infer HASA by Gibbs sampling in the same way used for sTOT without loss of generalization. The first step, defining the generative process of HASA for parameter estimation, is as follows:

1) For each document d, draw D multinomials θ_d from Dirichlet prior α;
2) For each aspect topic g, draw G multinomials ψ_g from Dirichlet prior β;
3) For each aspect topic g, draw G multinomials ω_g from Dirichlet prior η;
4) For each sentiment topic c, draw C beta distributions consist of both λ_c^1 and λ_c^2;
5) For each document d, draw D multinomials μ_d from Dirichlet prior ϵ;
6) For each aspect topic g, draw G+C+D+1 multinomials $\phi_{g(c,d,b)}$ from prior γ, (sentiment topic c, document topic d, background topic b);

For each document d:

1) For each token i in document d:
2) Draw aspect topic g_{di} from multinomial θ_d;
3) Draw item m_d from multinomial $\omega_{g_{di}}$;
4) Draw sentiment topic c_{di} from multinomial ψ_g;
5) Draw rating v_d from beta distribution with $\lambda_{c_{di}}^1$ and $\lambda_{c_{di}}^2$;
6) Draw switch r_d from multinomial μ_d;
   if r_{di} = 0 Draw word w_{di} from multinomial ϕ_b.
   else if r_{di} = 1 Draw word w_{di} from multinomial ϕ_d.
   else if r_{di} = 2 Draw word w_{di} from multinomial $\phi_{g_{di}}$.
   else Draw word w_{di} from multinomial $\phi_{c_{di}}$.

The generative model for HASA can be described as a Bayesian hierarchical model. In this inference, we need to calculate several conditional distributions. As shown in the previous subsection, the joint distribution of the entire corpus

is, therefore, the following mixture:

$$\begin{aligned} p(\mathbf{w},\mathbf{r},\mathbf{v},\mathbf{m},\mathbf{c},\mathbf{g},\phi,\mu,\psi,\theta,\omega|\alpha,\beta,\gamma,\eta,\epsilon,\lambda) &= p(\mathbf{w},\phi|\mathbf{g},\mathbf{c},\gamma) \\ \times p(\mathbf{g},\theta|\alpha)p(\mathbf{c},\psi|\mathbf{g},\beta)p(\mathbf{r},\mu|\epsilon)p(\mathbf{m},\omega|\mathbf{g})p(\mathbf{v}|\mathbf{c},\lambda) \\ = \prod_d^D \prod_i^{N_d}[P(w_{di}|\phi_{c(g,d,b)_{di}})P(c_{di}|\psi_{g_{di}})p(v_d|\lambda_{c_{di}}) \\ \times p(m_d|\omega_{g_d})P(g_{di}|\theta_d)P(r_{di}|\mu_d)] \times p(\theta|\alpha)\prod_d^D p(\mu_d|\epsilon) \\ \times \prod_g^G p(\psi_g|\beta)p(\omega_g|\eta)\prod_g^{G+C+D+1} p(\phi_g|\gamma). \end{aligned} \quad (1)$$

In this eq (1), multinomials ϕ_c, ψ_g, μ_d, and θ_{di} can be adapted by the conjugate prior and then integrated out analytically. In the Gibbs sampling procedure, we need to calculate the conditional distributions $P(g_{di}, c_{di}, r_{di}|\mathbf{g}_{\setminus di}, \mathbf{c}_{\setminus di}, \mathbf{r}_{\setminus di}, \mathbf{w}, \mathbf{m}, \mathbf{v}, \alpha, \beta, \gamma, \epsilon)$. In these distributions, $\mathbf{g}_{\setminus di}$ represent the aspect class assignments for all tokens except g_{di}, $\mathbf{s}_{\setminus di}$ represents the topic assignments for all tokens except c_{di}, and $\mathbf{r}_{\setminus di}$ represents the topic assignments for all tokens except r_{di}. Details of the derivation of Gibbs sampling for HASA are given below.

Starting from the joint distribution $P(\mathbf{w},\mathbf{r},\mathbf{v},\mathbf{m},\mathbf{c},\mathbf{g},\phi,\mu,\psi,\theta,\omega|\alpha,\beta,\gamma,\eta,\epsilon,\lambda)$, we can work out the conditional distribution $P(g_{di}=j, c_{di}=k, r_{di}=r|\mathbf{g}_{\setminus di}, \mathbf{c}_{\setminus di}, \mathbf{r}_{\setminus di}, \mathbf{v}, \alpha, \beta, \gamma, \delta, \epsilon, \lambda)$ as

$$P(j,k,r|\cdots) \propto \frac{n_{dj\setminus di}+\alpha_j}{\sum_g^G (n_{dg\setminus di}+\alpha_g)} \frac{n_{jk\setminus di}+\beta_k}{\sum_c^C n_{jc\setminus di}+\beta_c)} \times \begin{cases} \frac{n_{d0\setminus di}+\epsilon_0}{\sum_r^R n_{dr\setminus di}+\epsilon_r)} \frac{n_{bv\setminus di}+\gamma_v}{\sum_w^W n_{bw\setminus di}+\gamma_w)}, \text{if} \quad r_j=0, \\ \frac{n_{d1\setminus di}+\epsilon_1}{\sum_r^R n_{dr\setminus di}+\epsilon_r)} \frac{n_{dv\setminus di}+\gamma_v}{\sum_w^W n_{dw\setminus di}+\gamma_w)}, \text{if} \quad r_j=1, \\ \frac{n_{d2\setminus di}+\epsilon_2}{\sum_r^R n_{dr\setminus di}+\epsilon_r)} \frac{n_{jv\setminus di}+\gamma_v}{\sum_w^W n_{jw\setminus di}+\gamma_w)} \frac{n_{jm\setminus di}+\eta_m}{\sum_m^M n_{jm\setminus di}+\eta_m)}, \\ \text{if} \quad r_j=2, \\ \frac{n_{d1\setminus di}+\epsilon_3}{\sum_r^R n_{dr\setminus di}+\epsilon_r)} \frac{n_{kv\setminus di}+\gamma_v}{\sum_w^W n_{kw\setminus di}+\gamma_w)} \frac{(1-v_d)^{\lambda_k^1-1} v_d^{\lambda_k^2-1}}{B(\lambda_k^1,\lambda_k^2)}, \\ \text{if} \quad r_j=3. \end{cases} \quad (2)$$

where $n_{dj\setminus d}$ represents the number of tokens that have been assigned to aspect j in document d, except di, $n_{jk\setminus di}$ represents the number of tokens assigned to sentiment k in the tokens associated with j, except di, $n_{jm\setminus di}$ represents the number of item m in the tokens associated with j, except di, $n_{d0(1,2,3)\setminus di}$ represents the number of tokens assigned to switch r = 0;background topic (1:document specific topic, 2:aspect, 3:sentiment) in document d, except di, and $n_{bv(dv,jv,kv)\setminus di}$ represents the number of tokens assigned to word v in background topic (document specific topic, aspect specific, sentiment specific), except di, and B is the beta function. Since each word pattern and its item (rating) are assumed to have been generated conditional on aspect topic g (sentiment topic c), the resulting multinomial (beta) parameters will correspond. An item (rating) will, with high probability under a certain aspect (sentiment) class, likely contain the set of words that co-occurred with high probability in the same aspect (sentiment) class. Each new word/rating is generated by again selecting from sentiment c and repeating the entire process. Thus, we can view g (c) as a high level representation of the ensemble of word/item (word/rating) pairs in terms of probability distributions over the factors that each word/item (word/rating) can be assembled from.

Algorithm 1 Inference on HASA

```
initialize assignment randomly for all tokens
for I=1 to N_iteration do
  for j = 1 to D do
    for i = 1 to N_d do
      drawn g_di, c_di and r_di using Eq (2)
      update n_dg, n_gc, n_bw, n_dw, n_gw, n_cw, n_gm, and n_dr
    end for
    for z = 1 to C do
      update λ̂_c^1 and λ̂_c^2
    end for
  end for
end for
compute the posterior estimates of θ̂, φ̂, μ̂, ψ̂ and ω̂
```

An overview of the Gibbs sampling procedure we use is shown in the Algorithm; the posterior estimates are as follows:

$$\begin{aligned} \hat{\theta}_{dg} &= \frac{n_{dg}+\alpha_g}{\sum_g^G (n_{dg}+\alpha_g)}, \hat{\phi}_{g(c,b,d)w} = \frac{n_{g(c,b,d)w}+\gamma_w}{\sum_w^W (n_{g(c,b,d)w}+\gamma_w)}, \\ \hat{\mu}_{dr} &= \frac{n_{dr}+\epsilon_r}{\sum_r^R (n_{dr}+\epsilon_r)}, \hat{\psi}_{gc} = \frac{n_{gc}+\beta_z}{\sum_z^Z (n_{gc}+\beta_z)}, \hat{\omega}_{gm} = \frac{n_{gm}+\eta_m}{\sum_m^M (n_{gm}+\eta_m)}, \\ \hat{\lambda}_c^1 &= \bar{v}_c(\frac{\bar{v}_c(1-\bar{v}_c)}{s_c^2}-1), \hat{\lambda}_c^2 = (1-\bar{v}_c)(\frac{\bar{v}(1-\bar{v}_c)}{s_c^2}-1) \end{aligned} \quad (3)$$

IV. Experiments

A. Data sets

We focus here on the extraction of sentiments and topics, and use the following data set to quantitatively and qualitatively evaluate the proposed model.

Amazon review data[3]: This is a biggest public review data set currently available and has been used often in sentiment analysis research [4], [7]. Each entry consists of member id, product id, date, number of helpful feed back comments, number of feed back comments, rating, title and body. In these reviews, the rating score is a discrete value and varies from 1.0 to 5.0 in 0.5 increments. We normalized this value and assigned rating v in the range [0,1] to each article. We first selected the top 10000 products in terms of the number of members who reviewed the corresponding products from 01/2000 to 10/2005. This yielded a total set of 1384931 articles. Next, we selected 3 categories based on the number of reviewed products, and then split this data set into three data sets according to the product type, where product type was determined from the attached product info file by matching the product id. One set includes 4269 Book reviews, the second set includes 4158 Music reviews, and the third set includes 4360 DVD reviews. The data sets was tokenized automatically without using a stop word list. This means that we created one complete data set and three split data sets.

[3]Amazon Product Review Data (Huge): http://www.cs.uic.edu/l̃iub/FBS/sentiment-analysis.html

TABLE II
DETAILS OF DATA SETS

	DVD	Books	Music
# reviews	4360	4269	4158
# items	15	15	15
# words	11875	12512	13523

In our evaluation, the smoothing parameters α, β, γ, δ and ϵ were set to 1/Z, 0.1, 0.1, 1/Z and 0.25, respectively (all weak symmetric priors following previous work [2], [5]). The number of aspect topics, $|Z|$, was set to 15; a preliminary experiment confirmed that just one topic is enough for generating each item specific word. Additionally, we doubled the number of aspect topics so each topic with a high rating corresponds to a positive topic and one with a low score to a negative topic. As these models use the Beta distribution over rating, we normalized these values to lie in the range of 0 to 1.

Since original sTOT does not have either item or its rating score, we divide trend topics into aspect topics z_a and sentiment topics z_s, and then insert this value in each token of sTOT as the observed item m_d (score v_d) that is conditioned on each topic z_a (z_s) and can be sampled from the aspect (sentiment) topic specific multinomial distribution ($m_d \sim \omega_{z_a}$) (Beta distribution ($v_d \sim \lambda_z$)), like HASA. Additionally, we extend r as a switch to distinguish the differences in word tokens. If r=0, sTOT selects the background topic as responsible for generating a word. If r=1, sTOT selects the document-specific topic as responsible for generating a word. If r=2, sTOT selects the aspect topic as responsible for generating an aspect word associated with each item. If r=3, sTOT selects the sentiment topic as responsible for generating a sentiment word associated with each rating score. Consequently, sTOT predicts both the item and its rating score. We ran the experiments on PCs with Dual Core 2.66 GHz Xeon processors; the number of Gibbs sampling iterations was set to 1000.

B. Quantitative Evaluation

To assess the ability of our approach as a generative model to handle these review sets, we compared it against sTOT. We measure the ratio of the switch variable, evaluate the average word distribution separations between all pairs of different classes, and discuss the effect of the switch variable on the detection of sentiment words in a fully probabilistic and automated manner.

First, we measure the ratio of the switch variable to determine which kind of topic each document has in each token; the ratio of this variable informs us how many of the corresponding topics each document has. In particular, our main concern is 1)How does the ratio of this variable change with the number of topics? and 2)How much does the ratio differ between two models? We gauge their proportion and show measurements in Table III. From this table, we observe the following: (1)Although we increased the number of topics for each data set, we can not see any significant change due to this increment; about 80% of all tokens are occupied by topics in each document. This characteristic is common to both models and all data sets. (2)The ratio of sentiment topics is more insensitive to the number of sentiment classes in HASA than that in sTOT, where the increase in the number of sentiment topics does not significantly impact the ratio of sentiment topics in HASA. HASA differs from sTOT in capturing the relationships between aspect topics and sentiment topics at the document level. This indicates that sentiment analysis requires the fine-grained tightly coherent structure to describe the generative process.

Second, we measure the quality of topic models from the viewpoint of stochastic generative modeling. We can evaluate the generative ability of topic models by measuring the numerical distances between topics. The greater this distance is, the more the corresponding topics have specific and distinct word distributions. As these are probabilistic distributions over words, we can use KL-Divergence as the measure of distance. Since KL-divergence, $D_{kl}(P||Q)$, can measure the expected number of extra bits required to code samples from P when using a code based on Q, rather than using a code based on P, a natural measure of the evolution distance between two topics is given as follows.

$$D_{KL}(z_1||z_2) = \sum_{i=1}^{W} p(z_i|z_1) log \frac{p(z_i|z_1)}{p(z_i|z_2)}. \quad (4)$$

Both sTOT and HASA learn more various latent variables than other topic models, where each latent variable has a word distribution. Therefore, we also measure the difference between these word distributions. Table IV shows the results of the distance comparison. This table shows that HASA can more distinctly identify topics than sTOT, as the switching variable distinguishes words in each token and thus assigns words to the appropriate latent class.

Thirdly, we evaluate the ability of item identification given words in reviews, as HASA yields word distributions that can be applied to text classification. For each item m (class), we learn probabilistic model $p(\mathbf{w}|m)$ of the review in that item. An unseen review is classified by taking $argmax_m p(m|\mathbf{w}) = argmax_m p(\mathbf{w}|m)p(m) = argmax_m \prod_{i=1}^{n_d} p(w_{di}|m)p(m)$. This evaluation compares HASA with Naive Bayes (NB), and sTOT, by using the classification accuracy.

Similarly, we evaluate the rating prediction ability given the words/sentiment words in a review. This evaluation aims to determine which model more precisely infers the rating score from just the word distributions. We predict the rating of a given review by choosing the discretized rating that maximizes the posterior, which is calculated by multiplying all of rating probability of each word token from a topic-wise Beta distribution over rating $\prod_{i=1}^{n_d} p(v|\lambda_{z_{di}})$. As the baseline methods, we prepared Naive Bayes (NB), and sTOT, both model-based approaches, and then measured the difference between the predicted score and the correct rating score. To cover the two-stage approach, we calculated $p(w|v)$ and then predicted the score by using $argmax_v \prod_{i=1}^{n_d} p(v|w_{di})$.

As shown in Table VI, HASA achieves better accuracy than the others, and provides an average reduction in MAE relative error of 1.2%. These results also indicate that the manipulation of background topic and the hierarchical topic

TABLE III

THE RATIO OF SWITCH VARIABLE: IN THIS TABLE, r=0(r=1, r=2, r=3) DENOTES BACKGROUND (DOCUMENT, ASPECT TOPIC, SENTIMENT TOPIC), AND THE CORRESPONDING VALUE IS THE AVERAGE OF ITS COMPONENTS IN EACH REVIEW.

Product type	DVD				Book				Music			
Model	sTOT		HASA		sTOT		HASA		sTOT		HASA	
aspect topics (Z_a,G)	30	30	30	30	30	30	30	30	30	30	30	30
sentiment topics (Z_s,C)	15	30	15	30	15	30	15	30	15	30	15	30
$r = 0$	0.071	0.071	0.071	0.097	0.044	0.042	0.044	0.046	0.053	0.053	0.051	0.056
$r = 1$	0.012	0.012	0.014	0.015	0.015	0.015	0.013	0.012	0.013	0.014	0.015	0.015
$r = 2$	0.203	0.105	0.096	0.110	0.216	0.117	0.114	0.127	0.195	0.097	0.095	0.097
$r = 3$	0.714	0.812	0.817	0.777	0.725	0.826	0.827	0.815	0.734	0.836	0.839	0.832

TABLE IV

AVERAGE KL DIVERGENCE BETWEEN WORD DISTRIBUTIONS LEARNED FROM DATA: BOTH MODELS WERE LEARNED WITH THE NUMBER OF TOPICS Z_a (G) SET AT 30. RESULTS THAT DIFFER SIGNIFICANTLY, T-TEST $p < 0.01$, $p < 0.05$, FROM STOT ARE MARKED WITH '**', '*' RESPECTIVELY.

Product type	DVD				Book				Music			
	sTOT		HASA		sTOT		HASA		sTOT		HASA	
Z_s (C)	15	30	15	30	15	30	15	30	15	30	15	30
$D_{KL}(\hat{\phi}_{z_a(g)}\|\hat{\phi}_b)$	212.26	341.51	470.48**	568.32*	412.89	482.44	324.08**	388.96**	221.96	323.23	465.06**	632.04*
$D_{KL}(\hat{\phi}_{z_s(c)}\|\hat{\phi}_b)$	523.21	425.53	1829.00**	580.44*	282.83	303.25	3485.02**	694.38**	545.23	466.38	3236.16**	956.95*
$D_{KL}(\hat{\phi}_{z_a(g)}\|\hat{\phi}_{z_a(g)})$	9.25	10.88	13.65**	13.85**	10.34	11.65	12.85**	13.82**	10.13	11.03	13.57**	13.87**
$D_{KL}(\hat{\phi}_{z_s(c)}\|\hat{\phi}_{z_s(c)})$	10.27	11.01	16.24**	15.57**	11.22	11.73	16.75**	16.12**	11.12	11.77	16.52**	15.78**
$D_{KL}(\hat{\phi}_{z_a(g)}\|\hat{\phi}_{z_s(c)})$	19.76	20.24	30.94**	30.79**	19.22	20.73	31.15**	31.19**	20.12	20.73	31.18**	30.77**

TABLE V

EVALUATION OF REPRESENTATIVE CLASSIFICATION OF NB, STOT AND HASA UNDER DIFFERENT TOPIC NUMBERS: IN THESE EXPERIMENTS, THE NUMBER OF ASPECT TOPICS Z_a (G) IS SET TO 30. RESULTS THAT DIFFER SIGNIFICANTLY, T-TEST $p < 0.01$, $p < 0.05$, FROM THE OTHER MODELS ARE MARKED WITH '**', '*' RESPECTIVELY.

	Classification accuracy					
	DVD		Books		Music	
$Z_s(C)$	15	30	15	30	15	30
NB	0.62		0.48		0.51	
sTOT	0.69	0.73	0.51	0.55	0.58	0.64
HASA	0.72*	0.81*	0.52	0.55*	0.61**	0.68*

TABLE VI

MAE (RATING) COMPARISON OF NB, STOT AND HASA: MODEL-BASED APPROACHES (STOT AND HASA) WERE LEARNED WITH THE NUMBER OF TOPICS $Z_a(G)$ SET AT 30. RESULTS THAT DIFFER SIGNIFICANTLY, T-TEST $p < 0.01$, $p < 0.05$, FROM HASA ARE MARKED WITH '**', '*' RESPECTIVELY.

	Rating accuracy					
	DVD		Books		Music	
$Z_s(C)$	15	30	15	30	15	30
NB	0.317		2.98		3.21	
sTOT	0.241	0.228	0.231	0.212	0.236	0.225
HASA	0.222**	0.217*	0.211**	0.207	0.231*	0.217*

structure is essential for describing the generative process of review articles. Most background topic words are reused in almost all reviews regardless of content, referred item, and corresponding score. This result clearly confirms that HASA can realize automatic item and rating annotation and so offers aspect-sentiment-based item retrieval. We can not compare models with state-of-the-art rating prediction methods (e.g., collaborative filtering) due to the data sparseness of the Amazon review dataset; 97% of the reviewers evaluated only one item which prevents these methods from calculating the user-user/item-item relationships.

Finally, we numerically evaluated the models by computed test-set perplexity under the estimated parameters and compared the resulting values. Perplexity, which is a standard measure in the language modeling community to assess the predictive power of a model, is algebraically equivalent to the inverse of the geometric mean per-word likelihood (lower numbers are better). Perplexity was computed for all algorithms using 100 samples from 1000 different chains using

$$PPX = \exp(-\tfrac{1}{W}\textstyle\sum_{d\in D_{test}}^{|D_{test}|}\sum_{w\in d}^{|D|}\tfrac{1}{H}\log(\mu_0^h\phi_{bw}^h + \mu_1^h\phi_{dw}^h + \mu_2^h\sum_g^G\theta_{dg}^h\phi_{gw}^h + \mu_3^h\sum_g^G\theta_{dg}^h\sum_c^C\psi_{gc}^h\phi_{cw}^h)), \quad (5)$$

where W is the number of test words, H is the number of samples (from H different chains), θ_{dg}^h (ψ_{gc}^h) is the probability that g (c) will be assigned by a model to document d (g) in h and $\phi_{b(d,g,c)w}^h$ is the probability assigned by the model to word v conditioned on b (d, g, c)in h. A lower score implies that word w_d is less surprising to the model.

We computed the perplexity as follows. First, we randomly took 10% words from each article to create a test part; the remainder was used as the learning part. For every list, the test part was held out to compute perplexity. Second, the learning part was used for estimating the parameters by Gibbs sampling. Finally, a single set of topic counts was saved when a sample was taken; the log probability of test products that had never been seen before was computed in the same way as the perplexity computation. Table VII shows the perplexity comparison results; the averages of ten-fold cross validation.

This table shows that HASA offers lower perplexity than sTOT, and supports our idea that sentiments are best captured by capturing both words and the corresponding ratings, which reduces perplexity. Throughout these experiments, there was no significant difference between sTOT and HASA in terms of computation cost, as all models are inferred by using a

TABLE VIII
WORD DISTRIBUTION LEARNED FROM THE DATA SET BY HASA: WE LIST TOP 10 WORDS (BIGRAM WORDS) FROM TNG, STOT AND HASA WITH HIGH RATING. THE VALUES IN THE COLUMNS ARE THE CORRESPONDING PROBABILITIES.

DVD					
aspect word $\hat{\phi}_{gw}$	sentiment word $\hat{\phi}_{cw}$	aspect word $\hat{\phi}_{gw}$	sentiment word $\hat{\phi}_{cw}$	aspect word $\hat{\phi}_{gw}$	sentiment word $\hat{\phi}_{cw}$
wars,0.073	great, 0.083	version, 0.1476	movie,0.072	murray,0.0191	movie,0.072
star,0.052	like,0.083	extended,0.1379	film,0.062	lost,0.0299	film,0.062
lucas,0.016	good,072	dvd, 0.1289	great,0.054	bob,0.0135	great,0.054
jar,0.012	best,0.068	ring,0.1189	good,0.051	tokyo,0.0118	good,0.051
episode,0.012	feel,0.033	fellowship,0.1134	story,0.037	translation,0.0067	story,0.037
phantom,0.011	better,0.021	book,0.0913	think,0.026	charlotte,0.0053	think,0.026
menace,0.009	know,0.016	jackson,0.014	like,0.024	film,0.0042	like,0.024
effects,0.008	love,0.015	rings,0.018	movies,0.021	japanese,0.0029	movies,0.021
trilogy,0.005	really,0.011	peter,0.013	just,0.018	life,0.0017	just,0.018
dvd,0.004	just,0.006	lord,0.008	did,0.009	scarlett,0.0014	did,0.009
Books					
aspect word $\hat{\phi}_{gw}$	sentiment word $\hat{\phi}_{cw}$	aspect word $\hat{\phi}_{gw}$	sentiment word $\hat{\phi}_{cw}$	aspect word $\hat{\phi}_{gw}$	sentiment word $\hat{\phi}_{cw}$
harry, 0.165	best,0.057	amir,0.098	best,0.057	brown, 0.057	best,0.057
potter,0.077	like,0.021	afghanistan,0.073	like,0.021	da, 0,041	like,0.021
goblet,0.015	read,0.016	hassan,0.064	read,0.016	vinci, 0.040	read,0.016
series,0.014	novel,0.014	kite,0.056	novel,0.014	history,0.037	novel,0.014
rowling,0.005	good,0.012	novel,0.052	good,0.012	code,0.034	good,0.012
books,0.002	life,0.011	book,0.048	life,0.011	church,0.029	life,0.011
hogwarts,0.002	great,0.010	runner,0.033	great,0.010	dan,0.022	great,0.010
voldemort,0.001	really,0.009	father,0.025	really,0.009	grail,0.019	really,0.009
quidditch,0.001	better,0.008	afghan,0.021	better,0.008	historical,0.012	better,0.008
fourth,0.001	way,0.006	son,0.016	way,0.006	jesus,0.008	way,0.006
Music					
aspect word $\hat{\phi}_{gw}$	sentiment word $\hat{\phi}_{cw}$	aspect word $\hat{\phi}_{gw}$	sentiment word $\hat{\phi}_{cw}$	aspect word $\hat{\phi}_{gw}$	sentiment word $\hat{\phi}_{cw}$
madonna,0.134	like,0.113	day,0.154	like,0.033	u2,0.151	like,0.033
dance,0.126	great,0.107	green,0.129	music,0.027	album,0.086	music,0.027
hung,0.094	album,0.054	album,0.052	just,0.021	bono,0.084	just,0.021
madonnas,0.085	heard,0.047	punk,0.030	good,0.017	pop,0.073	good,0.017
new,0.076	pop,0.033	joe,0.016	band,0.021	vertigo,0.062	band,0.021
confessions,0.068	know,0.028	idiot,0.14	dont,0.019	way,0.059	dont,0.019
ray,0.032	better,0.029	suburbia,0.016	new,0.016	voice,0.043	new,0.016
light,0.028	fan,0.025	jesus,0.013	time,0.014	albums,0.037	time,0.014
love,0.023	bought,0.021	home,0.013	listen,0.013	atyclb,0.028	listen,0.013
york,0.015	long,0.015	dookie,0.012	want,0.012	achtung,0.022	want,0.012

TABLE VII
PERPLEXITY COMPARISON OF STOT AND HASA: BOTH MODELS WERE LEARNED WITH THE NUMBER OF TOPICS Z_a (G) SET AT 30. RESULTS THAT DIFFER SIGNIFICANTLY, T-TEST $p < 0.01$, $p < 0.05$, FROM HASA ARE MARKED WITH '**', '*' RESPECTIVELY.

	DVD		Books		Music	
Z_s (C)	15	30	15	30	15	30
sTOT	1195	1071	1192	1127	1312	1231
HASA	1173*	1051*	1144**	1082*	1275**	1216*

similar sampling approach. Consequently, HASA is useful as a generative model as shown in Table IV and VII, and to identify the item and the corresponding rating from a given review as shown in Table V and VI.

C. Qualitative Evaluation

Table VIII provides an example of the words of topics extracted by HASA. These three selected aspect topics have the highest probability in each category, where we selected the sentiment topic having a right skewed beta distribution in each aspect topic. In each topic, we list the top 10 words in decreasing order of topic-specific probability $\hat{\phi}_{gw}$ and $\hat{\phi}_{cw}$.

In the DVD category, the most popular aspect topic extracted by HASA is associated with "star wars", the next topic is associated with "The Lord of the Rings", and the last is associated with "Lost in Translation". Although all sentiment topics have a beta distribution for generating the highest rating in each aspect topic, the sentiment topic associated with "star wars" differs from that of 'The Lord of the Rings", and "Lost in Translation". Although we can find similar phenomena as well in the Music category, the sentiment topic is common for all three aspect topics in the Books category. The first aspect topic is associated with Madonna "Confessions On A Dance Floor" and its sentiment topic includes "pop". On the contrary, both the second aspect topic "green day" and the third aspect topic "U2" and their sentiment topics include the word "band". This table shows that sentiment topic words vary on the item and the category. Thus, HASA can provide us with sentiment words that are specific to both topics and the corresponding rating. Both benefits are useful for predicting the rating of a given review article. Because positive sentiment words are more various and interpretable than negative words in this data set, and space is limited, we show only the positive words due.

V. DISCUSSION

As shown in Table III, the background topic occupies about 6 % of each article over all data sets. This ratio is so high that topic models fail to detect meaningful words.

These background words are highly probable words in many various reviews and so do not have any strong relationship with the item/rating score of the article. Since these words differ depending on the corpus and data sets, it is difficult to manually create stop word lists that suit many data sets. HASA solves this problem by leveraging the switch variable and distinguishing these words for any given data set.

the models with Beta distribution show more distinct word distributions and thereby predict the rating more correctly than the other models.

VI. Conclusion

This paper showed how to simultaneously extract sentiment and aspect topic from review articles. We extended sTOT to incorporate rating scores instead of timestamps as the observed variable and enforcing the latent switching variable to concatenate topics. A novel feature of our approach is its inclusion of sentiment topic, aspect topic, observed item and its rating variable. One of the limitations of HASA is that it represents each document as a bag of words and thus ignores the word ordering and phrases. Future work is to extend HASA to handle aspect and sentiment phrases by detecting N-gram words.

References

[1] S. Brody and N. Elhadad. An unsupervised aspect-sentiment model for online reviews. In *HLT-NAACL*, pages 804–812, 2010.

[2] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. In *NIPS*, pages 475–482, 2005.

[3] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD*, pages 168–177, 2004.

[4] N. Jindal and B. Liu. Opinion spam and analysis. In *WSDM*, pages 219–230, 2008.

[5] N. Kawamae. Predicting future reviews: Sentiment analysis models for collaborative filtering. In *WSDM*, pages 605–614, 2011.

[6] N. Kawamae. Trend analysis model: Trend consists of temporal words, topics, and timestamps. In *WSDM*, pages 317–326, 2011.

[7] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw. Detecting product review spammers using rating behaviors. In *CIKM*, pages 939–948, 2010.

[8] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *CIKM*, pages 375–384, 2009.

[9] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *WWW*, pages 342–351, 2005.

[10] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *WWW*, pages 131–140, 2009.

[11] A. McCallum, A. Corrada-Emanuel, and X. Wang. Topic and role discovery in social networks. In *IJCAI*, pages 786–791, 2005.

[12] T. Minka and J. M. Winn. Gates. In *NIPS*, pages 1073–1080, 2008.

[13] S. Moghaddam and M. Ester. Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews. In *SIGIR*, pages 665–674, 2011.

[14] I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *ACL*, pages 308–316, 2008.

[15] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis on review text data: a rating regression approach. In *KDD*, pages 783–792, 2010.

[16] P. Wei and S. Xiaotong. Penalized model-based clustering with application to variable selection. *J. Mach. Learn. Res*, 8:1145–1164, 2007.

[17] D. Widyantoro, T. Ioerger, and J. Yen. Learning user interest dynamics with a three-descriptor representation. *Journal of the American Society for Information Science and Technology*, 52(3):212–225, 2001.

[18] J. Wiebe. Learning subjective adjectives from corpora. In *AAAI*, pages 735–740, 2000.

[19] T.-L. Wong, W. Lam, and T.-S. Wong. An unsupervised framework for extracting and normalizing product attributes from multiple web sites. In *SIGIR*, pages 35–42, 2008.

[20] J. Yu, Z.-J. Zha, M. Wang, and T.-S. Chua. Aspect ranking: identifying important product aspects from online consumer reviews. In *ACL*, pages 1496–1505, 2011.

[21] W. X. Zhao, J. Jiang, H. Yan, and X. Li. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *EMNLP*, pages 56–65, 2010.