Identifying Objective and Subjective Words via Topic Modeling

Hanqi Wang, Fei Wu, Weiming Lu, Yi Yang, Xi Li, Xuelong Li, *Fellow, IEEE*, and Yueting Zhuang

***Abstract*— It is observed that distinct words in a given document have either strong or weak ability in delivering facts (i.e., the *objective* sense) or expressing opinions (i.e., the *subjective* sense) depending on the topics they associate with. Motivated by the intuitive assumption that different words have varying degree of *discriminative* power in delivering the objective sense or the subjective sense with respect to their assigned topics, a model named as *i*dentified *o*bjective–*s*ubjective latent Dirichlet allocation (LDA) (*i*osLDA) is proposed in this paper. In the *i*osLDA model, the simple Pólya urn model adopted in traditional topic models is modified by incorporating it with a probabilistic generative process, in which the novel "*Bag-of-Discriminative-Words*" (BoDW) representation for the documents is obtained; each document has two different BoDW representations with regard to objective and subjective senses, respectively, which are employed in the joint objective and subjective classification instead of the traditional Bag-of-Topics representation. The experiments reported on documents and images demonstrate that: 1) the BoDW representation is more predictive than the traditional ones; 2) *i*osLDA boosts the performance of topic modeling via the joint discovery of latent topics and the different objective and subjective power hidden in every word; and 3) *i*osLDA has lower computational complexity than supervised LDA, especially under an increasing number of topics.**

***Index Terms*— Latent Dirichlet allocation (LDA), latent variable model, supervised learning, topic modeling.**

I. Introduction

THERE is a growing demand of automatical analysis on the multimodal data (e.g., electronic documents, images, audio and video data, and so on) that can be easily found and obtained from the Internet. So far, various machine learning algorithms have been employed in accessing, retrieving, clustering, and summarizing the data. Among them, topic models [1] are more and more popular due to their ability to efficiently discover the latent structure embedded over a group of documents and provide low-dimensional representation for large-scale data. The earliest topic model is probabilistic latent semantic analysis (pLSA) [2] that evolves from LSA [3]. As a *latent variable model* [4], it is the first to capture the hidden semantics (i.e., the *topics*) conveyed by different words during the modeling of documents. In pLSA, documents are projected into a low-dimensional topic space by assigning each word with a latent topic, where each topic is usually represented as a multinomial distribution over a fixed vocabulary. While various extensions of pLSA have been proposed in recent years [5]–[7], the most famous and successful one among them remains to be latent Dirichlet allocation (LDA) [8]. The LDA model inherits the notion of pLSA, but it employs a generative process on the topic proportion of each document and models the whole corpus via a hierarchical Bayesian framework [9]. In fact, pLSA turns out to be a special case of LDA with a uniform Dirichlet prior in a maximum $a\ posteriori$ model [10], while LDA has a better ability of modeling large-scale documents for its well-defined $a\ priori$. In the past decade, topic models, especially the LDA model, have been intensively studied [11]–[13] and widely applied for many different tasks [14]–[18].

As an unsupervised model, the original LDA model is built based on the "Bag-of-Words" (BoW) representation, where the documents are treated as *unordered* collections of words, disregarding any linguistic structures embedded in them. The BoW representation and the LDA framework have also been applied for image clustering after the low-level visual features of given images are extracted as the *visual words*. In spite of the convenience in modeling and computation, this traditional approach brings about, however, the latent representation learned by LDA has been criticized for several deficiencies [19], and it is often found not to be so strongly predictive [20]. As a matter of fact, the unsupervised manner employed in LDA unfortunately loses sight of the nature of various discriminative tasks, such as classification and regression, and thus provides no guarantee on the effectiveness of the learned representation. On the other side, it is often easy to obtain some useful auxiliary information [21] (e.g., the category labels or the ratings provided by the authors) along with the input documents in many practical applications. Therefore, much effort has been devoted to leveraging such auxiliary information and developing supervised extensions of the traditional LDA model in order to generate latent representation that is more predictive for the discriminative tasks [4]. In supervised topic models, such as supervised LDA (sLDA) [22], multiclass sLDA [23], and τLDA [24], each label attached to its corresponding document is modeled as the response variable predicted based on the latent representation

Manuscript received November 24, 2015; revised May 3, 2016; accepted October 26, 2016. This work was supported in part by the National Basic Research Program of China under Grant 2015CB352300, in part by NSFC under Grant U1509206, Grant 61572431, and Grant 61472353, in part by the China Knowledge Centre for Engineering Science and Technology, and in part by the Fundamental Research Funds for the Central Universities. *(Corresponding Author: F. Wu).*

H. Wang, F. Wu, W. Lu, Xi Li, and Y. Zhuang are with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China (e-mail: hanqiw@zju.edu.cn; wufei@zju.edu.cn; luwm@zju.edu.cn; xilizju@zju.edu.cn; yzhuang@zju.edu.cn).

Y. Yang is with the Center for Quantum Computation and Intelligent Systems, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: yi.yang@uts.edu.au).

Xuelong Li is with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an 710119, China (e-mail: xuelong_li@opt.ac.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2016.2626379

of the document that generated during the process of topic modeling.

So far, most supervised extensions of LDA utilize the Bag-of-Topics (BoT) representation of one document for the prediction of its corresponding label, in which the proportion of topics (instead of the word proportion in BoW) in the document is considered to be the predictive feature. That is, any two of the various words in the vocabulary are equal once they are assigned with the same topic. However, it is intuitive that distinct words in a given document have either strong or weak ability in delivering facts (i.e., the *objective* sense) or expressing opinions (i.e., the *subjective* sense) depending on their assigned topics [25], [26], which endows them with varying degree of discriminative power in terms of objective and subjective senses.

Three examples are presented here to illustrate this assumption.

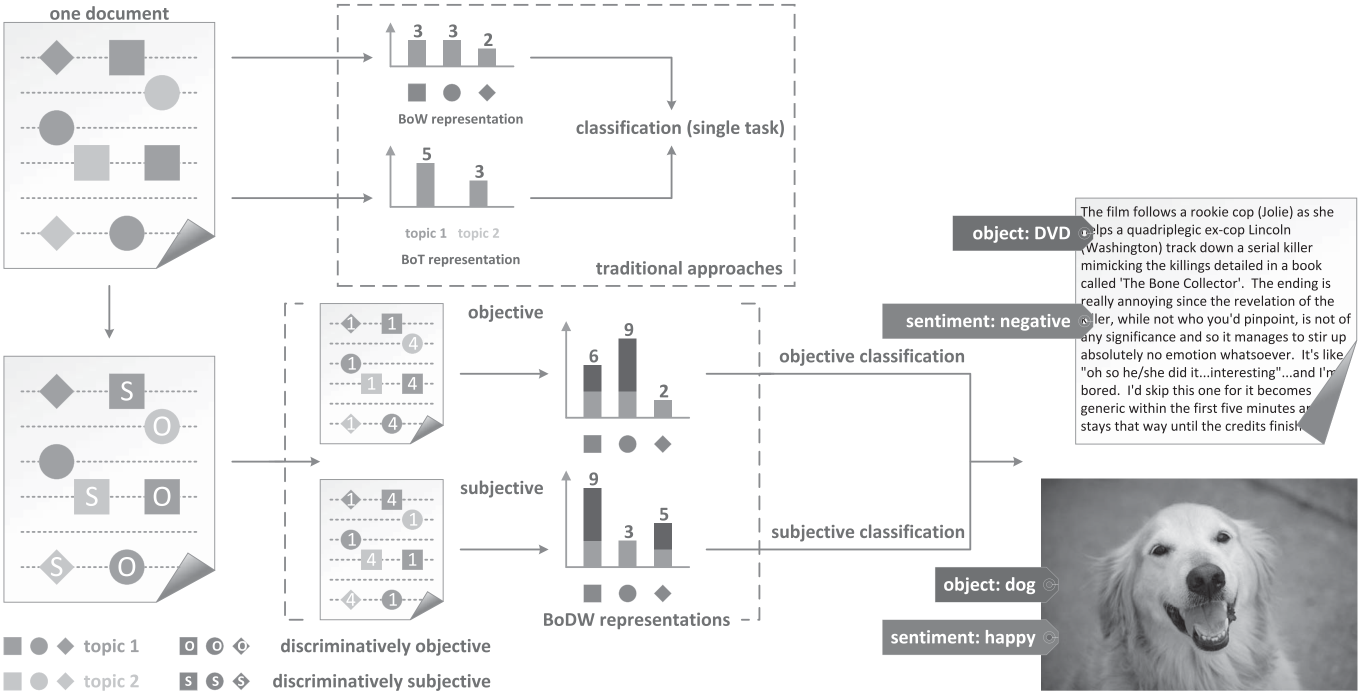
1) Given an article from the newspaper, the word "plant" in the topic "*nuclear crisis*" highly tends to indicate one particular object (a nuclear plant), and is, therefore, strongly predictive in the objective (category) classification of the article. On the contrary, when this word is assigned with the topic "*landscape*," it probably refers to the whole scene and is, therefore, weak in the objective sense.
2) Compare the aforementioned word "plant" with another word "reactor" after assigning both of them with the topic "*nuclear crisis*." They are both strongly discriminative in the objective sense, while the word "reactor" is more powerful than the word "plant," since the latter is also likely to remark plants growing near the nuclear reactors.
3) Given one word "bug" from a bunch of documents, it apparently remarks one object (one kind of insects) when assigned with the topic "*order Hemiptera*," while the same word under the topic "*software*" probably conveys a negative opinion in sentimental identification.

Thus, it becomes imperative to first deliberately characterize the different objectively or subjectively discriminative power of the words in the documents with respect to their involved topics, and then benefit from such identification in constructing a more predictive representation of each document. As a result, a supervised approach named as *i*dentified *o*bjective–*s*ubjective LDA (*i*osLDA) is proposed in this paper that extends the basic framework of multiclass sLDA in many aspects. In the *i*osLDA model, the simple Pólya urn (SPU) model followed by traditional topic models is modified by incorporating it with a probabilistic generative process to obtain the novel "*Bag-of-Discriminative-Words*" (BoDW) representation for the documents. Each document has two BoDW representations with regard to objective and subjective senses, respectively, which are then employed in the joint objective and subjective classification. The BoDW representation and the whole procedure of the *i*osLDA model are shown in Fig. 1 as well as the traditional methods. The *i*osLDA possesses the attractive ability of naturally tapping into the different powers of various words in delivering either an objective or a subjective sense in one given document, while it jointly imposes the auxiliary information in terms of both objective and subjective senses to boost the performance of latent representation (i.e., topic modeling). Results of several experiments demonstrate that BoDW is more predictive for discriminative tasks than the traditional BoW and BoT representation employed in the current methods.

II. Related Work

The sLDA [22] model is a natural supervised extension of the traditional LDA model. Inheriting the hierarchical Bayesian structure that adopted in traditional LDA, sLDA is capable to properly handle labeled documents by adding to the model a response variable associated with each document. As mentioned before, sLDA jointly models the documents and the responses, and then, the responses are predicted by the latent topics discovered in their corresponding documents (i.e., BoT). The sLDA is initially proposed for documents with unconstrained real-valued labels, where the response value is produced from a normal linear model. However, sLDA theoretically accommodates various types of response (e.g., real or discrete values, nonnegative values, multiclass labels, and so on) when cooperated by a *generalized linear model* [27], which makes it easily extended for many kinds of discriminative tasks. The *multiclass sLDA* model is implemented in [23] for analyzing images in different categories. To simultaneously model the visual words in images and the textual words annotated for each image while performing classification, the authors further proposed *multiclass sLDA with annotation* that combines *corresponding LDA* [14] and softmax regression in a joint framework. In such an approach, both the visual and textual words are latent variables while some of them share the same topic, based on which the aforementioned BoT representation is generated for prediction. As another famous variant of sLDA, τLDA [24] aims to stride across the language gap between documents with different technicalities (e.g., a news report and its related journal article). In the τLDA model, each word is assigned with a binary selector to determine whether it is a technical word. All the assigned selectors in one document form the latent representation, based on which the document technicality is predicted via a cosine regression model.

While the traditional sLDA model (including its multiclass variant) is capable for almost any kinds of discriminative tasks, such as classification and regression, it lacks the ability to perform both objective and subjective identification of given data at the same time. Several approaches, therefore, have been proposed to discover both topics (i.e., the objective sense) and sentiments (i.e., the subjective sense) in a collaborative manner. For instance, Mei *et al.* [25] propose an approach to model the mixture of topics and sentiments in weblogs, named as topic-sentiment mixture. Multigrain LDA [28] is built later that aims to extract and aggregate specific sentimental words related to different topics. The joint sentiment-topic (JST) model and its reparameterized version (reversed JST) [29] are designed to explicitly identify the sentimental polarities expressed by words in documents, and is capable for mining the content of different sentiments in terms of one given topic.



The film follows a rookie cop (Jolie) as she elps a quadriplegic ex-cop Lincoln Washington) track down a serial killer mimicking the killings detailed in a book called 'The Bone Collector'. The ending is really annoying since the revelation of the ller, while not who you'd pinpoint, is not of any significance and so it manages to stir up absolutely no emotion whatsoever. It's like "oh so he/she did it...interesting"...and I'm bored. I'd skip this one for it becomes generic within the first five minutes an stays that way until the credits finish

Fig. 1. Intuitive illustration of three document representations, namely, the BoW model, the BoT model, and the BoDW model proposed in this paper. Different shapes indicate distinct words, while different colors indicate different topics. The shapes marked with letter "O" indicate that these words have the discriminative power to deliver the objective sense, and those with letter "S" tend to convey the subjective sense. For simplicity, there are in total three words and two topics in this document, while the frequency of the words with discriminative power in the objective or the subjective sense is multiplied by four in the corresponding BoDW representation as a toy example. The document representation in terms of BoDW is particularly appropriate for discriminant analysis (e.g., document classification or sentiment identification) owing to its nature of disentangling the discriminative words with respect to their topics. In contrast, in the BoW and BoT representation, all of the words are equally employed to the classification tasks no matter how discriminative or trivial these words are with respect to their assigned topics (best viewed in color).

All these aforementioned methods conduct an unsupervised manner in discovering the latent sentiments as well as the topics, and represent the documents via BoT or its equivalent (i.e., the representation constituted by the proportion of latent variables). Thus, they are not as predictive as the traditional sLDA model in the discriminative tasks.

The Pólya urn model [30] is a type of statistical model that treats objects under analysis as colored balls and their groups or containers as urns. In the perspective of topic modeling, an individual word in the given document can be treated as a ball of a certain color indicating its uniqueness in the vocabulary, while each topic is seen as an urn. The word distribution in terms of different topics is the equivalent to the color proportion of balls in the corresponding urns. The original LDA model adopts the SPU model that when a ball in one certain color (i.e., one word) is drawn from an urn (i.e., its assigned topic), it is observed and then put back into the urn with an additional ball of the same color (i.e., the same word). The SPU model endows traditional LDA with a self-reinforcing property known as "the rich get richer." Instead of the SPU model, the generalized Pólya urn (GPU) model [31] is later introduced into topic modeling in order to incorporate the corpus-specific word co-occurrence information. The GPU model goes a step further that after a ball with certain color (i.e., one word) in an urn (i.e., its assigned topic) is drawn; not only two balls with the same color, but also some balls with other colors (i.e., some related words) are put back to the urn together. The interaction of balls with different colors can be configured by predefined rules or *a priori* knowledge [32]. In this paper, the employed Pólya urn model is modified based on the SPU model that the number of balls to put back after being observed is drawn by a certain kind of probabilistic distribution, which is a brand new attempt to the best of our knowledge.

Other than topic modeling, there are several kinds of approaches reported to have the ability in discovering and identifying the semantic information hidden in enormous numbers of documents and images. For instance, nonnegative matrix factorization (NMF) is another popular dimension-reduction method that widely applied to image processing and pattern recognition for its nonnegative constraints that allow only additive combinations and lead to naturally sparse and parts-based representations, while it deals with large-scale data sets in high efficiency with an online algorithm employed [33]. NMF essentially gives similar results of reduction as LDA, except for the specific assignment of topics to each word, while the results are not necessarily normalized [34]. Besides, some of the works in neural network also demonstrate their effectiveness in topic modeling [35]–[37]. As a matter of fact, such neural network models are either directed or undirected *latent variable models* as well as the Bayesian models, such as LDA, and the hidden layer in their model is commonly treated as the "topics." While the neural models find out the most representative words in each topic according to the largest weights connected to the hidden layer, they lack the ability to discover the significance of

TABLE I
NOTATIONS USED IN THE *i*osLDA MODEL

Notation	Description
D	the number of documents
N_d	the number of words in the d^{th} document
K	the number of the topics hidden in the corpus
V	the size of the vocabulary
O	the number of distinct objective labels
S	the number of distinct subjective labels
$w_{d,i}$	the i^{th} word in the d^{th} document
$z_{d,i}$	the topic assignment of $w_{d,i}$
y_d^O	the objective label of the d^{th} document
y_d^S	the subjective label of the d^{th} document
θ_d	the topic proportion specific to the d^{th} document
ϕ_k	the word proportion specific to the k^{th} topic
$\lambda_{k,v}^O$	the probability of different values that may taken by an objective impact scaler, whose corresponding word is v under the k^{th} topic
$\lambda_{k,v}^S$	the probability of different values that may taken by an subjective impact scaler, whose corresponding word is v under the k^{th} topic
α	the Dirichlet a priori of all the topic proportion $\boldsymbol{\theta}$
β	the Dirichlet a priori of all the word proportion $\boldsymbol{\phi}$
γ^O	the Dirichlet a priori of all the proportion $\boldsymbol{\lambda}^O$
γ^S	the Dirichlet a priori of all the proportion $\boldsymbol{\lambda}^S$
η^O	parameters of softmax regression
η^S	parameters of softmax regression

the topics in different documents, which may suggest that they are incomplete topic models. Finally, compared with all the aforementioned approaches, LDA and its variants can easily incorporate various kinds of prior knowledge in their model due to the Bayesian techniques they adopt, and avoid several kinds of drawbacks, such as the dependence on initialization, overfitting, and noise-level underestimation problems [38].

III. *i*osLDA MODEL

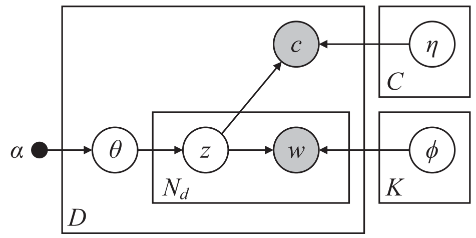
A. Formulation

Table I gives out the notations used in this paper. Assume that there are in total D documents in the data set. The dth document has N_d words, while all the distinct words in the whole data set form a vocabulary of size V. Each document has exactly one objective label (e.g., the mainly mentioned book or movie) and one subjective label (e.g., the delivered positive or negative sentiment, or some much more fine-grained categories). The whole data set, therefore, contains O different objective labels and S distinct subjective ones, respectively. In addition, we assume that the number of total topics hidden in the data set (i.e., K) is *a priori* specified and fixed, while it can be determined during model selection [39] in practice. As in the traditional LDA and sLDA model, topic proportion θ in terms of each document is repeatedly drawn from its K-dimensional Dirichlet *a priori* with parameter α, while the topic assignment z of each word is conditioned on θ. The word proportion ϕ specific to the topics is smoothed by endowing it with a symmetric *a priori*, i.e., the V-dimensional Dirichlet distribution with parameter β, and every word is then sampled according to ϕ_z, where z is its topic assignment.

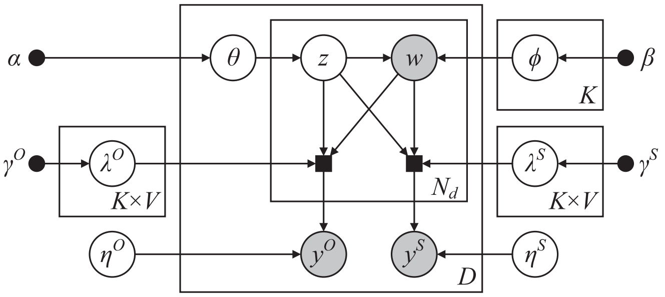
After all the topic assignments z are determined, traditional supervised models [22], [23] directly utilize them to form the BoT representation of the documents, and learn the parameters of the linear or softmax regression model under the assumption that the label of each document is drawn conditioned on the BoT representation. In contrast, *i*osLDA assumes that different words have their intrinsic different powers in delivering the objective or the subjective sense with respect to their assigned topics. Therefore, *i*osLDA modifies the SPU model, which is the original sampling process of words given their topic assignments: when a ball with one certain color (i.e., one word) is drawn from its urn (i.e., the assigned topic), the number of balls with the identical color to put back is determined conditioned on a probabilistic distribution specific to the urn and the color of the ball itself. That is, different words will have various *impact scalers* with respect to their assigned topics, that one word may be treated as two or three similar words, while another word is ignored if its corresponding scaler is zero. It is worth noting that the probabilistic distribution in the modified model can be any discrete one: the Poisson distribution, the multinomial distribution, or even a fixed nonnegative integer (employed as a toy example in Fig. 1). In the *i*osLDA model, the impact scalers in terms of objective and subjective senses both range from $[0, L]$, and are conditioned on the multinomial distribution with parameters λ^O and λ^S, respectively. Each of the two distribution has Dirichlet *a priori*, namely, γ^O and γ^S. Each word is multiplied by its objective impact scaler in the calculation of word frequency to constitute the BoDW representation of the given documents in terms of objective identification, and so do the subjective impact scalers work with their corresponding words.

Denoting Dirichlet and mutinomial distribution as "Dir" and "Multi," respectively, the whole generative process of the *i*osLDA model can be described as follows.

1) For each topic k, draw word proportion $\phi_k \sim \mathrm{Dir}(\beta)$.
2) For each topic k and word w, the following holds.
   a) Draw possibility $\lambda_{k,w}^O \sim \mathrm{Multi}(\gamma^O)$ for values of the impact scalers in terms of objective sense.
   b) Draw possibility $\lambda_{k,w}^S \sim \mathrm{Multi}(\gamma^S)$ for values of the impact scalers in terms of subjective sense.
3) For each document d, the following holds.
   a) Draw topic proportion $\theta_d \sim \mathrm{Dir}(\alpha)$.
   b) Sample each topic assignment $z_{d,i} \sim \mathrm{Mult}(\theta_d)$.
   c) Sample each word $w_{d,i} \sim \mathrm{Mult}(\phi_{z_{d,i}})$.
   d) Sample each objective impact scaler
   $$x_{d,i}^O \sim \mathrm{Mult}\big(\lambda_{z_{d,i},w_{d,i}}^O\big).$$
   e) Sample each subjective impact scaler
   $$x_{d,i}^S \sim \mathrm{Mult}\big(\lambda_{z_{d,i},w_{d,i}}^S\big).$$
   f) Draw the objective label y_d^O that
   $$p\big(y_d^O = o|\boldsymbol{w}_d, \boldsymbol{x}_d^O, \eta^O\big) = \frac{\exp\big(\eta_o^{O^T}\overline{w}_d^O\big)}{\sum_{i=1}^{O}\exp\big(\eta_i^{O^{\mathrm{T}}}\overline{w}_d^O\big)}$$
   where
   $$\overline{w}_d^O = \frac{1}{N_d}\sum_{i=1}^{N_d} w_{d,i}x_{d,i}^O.$$



(a)



(b)

Fig. 2. Graphical model representation of (a) multiclass sLDA and (b) *i*osLDA proposed in this paper. The central plates from outer to inner, respectively, represent the documents and the words in each document. The black round dots are the hyper parameters. The shaded nodes indicate observations, while the others represent the latent variables in the model. Besides, the black rectangles in (b) represent the modified Pólya urn model, in which the power (i.e., the frequency) of different words will be magnified or reduced by their specific impact scalers.

g) Draw the subjective label y_d^S that

$$p\left(y_d^S = s|\boldsymbol{w}_d, \boldsymbol{x}_d^S, \eta^S\right) = \frac{\exp\left(\eta_s^{S\mathrm{T}}\overline{w}_d^S\right)}{\sum_{i=1}^{S}\exp\left(\eta_i^{S\mathrm{T}}\overline{w}_d^S\right)}$$

where

$$\overline{w}_d^S = \frac{1}{N_d}\sum_{i=1}^{N_d} w_{d,i}x_{d,i}^S.$$

The graphical model representation of the proposed *i*osLDA is shown in Fig. 2(a), which is compared with multiclass sLDA in Fig. 2(b). Both of the two model are illustrated with the notations in Table I, except that the single label (either objective or subjective) attached to each document is modeled as a distinct value c in multiclass sLDA, $c \in [1, C]$, and multiclass sLDA predicts the labels based on a softmax regression of parameter η.

B. Training Process

An EM strategy is conducted to train the *i*osLDA model, which consists of *posterior inference* and *parameter estimation*. The posterior inference aims to obtain the conditional distribution of the latent variables $\{\boldsymbol{\theta}, \boldsymbol{\phi}, \boldsymbol{\lambda}^O, \boldsymbol{\lambda}^S, z, \boldsymbol{x}^O, \boldsymbol{x}^S\}$ given the observations $\{\boldsymbol{w}, \boldsymbol{y}^O, \boldsymbol{y}^S\}$. This is an intractable problem for most of topic models [8]; therefore, some approximate methods are often taken as substitutes, among which variational inference [8], Gibbs sampling [39], and expectation propagation [40] are the popular ones. As in the *i*osLDA model, the collapsed Gibbs sampling is conducted in the E-step of the training process, where all the proportion $\boldsymbol{\theta}$, $\boldsymbol{\phi}$, $\boldsymbol{\lambda}^O$, and $\boldsymbol{\lambda}^S$ are integrated out first and the assignments $\boldsymbol{z}$, $\boldsymbol{x}^O$, and $\boldsymbol{x}^S$ are then iteratively sampled. Denoting $\{\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{x}^O, \boldsymbol{x}^S, \boldsymbol{y}^O, \boldsymbol{y}^S\}$ as $\boldsymbol{\Phi}$ and $\{\alpha, \beta, \gamma^O, \gamma^S, \eta^O, \eta^S\}$ as Ψ, the rules for updating the assignments in the Markov chain can be derived as follows:

$$p(z_{d,i} = k|\boldsymbol{\Phi}_{-z_{d,i}}, \Psi) \propto (\alpha_k + n_{d,k})\frac{\beta_{w_{d,i}} + n_{k,w_{d,i}}}{\sum_{v=1}^{V}(\beta_v + n_{k,v})} \tag{1}$$

$$p\left(x_{d,i}^O = l|\boldsymbol{\Phi}_{-x_{d,i}^O}, \Psi\right) \propto \frac{\gamma_l^O + n_{k,w_{d,i},l}^O}{\sum_{j=0}^{L}\gamma_j^O + n_{k,w_{d,i}}} \times \frac{\exp\left(\sum_{i=1}^{N_d}\eta_{y_d^O,w_{d,i}}^O w_{d,i}x_{d,i}^O\right)}{\sum_{o=1}^{O}\exp\left(\sum_{i=1}^{N_d}\eta_{o,w_{d,i}}^O w_{d,i}x_{d,i}^O\right)} \tag{2}$$

$$p\left(x_{d,i}^S = l|\boldsymbol{\Phi}_{-x_{d,i}^S}, \Psi\right) \propto \frac{\gamma_l^S + n_{k,w_{d,i},l}^S}{\sum_{j=0}^{L}\gamma_j^S + n_{k,w_{d,i}}} \times \frac{\exp\left(\sum_{i=1}^{N_d}\eta_{y_d^S,w_{d,i}}^S w_{d,i}x_{d,i}^S\right)}{\sum_{s=1}^{S}\exp\left(\sum_{i=1}^{N_d}\eta_{s,w_{d,i}}^S w_{d,i}x_{d,i}^S\right)}. \tag{3}$$

Here, $n_{d,k}$ denotes the number of words in the dth document assigned with the kth topic, while $n_{k,v}$ denotes the times that distinct word v occurring in the kth topic. $n_{k,v,l}^O$ and $n_{k,v,l}^S$ represent the number of the word v assigned with the kth topic and impact selectors of value l in terms of objective and subjective senses, respectively, where $l \in [0, L]$. $\Gamma(\cdot)$ is the Gamma function that $\Gamma(x) = \int_0^{+\infty} t^{x-1}e^{-t}\mathrm{d}t$.

The full derivations of the sampling rules above are presented in the Appendix. It is worth noting that though the generative process of *i*osLDA is much more sophisticated than the traditional models, the topics are sampled as simply as those in the original LDA model (and more conveniently than sLDA), which endows it with high efficiency in calculation.

In the M-step of the training process, parameters η^O and η^S are estimated via minimizing their corresponding softmax cost functions. Denoting $1(\cdot)$ as the indicator function that $1(\text{true statement}) = 1$, $1(\text{false statement}) = 0$, the $\{c, v\}$ elements of their derivative matrices are taken as

$$-\frac{1}{D}\sum_{d=1}^{D}\frac{\sum_{l=0}^{L} n_{d,v,l}^O}{N_d} \times \left[1\left(y_d^O = l\right) - \frac{\exp\left(\sum_{i=1}^{N_d}\eta_{l,w_{d,i}}^O w_{d,i}x_{d,i}^O\right)}{\sum_{o=1}^{O}\exp\left(\sum_{i=1}^{N_d}\eta_{o,w_{d,i}}^O w_{d,i}x_{d,i}^O\right)}\right] \tag{4}$$

and

$$-\frac{1}{D}\sum_{d=1}^{D}\frac{\sum_{l=0}^{L} n_{d,v,l}^S}{N_d} \times \left[1\left(y_d^S = l\right) - \frac{\exp\left(\sum_{i=1}^{N_d}\eta_{l,w_{d,i}}^S w_{d,i}x_{d,i}^S\right)}{\sum_{s=1}^{S}\exp\left(\sum_{i=1}^{N_d}\eta_{s,w_{d,i}}^S w_{d,i}x_{d,i}^S\right)}\right] \tag{5}$$

respectively, a gradient descent process can efficiently and effectively accomplish the parameter estimation.

C. Prediction

After properly trained, the proposed *i*osLDA is able to predict both the objective and subjective labels of documents that are unseen before. The prediction process is equal to calculating the expectation of response $\{\boldsymbol{y}^O, \boldsymbol{y}^S\}$ given $\{\boldsymbol{\phi}, \boldsymbol{\lambda}^O, \boldsymbol{\lambda}^S\}$. In practice, we first run collapsed Gibbs sampling on the unseen documents, and then take

$$E\left[y_d^O\right] = \max_o p\left(y_d^O = o \middle| \boldsymbol{\Phi}_{-y_d^O}, \Psi, \boldsymbol{\phi}, \boldsymbol{\lambda}^O\right) \tag{6}$$

and

$$E\left[y_d^S\right] = \max_s p\left(y_d^S = s \middle| \boldsymbol{\Phi}_{-y_d^S}, \Psi, \boldsymbol{\phi}, \boldsymbol{\lambda}^S\right) \tag{7}$$

for the dth document. Here, the probability of $\boldsymbol{y}^O$ and $\boldsymbol{y}^S$ is as the same as in the generative process.

IV. Experiments

A. Data Sets

Data sets of two modalities (i.e., textual and visual data) are conducted in our experimental comparisons.

1) *Document Data Set:* The Multidomain Sentiment Data Set[1] is employed, which consists of a large number of reviews about products (objective sense) as well as their sentimental ratings (subjective sense) from Amazon.com. English stop words and words occurring fewer than ten times are removed during the preprocess, which generates a vocabulary that is more effective.
2) *Image Data Sets:* Two data sets, namely, the Flickr Data Set and the Twitter Data Set, are utilized for evaluation. They are both proposed in [41]. The images in Flickr Data Set are labeled by *adjective–noun pairs* (ANP) (e.g., "sad man" or "happy family"), where the adjective words are treated as the subjective descriptions of the images and the noun words indicate their objective senses. To make the data set more balanced, we use the queries belonging to 24 ANPs to retrieve images from the Internet. As for the Twitter Data Set, the hashtags of images are taken as objective labels and those images have been manually labeled as one of the sentimental senses (i.e., positive, negative, or neutral). On both of these data sets, the SIFT descriptors are extracted as local features on each image, and then, respectively, obtained a collection of 1000 visual words as the codebook.

More detailed description of all the data sets can be found in Table II. In the experiments, half of the documents and images are chosen as the training data, while others are employed for testing.

TABLE II

Statistics of Data Sets

Multi-Domain Sentiment Dataset	
documents	8,000
words	2,625,094
vocabulary size	32,347
objective labels	4: books, dvd, electronics, kitchen
subjective labels	2: positive, negative
Flickr Dataset	
images	1,323
visual words	1,177,027
codebook size	1,000
adjective-noun pairs	24: angry dog, angry man, bright moon, broken glass, crying baby, disgusted man, fat girl, fat cat, fearful man, happy baby, happy family, joyful dog, joyful man, misty forest, misty lake, sad dog, sad man, scary tree, sparkling water, surprised man, wet cat, wet dog, scary monster, wet grass
objective labels	14
subjective labels	15
Twitter Dataset	
images	595
visual words	855,268
codebook size	1,000
objective labels	22: decemberwish, election, sandy, cancer, blackfriday, religion, android, aids, nfl, abortion, police, obama, globalwarming, gaymarriage, championsleague, cairo, agt, applefanboy, memoriesiwontforget, hurricanesandy, newyork, zimmerman
subjective labels	3: positive, negative, neutral

B. Comparative Approaches

Seven state-of-the-art methods, which are all discriminative extensions of topic models, are involved in the experiments. They are compared with three *i*osLDA-based models, two of which are the variants of the complete *i*osLDA that proposed in this paper.

1) *BoW + Support Vector Machine (SVM):* This is one of the traditional approaches to classify the documents, in which the SVM is learned for classification based on the BoW representation of the given documents.
2) *BoW + LR:* Another traditional method for discriminative tasks that similar to BoW+SVM, while logistic regression is employed here instead of SVM.
3) *LDA + SVM:* LDA is originally an unsupervised model, while the BoT representation it generates on given data can be imposed by discriminative methods as the features for prediction. LDA+SVM is a two-step approach that LDA obtains BoT representation on both training and test sets, and an SVM model is then learned and utilized for classification.
4) *LDA + LR:* The framework of LDA+LR is exactly the same to LDA+SVM, except that logistic regression is employed here instead of the SVM model.
5) *BoW + LDA + SVM:* In this approach, both the traditional BoW representation and the BoT representation generated by the original LDA model are employed, which are combined as the feature of the given data. Based on them, an SVM model is learned for classification.
6) *BoW + LDA + LR:* The framework of this method is exactly the same to BoW+LDA+SVM, except that logistic regression is employed here instead of the SVM model.
7) *sLDA:* The multiclass version of sLDA [23] is employed, which is capable to jointly perform topic modeling and

[1] http://www.cs.jhu.edu/∼mdredze/datasets/sentiment/

document classification. In the experiments, it takes either the objective or the subjective label of given documents as the response value, and predict the label on testing data after it converges on the training set.

8) i*osLDA-Single:* In *i*osLDA, words have different powers, respectively, in delivering objective and subjective senses. In order to validate that the identification of objective and subjective power is attractive and effective for different discriminant tasks, a variant of *i*osLDA denoted as *i*osLDA-single is evaluated, in which the objective and the subjective powers of each word are not differentiated, and the two impact scalers of one word are added up to be the total scaler of the word.
9) i*osLDA-Binary:* Another derived variant of *i*osLDA aiming to examine the effectiveness of discovering different degrees of discriminative power. In *i*osLDA-binary, all the impact scalers that are either objective or subjective have only two possible values: 0 and 1. Thus, all the discriminative words have the equal power, while the trivial words are simply ignored.
10) i*osLDA-Complete:* In the following part of this paper, the complete model of *i*osLDA that combines *i*osLDA-single and *i*osLDA-binary is denoted as *i*osLDA-complete.

The Gaussian kernel is adopted in the implemented SVM algorithm. As the result of model selection, the number of topics is set to be 20, and the hyper parameters are configured that $\alpha = 0.01$ and $\beta = 0.1$ for all the approaches; particularly, for the three models based on *i*osLDA, the $a\ priori$ of impact scaler proportion is fixed that $\gamma^O = 0.8$ and $\gamma^S = 0.5$, while the maximum value of the scalers L is set to be ten for simplicity. Initially, all the topic assignments are generated randomly, while the impact scalers (except for those in *i*osLDA-binary) are sampled from a Gaussian distribution with parameters $\mu = 5$ and $\sigma^2 = (5/3)$, and then adjusted to the nearest integer in $[0, 10]$. All the results reported in comparisons are the average performance of each model after ten repeated random experiments.

C. Object and Sentiment Classification

The performance of different approaches is first evaluated by objective and subjective classifications. Five metrics are utilized for the comparisons, i.e., accuracy, microaveraged Area Under Curve (AUC), macroaveraged AUC, microaveraged F1, and macro-averaged F1. It is worth noting that AUC has the ability to depict the tradeoff between true-positive and false-negative results in classification [42] and, therefore, is a good supplement to other metrics, such as accuracy and the F1 score.

Tables III and IV, respectively, report the performance of classification in terms of objective and subjective senses on all the data sets, including their means and standard deviations. It is worth noting that the *i*osLDA model and its variants have the ability to simultaneously predict object labels and sentiment labels, while other methods have to complete them separately. The best result in each metric is shown in bold. It is observed that methods with logistic regression obtain the worst results, since logistic regression is fit for binary classification tasks, but not so capable in the multiclass identification; multiclass sLDA performs better than logistic regression but falls behind with SVM, which probably due to the fact that SVM is an extremely discriminative method, but the softmax regression embedded in the model of multiclass sLDA is less predictive. All the aforementioned methods employ BoW or BoT representation to learn a classifier, while *i*osLDA-single, *i*osLDA-binary, and *i*osLDA-complete adopt the BoDW representation and outperform the traditional methods in terms of almost every metric on the three data sets. Taking a closer look at the performance of *i*osLDA and its two variants, it is also observed that *i*osLDA-binary achieves better performance than *i*osLDA-single to some extent, demonstrating the necessity of, respectively, discovering discriminative power of different words in delivering the objective and subjective senses. Meanwhile, the complete version of the *i*osLDA model has the best performance among all the approaches, which indicates that it is more effective to model the various degree of the discriminative power (i.e., the impact factors) than a simple binary assumption on it.

D. Generalization Ability

The *perplexity* on testing data is commonly exploited in measuring the generalization ability of topic models; generally, a lower perplexity indicates a better generalization ability [8], which suggests higher performance in topic modeling. The perplexity obtained over all the three data sets by LDA, multiclass sLDA, *i*osLDA-binary, and *i*osLDA-complete is shown in Fig. 3, where the performance of *i*osLDA-single is omitted, since it is extremely close to *i*osLDA-complete. It is observed that multiclass sLDA gains a higher perplexity than the original LDA in most cases, for sLDA aims to find the best representation over topics of the documents for classification rather than the generalization of unseen documents. In contrast, the *i*osLDA model (as well as its simplified variant), though having the similar mechanism to sLDA in generating the low-dimensional representation in a supervised manner, still obtains the lowest perplexity regardless of the number of topics. This may owe to the joint modeling of latent topics and the auxiliary information in terms of both objective and subjective senses, while the discovery of the discriminative and trivial words partially eliminates the compromise between topic modeling and the prediction of the response values.

E. Computational Efficiency

When the dimension of the latent representation (i.e., the best number of topics) required by the hidden structure of given data increases, traditional methods using the BoT representation for discriminative tasks suffer from the rise of computational complexity. Fig. 4 gives out the performance in terms of the runningtime of traditional multiclass sLDA for a single experiment on different data sets, which is compared with *i*osLDA-binary and *i*osLDA-complete proposed in this paper. Here, the performance of *i*osLDA-single is omitted for that it is theoretically identical to the one of *i*osLDA-complete. As a joint model for topic modeling

This article has been accepted for inclusion in a future issue of this journal. Content is final as presented, with the exception of pagination.

TABLE III

COMPARISONS OF OBJECT CLASSIFICATION

	BoW		LDA		BoW + LDA		sLDA	*i*osLDA		
	+SVM	+LR	+SVM	+LR	+SVM	+LR		single	binary	complete
Multi-Domain Sentiment Dataset										
Accuracy	0.6870	0.6150	0.6825$_{\pm 0.0014}$	0.6185$_{\pm 0.0028}$	0.6943$_{\pm 0.0013}$	0.6231$_{\pm 0.0026}$	0.6352$_{\pm 0.0067}$	0.6881$_{\pm 0.0037}$	0.6950$_{\pm 0.0050}$	**0.7076**$_{\pm 0.0042}$
Micro-AUC	0.7996	0.7043	0.7957$_{\pm 0.0054}$	0.7099$_{\pm 0.0052}$	0.8090$_{\pm 0.0048}$	0.7279$_{\pm 0.0048}$	0.7518$_{\pm 0.0087}$	**0.8122**$_{\pm 0.0074}$	0.8072$_{\pm 0.0081}$	0.8089$_{\pm 0.0076}$
Macro-AUC	0.7767	0.6716	0.7669$_{\pm 0.0028}$	0.6813$_{\pm 0.0075}$	0.7834$_{\pm 0.0027}$	0.6854$_{\pm 0.0072}$	0.7297$_{\pm 0.0085}$	0.7697$_{\pm 0.0060}$	0.7743$_{\pm 0.0062}$	**0.7911**$_{\pm 0.0070}$
Micro-F1	0.7896	0.7526	0.8113$_{\pm 0.0093}$	0.7635$_{\pm 0.0112}$	0.7952$_{\pm 0.0082}$	0.7842$_{\pm 0.0103}$	0.7768$_{\pm 0.0121}$	0.8009$_{\pm 0.0102}$	0.8201$_{\pm 0.0125}$	**0.8270**$_{\pm 0.0096}$
Macro-F1	0.7919	0.6694	0.7672$_{\pm 0.0084}$	0.6889$_{\pm 0.0092}$	0.7920$_{\pm 0.0078}$	0.6947$_{\pm 0.0088}$	0.7138$_{\pm 0.0096}$	0.7398$_{\pm 0.0108}$	0.7764$_{\pm 0.0099}$	**0.7996**$_{\pm 0.0090}$
Flickr Dataset										
Accuracy	0.5077	0.4532	0.5287$_{\pm 0.0041}$	0.5196$_{\pm 0.0035}$	0.5486$_{\pm 0.0041}$	0.4870$_{\pm 0.0032}$	0.5453$_{\pm 0.0141}$	0.5602$_{\pm 0.0043}$	0.5725$_{\pm 0.0053}$	**0.5842**$_{\pm 0.0048}$
Micro-AUC	0.6163	0.5459	0.5587$_{\pm 0.0082}$	0.5485$_{\pm 0.0097}$	0.5691$_{\pm 0.0085}$	0.5786$_{\pm 0.0104}$	0.5944$_{\pm 0.0122}$	0.6037$_{\pm 0.0084}$	0.6272$_{\pm 0.0099}$	**0.6275**$_{\pm 0.0093}$
Macro-AUC	0.6782	0.6225	0.6614$_{\pm 0.0078}$	0.6461$_{\pm 0.0081}$	0.6637$_{\pm 0.0084}$	0.6609$_{\pm 0.0084}$	0.6728$_{\pm 0.0098}$	0.6875$_{\pm 0.0087}$	0.6928$_{\pm 0.0092}$	**0.7128**$_{\pm 0.0081}$
Micro-F1	0.6280	0.6237	0.6917$_{\pm 0.0095}$	0.6839$_{\pm 0.0087}$	0.6968$_{\pm 0.0077}$	0.6676$_{\pm 0.0079}$	0.7058$_{\pm 0.0107}$	0.7237$_{\pm 0.0096}$	0.7281$_{\pm 0.0122}$	**0.7411**$_{\pm 0.0114}$
Macro-F1	0.7467	0.7355	0.8015$_{\pm 0.0079}$	0.8056$_{\pm 0.0081}$	0.8251$_{\pm 0.0076}$	0.7484$_{\pm 0.0080}$	0.8152$_{\pm 0.0114}$	0.8406$_{\pm 0.0077}$	0.8270$_{\pm 0.0100}$	**0.8424**$_{\pm 0.0094}$
Twitter Dataset										
Accuracy	0.4195	0.3356	0.4436$_{\pm 0.0026}$	0.4076$_{\pm 0.0062}$	0.4531$_{\pm 0.0022}$	0.4329$_{\pm 0.0058}$	0.4369$_{\pm 0.0070}$	0.4722$_{\pm 0.0045}$	0.4738$_{\pm 0.0030}$	**0.4851**$_{\pm 0.0034}$
Micro-AUC	0.5916	0.5453	0.6327$_{\pm 0.0038}$	0.5871$_{\pm 0.0069}$	0.6412$_{\pm 0.0038}$	0.6101$_{\pm 0.0066}$	0.6335$_{\pm 0.0075}$	0.6623$_{\pm 0.0048}$	0.6542$_{\pm 0.0057}$	**0.6812**$_{\pm 0.0049}$
Macro-AUC	0.7221	0.5944	0.7423$_{\pm 0.0030}$	0.7136$_{\pm 0.0058}$	0.7428$_{\pm 0.0028}$	0.7182$_{\pm 0.0062}$	0.7189$_{\pm 0.0034}$	0.7424$_{\pm 0.0074}$	0.7517$_{\pm 0.0044}$	**0.7785**$_{\pm 0.0031}$
Micro-F1	0.5982	0.5272	0.6592$_{\pm 0.0055}$	0.6096$_{\pm 0.0035}$	0.6656$_{\pm 0.0047}$	0.6294$_{\pm 0.0035}$	0.6487$_{\pm 0.0027}$	0.6691$_{\pm 0.0068}$	0.6743$_{\pm 0.0046}$	**0.6762**$_{\pm 0.0054}$
Macro-F1	0.7071	0.6367	0.7735$_{\pm 0.0049}$	0.7348$_{\pm 0.0043}$	0.7744$_{\pm 0.0042}$	0.7615$_{\pm 0.0039}$	0.7681$_{\pm 0.0053}$	0.7728$_{\pm 0.0033}$	0.7760$_{\pm 0.0035}$	**0.7997**$_{\pm 0.0036}$

TABLE IV

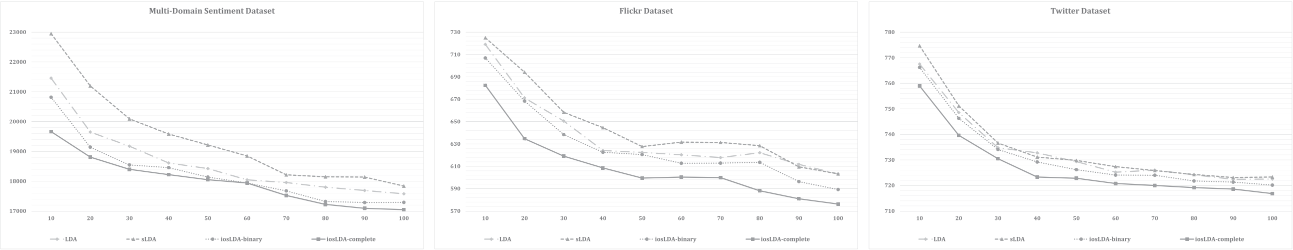
COMPARISONS OF SENTIMENT IDENTIFICATION

	BoW		LDA		BoW + LDA		sLDA	*i*osLDA		
	+SVM	+LR	+SVM	+LR	+SVM	+LR		single	binary	complete
Multi-Domain Sentiment Dataset										
accuracy	0.7320	0.6225	0.7952$_{\pm 0.0023}$	0.7656$_{\pm 0.0016}$	0.8038$_{\pm 0.0025}$	0.7662$_{\pm 0.0016}$	0.7675$_{\pm 0.0031}$	0.7964$_{\pm 0.0032}$	0.8124$_{\pm 0.0040}$	**0.8155**$_{\pm 0.0017}$
Micro-AUC	0.8020	0.6684	**0.8698**$_{\pm 0.0039}$	0.8521$_{\pm 0.0056}$	0.8635$_{\pm 0.0038}$	0.8626$_{\pm 0.0060}$	0.8569$_{\pm 0.0087}$	0.8669$_{\pm 0.0066}$	0.8613$_{\pm 0.0071}$	0.8572$_{\pm 0.0075}$
Macro-AUC	0.7269	0.6717	0.7781$_{\pm 0.0044}$	0.7719$_{\pm 0.0065}$	0.7918$_{\pm 0.0042}$	0.7891$_{\pm 0.0069}$	0.7684$_{\pm 0.0096}$	0.7936$_{\pm 0.0054}$	0.7920$_{\pm 0.0090}$	**0.7964**$_{\pm 0.0063}$
Micro-F1	0.8451	0.7857	0.8725$_{\pm 0.0078}$	0.8669$_{\pm 0.0090}$	0.8854$_{\pm 0.0083}$	0.8748$_{\pm 0.0087}$	0.8685$_{\pm 0.0123}$	0.8802$_{\pm 0.0094}$	0.8826$_{\pm 0.0099}$	**0.8886**$_{\pm 0.0092}$
Macro-F1	0.7453	0.6723	0.7806$_{\pm 0.0085}$	0.7224$_{\pm 0.0081}$	0.7899$_{\pm 0.0087}$	0.7598$_{\pm 0.0083}$	0.7120$_{\pm 0.0088}$	0.7866$_{\pm 0.0080}$	0.7897$_{\pm 0.0085}$	**0.7916**$_{\pm 0.0078}$
Flickr Dataset										
accuracy	0.4837	0.4397	0.4970$_{\pm 0.0065}$	0.4789$_{\pm 0.0070}$	0.5039$_{\pm 0.0058}$	0.4801$_{\pm 0.0065}$	0.5015$_{\pm 0.0115}$	0.5068$_{\pm 0.0087}$	0.5151$_{\pm 0.0084}$	**0.5208**$_{\pm 0.0073}$
Micro-AUC	0.6417	0.5929	0.6431$_{\pm 0.0096}$	0.6186$_{\pm 0.0077}$	0.6522$_{\pm 0.0075}$	0.6489$_{\pm 0.0071}$	0.6402$_{\pm 0.0078}$	0.6522$_{\pm 0.0093}$	0.6612$_{\pm 0.0090}$	**0.6680**$_{\pm 0.0094}$
Macro-AUC	0.7388	0.6902	0.7497$_{\pm 0.0079}$	0.7436$_{\pm 0.0080}$	0.7564$_{\pm 0.0062}$	0.7600$_{\pm 0.0073}$	0.7429$_{\pm 0.0087}$	0.7611$_{\pm 0.0079}$	0.7541$_{\pm 0.0098}$	**0.7692**$_{\pm 0.0065}$
Micro-F1	0.6046	0.5615	0.6642$_{\pm 0.0099}$	0.6476$_{\pm 0.0121}$	0.6643$_{\pm 0.0087}$	0.6591$_{\pm 0.0099}$	0.6683$_{\pm 0.0107}$	0.6725$_{\pm 0.0106}$	0.6803$_{\pm 0.0112}$	**0.6814**$_{\pm 0.0095}$
Macro-F1	0.6502	0.6109	0.7015$_{\pm 0.0097}$	0.7016$_{\pm 0.0114}$	0.7045$_{\pm 0.0089}$	0.7023$_{\pm 0.0105}$	0.7024$_{\pm 0.0093}$	0.7067$_{\pm 0.0114}$	0.7185$_{\pm 0.0102}$	**0.7251**$_{\pm 0.0081}$
Twitter Dataset										
accuracy	0.7349	0.6879	0.7454$_{\pm 0.0027}$	0.7117$_{\pm 0.0034}$	0.7811$_{\pm 0.0030}$	0.7225$_{\pm 0.0031}$	0.7587$_{\pm 0.0051}$	0.7727$_{\pm 0.0035}$	0.7871$_{\pm 0.0032}$	**0.7907**$_{\pm 0.0029}$
Micro-AUC	0.6764	0.6102	0.7080$_{\pm 0.0041}$	0.6075$_{\pm 0.0056}$	0.7242$_{\pm 0.0042}$	0.6093$_{\pm 0.0061}$	0.6281$_{\pm 0.0044}$	0.6922$_{\pm 0.0062}$	**0.7381**$_{\pm 0.0070}$	0.7255$_{\pm 0.0055}$
Macro-AUC	0.5742	0.5346	0.6125$_{\pm 0.0026}$	0.5868$_{\pm 0.0039}$	0.6253$_{\pm 0.0026}$	0.5908$_{\pm 0.0040}$	0.5823$_{\pm 0.0038}$	0.6206$_{\pm 0.0042}$	**0.6442**$_{\pm 0.0066}$	0.6216$_{\pm 0.0026}$
Micro-F1	0.7978	0.7153	0.8322$_{\pm 0.0046}$	0.8148$_{\pm 0.0066}$	0.8503$_{\pm 0.0041}$	0.8172$_{\pm 0.0068}$	0.8442$_{\pm 0.0040}$	0.8529$_{\pm 0.0054}$	0.8683$_{\pm 0.0069}$	**0.8687**$_{\pm 0.0046}$
Macro-F1	0.7380	0.6447	0.7526$_{\pm 0.0033}$	0.6645$_{\pm 0.0074}$	0.8185$_{\pm 0.0031}$	0.6737$_{\pm 0.0071}$	0.7664$_{\pm 0.0054}$	0.7776$_{\pm 0.0030}$	0.8112$_{\pm 0.0038}$	**0.8263**$_{\pm 0.0041}$

and document classification, *i*osLDA also proved to be more efficient than the traditional methods, such as sLDA. With the increase of the number of topics, the time consumption of sLDA grows in a linear manner, while it takes far less extra seconds for *i*osLDA to complete its computation. It is because, though having more latent variables in its model, the sampling of topics in *i*osLDA is actually identical to the standard LDA [as mentioned in the derivation of (1)], while the sampling of impact factors (4) and (5) is also efficient enough. As for sLDA, each probable topic assignment in a document will change the topic proportions employed to generate the response values (i.e., the object or sentiment labels), making the sampling process much more complex than *i*osLDA.

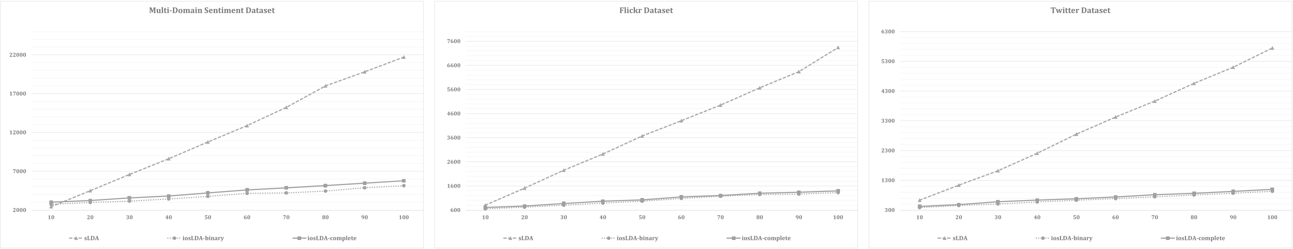
F. Discovery of Discriminative Words in Documents

We demonstrated before that different words in a given document have their varying degree of power to describe the facts (i.e., the objective senses) or convey the personal opinions (i.e., the subjective senses) contained in this document, and that *i*osLDA is capable to discern such discriminative power of the words under a specified topic. Thus, it is necessary and interesting to observe how our approach discover and measure the discrimination hidden in the textual words. In Table V,



(a) (b) (c)

Fig. 3. Comparisons of perplexity obtained by different models on (a) Multidomain Sentiment Data Set, (b) Flickr Data Set, and (c) Twitter Data Set. The vertical axis is the perplexity and the horizontal axis is the number of topics. Note that the perplexity obtained by *i*osLDA-single on these data sets is omitted for its high similarity to the one of *i*osLDA-complete.



(a) (b) (c)

Fig. 4. Performance in terms of the running-time of the traditional sLDA model compared with *i*osLDA-binary and *i*osLDA-complete for a single experiment on (a) Multidomain Sentiment Data Set, (b) Flickr Data Set, and (c) Twitter Data Set. The vertical axis is the elapsed time in seconds and the horizontal axis is the number of topics.

TABLE V

OBJECTIVELY DISCRIMINATIVE WORDS IN TERMS OF EACH TOPIC

topic: *book*		topic: *movie*		topic: *camera*		topic: *mobile phone*	
author	8.4423	acting	9.0196	kodak	8.9478	software	9.4017
diet	8.2746	character	8.9232	digital	8.2315	card	8.9712
written	8.2152	dvd	8.4063	product	8.1262	phone	8.4262
pages	8.1850	movie	8.1162	setup	7.9676	battery	8.1396
novel	7.9304	film	7.7558	power	7.6703	earphones	7.8751
book	7.9085	funny	7.7165	product	7.4953	audio	7.6149
interesting	7.5954	music	7.4233	camera	7.3832	smart	7.4981
opinions	7.4841	series	7.1150	card	7.3748	sony	7.4901
character	7.2010	season	6.6887	photos	6.9974	buy	7.4756
library	7.0926	watch	6.5792	quality	6.8594	camera	7.4506

the most discriminative words in delivering the objective sense under four topics (i.e., "book," "movie," "camera," and "mobile phone," which are automatically mined and manually named) in the document data set (i.e., Multidomain Sentiment Data Set) are presented, while those with the greatest discrimination in terms of the subjective sense are shown in Table VI. The objectively and subjectively powerful words are listed in the descending order in terms of their corresponding discrimination, where the discrimination of one given word v under the topic k in terms of objective or subjective sense is evaluated as

$$\operatorname{disc}(k, w) = \sum_{l=0}^{L} \frac{l \cdot n_{k,v,l}}{n_{k,v}}. \tag{8}$$

Here, $n_{k,v,l}$ denotes the number of word v under topic k that has an objectively or subjectively discriminative power l.

Several conclusions can be drawn from the results in experiments. First, as observed in Tables V and VI, many powerful words delivering an objective sense are nouns, whereas those subjectively discriminative words are mainly adjectives; however, several important nouns (e.g., "anything," "fan," and "problem") actively take part in the identification of subjective senses, and vice versa ("interesting," "funny," "smart," and so on), for the product reviewers tend to use these words in the description of a particular object or sentiment. In the second place, while the objectively powerful words have relatively higher discrimination (note that the maximum of discriminative power is ten with regard to the settings of the experiments), words that carrying subjective senses commonly have discrimination that are more moderately (for five would be the theoretical mean and the border line between discriminative and trivial words), while *i*osLDA still mines out words that are very useful in sentiment classification (e.g., "love," "beautiful," and "expensive"). Finally, it is observed that some words bear the objectively or subjective discriminative power across various topics (like "power" and "card" in terms of objective sense, or "well," "price," and "beautiful" in terms of subjective sense), indicating the similarities between their corresponding topics; meanwhile, some words are discovered to be discriminative in terms of both objective and subjective senses (e.g., "character," "photos," and "funny"), suggesting that they not only describe a specific object, but also capable to be the indicators of the sentiments. These conclusions are reasonably in accord with human intuition, which demonstrates the effectiveness of *i*osLDA to some degree.

G. Localization of Objective and Sentimental Regions

It is quite a natural ability for human beings to disentangle a discriminative subset of sensory information from their surrounding visual field before interpreting a complex scene, which is often named as "focus of attention" or "visual

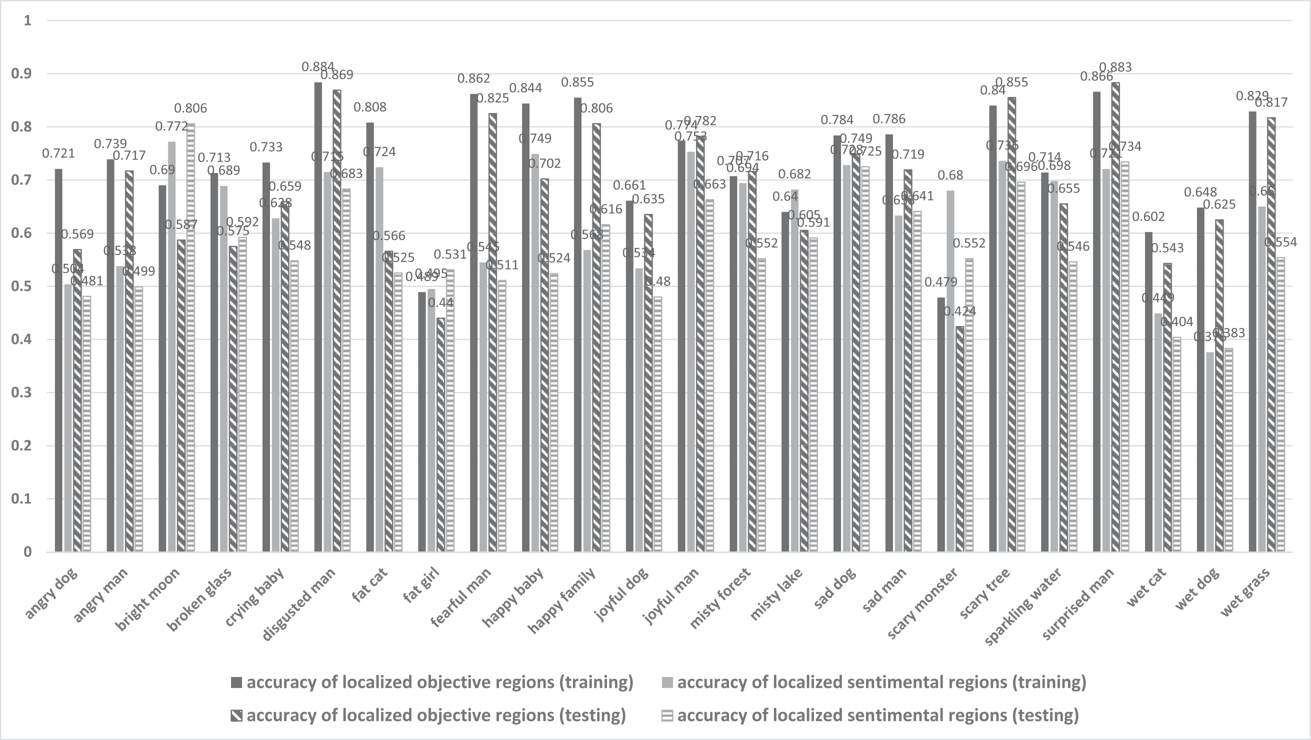


Fig. 5. Accuracy of detected discriminative visual words in terms of either objective or subjective sense on the training/testing set, respectively.

TABLE VI
SUBJECTIVELY DISCRIMINATIVE WORDS IN TERMS OF EACH TOPIC

topic: *book*		topic: *movie*		topic: *camera*		topic: *mobile phone*	
great	7.8309	love	8.3411	beautiful	7.7640	expensive	8.0066
history	7.6622	like	7.7192	expensive	7.5857	well	7.4398
point	7.2609	good	7.5722	broken	7.4056	battery	7.2120
anything	6.9976	character	7.4872	photos	7.3612	great	7.0873
favorite	6.6811	boring	7.0960	cheap	7.2267	big	7.0358
food	6.6202	actress	6.9403	quickly	6.9912	beautiful	6.9364
well	6.5743	funny	6.7630	price	6.9414	power	6.8520
recipes	6.5037	great	6.7030	well	6.9277	broken	6.7423
really	6.4224	fan	6.6784	power	6.6703	price	6.6812
best	6.2684	story	6.4272	works	6.4210	problem	6.5449

attention" [43]. Various models have been proposed to mimic such attractive characteristic of human eyes in order to acquire relevant stimuli from images that may contain complex and even obscure scenes. However, given one image, traditional "focus of attention" or "visual attention" is incapable of detecting objective regions (usually objects) and visual effects (usually emotions and sentiments) [44], [45]. The *i*osLDA model proposed in this paper has the inherent ability to locate both objective and sentimental regions in one given image. If the visual words that represents one region in the given image is identified as strongly objective (in practice, this equals to that the objective impact scaler of this word is larger than five), this region is detected to be a descriptor of the dominant object (e.g., "face" or "car") in the image. On the other hand, the regions consisting of visual words that are strongly subjective will be identified as a region that mainly describes the overall sentiment (e.g., "sad" or "surprise").

In order to validate the underlying ability of *i*osLDA, each image in the Flickr Data Set is manually labeled with two kinds of regions (i.e., objective region and sentimental region), and every identified region is marked by a square bounding box. The *i*osLDA is conducted over images from both training set and testing set, and the detected discriminative regions are compared with the ground truth. The detected discriminative visual words in terms of either objective or subjective sense are examined to see whether they are in the bounding boxes provided by the ground truth. Fig. 5 gives out the detection accuracy of discriminatively objective or subjective visual words in terms of different ANPs on both training and testing sets. In order to visually illustrate the detected object regions and sentimental regions, we mark circles having discriminative visual words to be the centers and radius of 25 pixels as objective or sentimental regions, and then compare them with the ground truth. Some comparisons are given in Fig. 6.

V. CONCLUSION

In this paper, a supervised topic model named as *i*osLDA is proposed to discover the words that either discriminative or trivial in delivering an objective or a subjective sense with respect to their assigned topics. To achieve this goal, first, the SPU model adopted in traditional topic models is modified by incorporating it with a probabilistic generative process, making it possible to obtain the novel BoDW representation for the documents; after that, each document is defined to have two different BoDW representations with regard to objective and subjective senses, respectively, which are employed in the joint objective and subjective classification instead of the traditional BoT representation. Results of various experiments

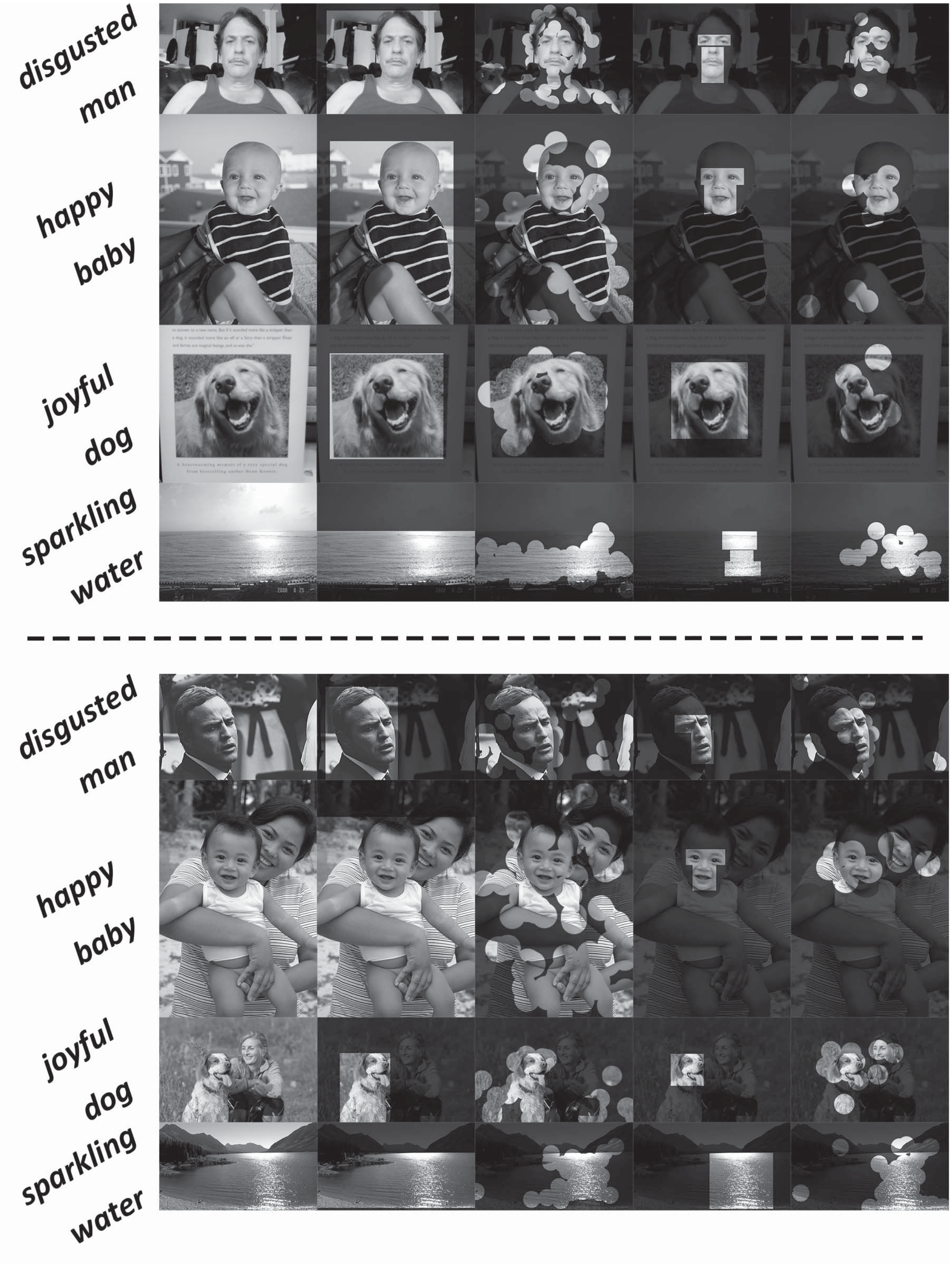


Fig. 6. Illustration of detected objective and sentimental regions by *i*osLDA. Columns from left to right are the ANPs, the original images, the ground truth of objective regions, the detected objective regions, the ground truth of sentimental regions, and the detected sentimental regions. The four rows on the top are generated during training, while the others are results from the testing set (best viewed in color).

indicate that: 1) the BoDW representation is more predictive than the traditional ones; 2) *i*osLDA boosts the performance of topic modeling via the joint discovery of latent topics and the different objective and subjective power hidden in every word; and 3) *i*osLDA has lower computational complexity than sLDA, especially under an increasing number of topics.

APPENDIX
DERIVATIONS IN POSTERIOR INFERENCE

The full derivations of 1–3 are given in this section. First, the joint distribution of *i*osLDA can be written as

$$\begin{aligned}p(\mathbf{\Phi}|\Psi) = \; & p(\boldsymbol{z}|\alpha)p(\boldsymbol{w}|\boldsymbol{z},\beta)\\ & \times p(\boldsymbol{x}^O|\gamma^O,\boldsymbol{z},\boldsymbol{w})p(\boldsymbol{x}^S|\gamma^S,\boldsymbol{z},\boldsymbol{w})\\ & \times p(\boldsymbol{y}^O|\eta^O,\boldsymbol{w},\boldsymbol{x}^O)p(\boldsymbol{y}^S|\eta^S,\boldsymbol{w},\boldsymbol{x}^S) \qquad (9)\end{aligned}$$

while the six terms in the right-hand side of (9) can be further expanded that

$$\begin{aligned}& p(\boldsymbol{z}|\alpha)\\ & \quad = \int p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha)d\boldsymbol{\theta}\\ & \quad = \prod_{d=1}^{D} \frac{\Gamma\big(\sum_{k=1}^{K}\alpha_k\big)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \frac{\prod_{k=1}^{K}\Gamma(\alpha_k+n_{d,k})}{\Gamma\big(\sum_{k=1}^{K}(\alpha_k+n_{d,k})\big)} \qquad (10a)\end{aligned}$$

$$\begin{aligned}& p(\boldsymbol{w}|\boldsymbol{z},\beta)\\ & \quad = \int p(\boldsymbol{w}|\boldsymbol{z},\boldsymbol{\phi})p(\boldsymbol{\phi}|\beta)d\boldsymbol{\phi}\\ & \quad = \prod_{k=1}^{K} \frac{\Gamma\big(\sum_{v=1}^{V}\beta_v\big)}{\prod_{v=1}^{V}\Gamma(\beta_v)} \frac{\prod_{v=1}^{V}\Gamma(\beta_v+n_{k,v})}{\Gamma\big(\sum_{v=1}^{V}(\beta_v+n_{k,v})\big)} \qquad (10b)\end{aligned}$$

$$\begin{aligned}& p(\boldsymbol{x}^O|\gamma^O,\boldsymbol{z},\boldsymbol{w})\\ & \quad = \int p(\boldsymbol{x}^O|\boldsymbol{z},\boldsymbol{w},\boldsymbol{\lambda}^O)p(\boldsymbol{\lambda}^O|\gamma^O)d\boldsymbol{\lambda}^O\\ & \quad = \prod_{k=1}^{K}\prod_{v=1}^{V} \frac{\Gamma\big(\sum_{l=0}^{L}\gamma_l^O\big)}{\prod_{l=0}^{L}\Gamma(\gamma_l^O)} \frac{\prod_{l=0}^{L}\Gamma\big(\gamma_l^C+n_{k,v,l}^C\big)}{\Gamma(\sum_{l=0}^{L}\gamma_l^C+n_{k,v})} \qquad (10c)\end{aligned}$$

$$\begin{aligned}& p(\boldsymbol{x}^S|\gamma^S,\boldsymbol{z},\boldsymbol{w})\\ & \quad = \int p(\boldsymbol{x}^S|\boldsymbol{z},\boldsymbol{w},\boldsymbol{\lambda}^S)p(\boldsymbol{\lambda}^S|\gamma^S)\mathrm{d}\boldsymbol{\lambda}^S\\ & \quad = \prod_{k=1}^{K}\prod_{v=1}^{V} \frac{\Gamma\big(\sum_{l=0}^{L}\gamma_l^S\big)}{\prod_{l=0}^{L}\Gamma\big(\gamma_l^S\big)} \frac{\prod_{l=0}^{L}\Gamma\big(\gamma_l^S+n_{k,v,l}^S\big)}{\Gamma\big(\sum_{l=0}^{L}\gamma_l^S+n_{k,v}\big)} \qquad (10d)\end{aligned}$$

$$\begin{aligned}& p(\boldsymbol{y}^O|\boldsymbol{w},\boldsymbol{x}^O,\eta^O)\\ & \quad = \prod_{d=1}^{D} \frac{\exp\big(\sum_{i=1}^{N_d}\eta^O_{y_d^O,w_{d,i}}w_{d,i}x_{d,i}^O\big)}{\sum_{o=1}^{O}\exp\big(\sum_{i=1}^{N_d}\eta^O_{o,w_{d,i}}w_{d,i}x_{d,i}^O\big)} \qquad (10e)\end{aligned}$$

$$\begin{aligned}& p(\boldsymbol{y}^S|\boldsymbol{w},\boldsymbol{x}^S,\eta^S)\\ & \quad = \prod_{d=1}^{D} \frac{\exp\big(\sum_{i=1}^{N_d}\eta^S_{y_d^S,w_{d,i}}w_{d,i}x_{d,i}^S\big)}{\sum_{s=1}^{S}\exp\big(\sum_{i=1}^{N_d}\eta^S_{s,w_{d,i}}w_{d,i}x_{d,i}^S\big)}. \qquad (10f)\end{aligned}$$

Then, it is able to derive rules for updating the Markov chain. First, with other variables fixed, a specific topic assignment $z_{d,i}$ is sampled following the derivation that:

$$\begin{aligned}& p(z_{d,i}=k|\mathbf{\Phi}_{-z_{d,i}},\Psi)\\ & \quad = \frac{p(\mathbf{\Phi}|\Psi)}{p(\mathbf{\Phi}_{-\{z_{d,i},w_{d,i}\}}|\Psi)p(w_{d,i}|\Psi)\sum_{z_{d,i}=1}^{K}p(w_{d,i}|\beta,z_{d,i})}\\ & \quad \propto \frac{p(\mathbf{\Phi}|\Psi)}{p(\mathbf{\Phi}_{-\{z_{d,i},w_{d,i}\}}|\Psi)}. \qquad (11)\end{aligned}$$

After applying (10a)–(10f) into the derivation and omitting terms that are equal in the fraction, it can be further simplified that

$$\begin{aligned}p(z_{d,i}=k|\mathbf{\Phi}_{-z_{d,i}},\Psi) & \propto \frac{p(\boldsymbol{z}|\alpha)}{p(\boldsymbol{z}_{-\{d,i\}}|\alpha)} \frac{p(\boldsymbol{w}|\beta,\boldsymbol{z})}{p(\boldsymbol{w}_{-\{d,i\}}|\beta,\boldsymbol{z}_{-\{d,i\}})}\\ & \propto \frac{\alpha_k+n_{d,k}}{\sum_{k=1}^{K}(\alpha_k+n_{d,k})} \frac{\beta_{w_{d,i}}+n_{k,w_{d,i}}}{\sum_{v=1}^{V}(\beta_v+n_{k,v})}\\ & \propto (\alpha_k+n_{d,k})\frac{\beta_{w_{d,i}}+n_{k,w_{d,i}}}{\sum_{v=1}^{V}(\beta_v+n_{k,v})}. \qquad (12)\end{aligned}$$

While the topic assignments $\boldsymbol{z}$ are determined, the rules for generating the impact scalers can be obtained from similar derivations. To sample the specific scaler $x_{d,i}^O$, we have

$$\begin{aligned}& p\big(x_{d,i}^O=l|\mathbf{\Phi}_{-x_{d,i}^O},\Psi\big)\\ & \quad \propto \frac{p(\boldsymbol{x}^O|\gamma^O,\boldsymbol{z},\boldsymbol{w})}{p\big(\boldsymbol{x}^O_{-\{d,i\}}|\gamma^O,\boldsymbol{z}_{-\{d,i\}}\boldsymbol{w}_{-\{d,i\}}\big)}p\big(y_d^O|\eta^O,\boldsymbol{x}^O,\boldsymbol{w}\big)\\ & \quad \propto \frac{\gamma_l^O+n^O_{k,w_{d,i},l}}{\sum_{j=0}^{L}\gamma_j^O+n_{k,w_{d,i}}} \frac{\exp\big(\sum_{i=1}^{N_d}\eta^O_{y_d^O,w_{d,i}}w_{d,i}x_{d,i}^O\big)}{\sum_{o=1}^{O}\exp\big(\sum_{i=1}^{N_d}\eta^O_{o,w_{d,i}}w_{d,i}x_{d,i}^O\big)} \qquad (13)\end{aligned}$$

where $l \in [0, L]$. The generation of $x_{d,i}^S$ follows a similar rule.

REFERENCES

[1] D. M. Blei, L. Carin, and D. Dunson, "Probabilistic topic models," *IEEE Signal Process. Mag.*, vol. 27, no. 6, pp. 55–65, Nov. 2010.

[2] T. Hofmann, "Probabilistic latent semantic indexing," in *Proc. 22nd Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 1999, pp. 50–57.

[3] M. W. Berry, S. T. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," *SIAM Rev.*, vol. 37, no. 4, pp. 573–595, 1995.

[4] N. Chen, J. Zhu, F. Sun, and B. Zhang, "Learning harmonium models with infinite latent features," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 3, pp. 520–532, Mar. 2014.

[5] T.-C. Chou and M. C. Chen, "Using incremental PLSI for threshold-resilient online event analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 3, pp. 289–299, Mar. 2008.

[6] J. T. Chien and M. S. Wu, "Adaptive Bayesian latent semantic analysis," *IEEE Trans. Audio, Speech, Language Process.*, vol. 16, no. 1, pp. 198–207, Jan. 2008.

[7] N. K. Bassiou and C. L. Kotropoulos, "Online PLSA: Batch updating techniques including out-of-vocabulary words," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 11, pp. 1953–1966, Nov. 2014.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[9] D. M. Blei, M. I. Jordan, and A. Y. Ng, "Hierarchical Bayesian models for applications in information retrieval," in *Proc. 7th Valencia Int. Meeting Bayesian Statist.*, vol. 7, Sep. 2003, pp. 25–43.

[10] M. Girolami and A. Kabán, "On an equivalence between PLSI and LDA," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2003, pp. 433–434.

[11] T. L. Griffiths, M. I. Jordan, J. B. Tenenbaum, and D. M. Blei, "Hierarchical topic models and the nested chinese restaurant process," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 16. 2004, pp. 17–24.

[12] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *J. Amer. Statist. Assoc.*, vol. 101, no. 476, pp. 1566–1581, Dec. 2006.

[13] D. M. Blei and P. I. Frazier, "Distance dependent chinese restaurant processes," *J. Mach. Learn. Res.*, vol. 12, pp. 2461–2488, Aug. 2011.

[14] D. M. Blei and M. I. Jordan, "Modeling annotated data," in *Proc. 26th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2003, pp. 127–134.

[15] C. Wang, D. M. Blei, and D. Heckerman, "Continuous time dynamic topic models," in *Proc. 24th Conf. Uncertainty Artif. Intell.*, 2008, pp. 579–586.

[16] D. Wang, S. Zhu, T. Li, and Y. Gong, "Multi-document summarization using sentence-based topic models," in *Proc. ACL-IJCNLP*, 2009, pp. 297–300.

[17] Z. Chen, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Leveraging multi-domain prior knowledge in topic models," in *Proc. 23th Int. Joint Conf. Artif. Intell.*, 2013, pp. 2071–2077.

[18] J.-T. Chien, "Hierarchical theme and topic modeling," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 3, pp. 565–578, Mar. 2016.

[19] M. M. Shafiei and E. E. Milios, "Latent Dirichlet co-clustering," in *Proc. 6th Int. Conf. Data Mining*, 2006, pp. 542–551.

[20] E. P. Xing, R. Yan, and A. G. Hauptmann, "Mining associated text and images with dual-wing harmoniums," in *Proc. 24th Conf. Uncertainty Artif. Intell.*, 2005, pp. 1–9.

[21] X. Wang, N. Mohanty, and A. McCallum, "Group and topic discovery from relations and their attributes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 18. 2006, pp. 1449–1456.

[22] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 20. 2008, pp. 121–128.

[23] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1903–1910.

[24] S. Yang, S. P. Crain, and H. Zha, "Bridging the language gap: Topic adaptation for documents with different technicality," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, Apr. 2011, pp. 823–831.

[25] Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai, "Topic sentiment mixture: Modeling facets and opinions in weblogs," in *Proc. 16th Int. Conf. World Wide Web*, 2007, pp. 171–180.

[26] B. Lu, M. Ott, C. Cardie, and B. K. Tsou, "Multi-aspect sentiment analysis with topic models," in *Proc. 11th Int. Conf. Data Mining Workshops*, 2011, pp. 81–88.

[27] P. McCullagh, "Generalized linear models," *Eur. J. Oper. Res.*, vol. 16, no. 3, pp. 285–292, 1984.

[28] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 111–120.

[29] C. Lin, Y. He, R. Everson, and S. Ruger, "Weakly supervised joint sentiment-topic detection from text," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1134–1145, Jun. 2012.

[30] D. A. Sprott, "Urn models and their application—An approach to modern discrete probability theory," *Technometrics*, vol. 20, no. 4, p. 501, 1978.

[31] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 262–272.

[32] Z. Chen and B. Liu, "Topic modeling using topics from many domains, lifelong learning and big data," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 703–711.

[33] N. Guan, D. Tao, Z. Luo, and B. Yuan, "Online nonnegative matrix factorization with robust stochastic approximation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 7, pp. 1087–1099, Jul. 2012.

[34] C. Jaegul, L. Changhyun, C. K. Reddy, and P. Haesun, "Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization," *IEEE Trans. Vis. Comput. Graphics*, vol. 19, no. 12, pp. 1992–2001, Dec. 2013.

[35] R. R. Salakhutdinov and G. E. Hinton, "Replicated softmax: An undirected topic model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 22. 2009, pp. 1607–1614.

[36] H. Larochelle and S. Lauly, "A neural autoregressive topic model," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25. 2012, pp. 2708–2716.

[37] Y. Zheng, Y.-J. Zhang, and H. Larochelle, "Topic modeling of multimodal data: An autoregressive approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1370–1377.

[38] W. Fan, N. Bouguila, and D. Ziou, "Variational learning for finite Dirichlet mixture models and applications," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 5, pp. 762–774, May 2012.

[39] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proc. Nat. Acad. Sci.*, vol. 101. 2004, pp. 5228–5235.

[40] T. P. Minka, "Expectation propagation for approximate Bayesian inference," in *Proc. 17th Conf. Uncertainty Artif. Intell.*, 2001, pp. 362–369.

[41] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in *Proc. 21st ACM Int. Conf. Multimedia*, 2013, pp. 223–232.

[42] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[43] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[44] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 83–92.

[45] J. Jia, S. Wu, X. Wang, P. Hu, L. Cai, and J. Tang, "Can we understand van gogh's mood?: Learning to infer affects from images in social networks," in *Proc. 20th ACM Int. Conf. Multimedia*, 2012, pp. 857–860.

Hanqi Wang received the B.Sc. degree from the Computer Science and Technology College, Zhejiang University, Hangzhou, China, in 2012, where he is currently pursuing the Ph.D. degree with the College of Computer Science.

His current research interests include topic modeling and Bayesian statistics.

Fei Wu received the B.S. degree from Lanzhou University, Lanzhou, China, the M.S. degree from University of Macau, Zhuhai, China, and the Ph.D. degree from Zhejiang University, Hangzhou, China.

He was a Visiting Scholar with the Prof. B. Yu's Group, University of California at Berkeley, Berkeley, CA, USA, from 2009 to 2010. He is currently a Full Professor with the College of Computer Science and Technology, Zhejiang University. His current research interests include multimedia retrieval, sparse representation, and machine learning.

Weiming Lu received the B.Sc. and Ph.D. degrees in computer science from Zhejiang University, Hangzhou, China, in 2003 and 2009, respectively.

He is currently an Assistant Professor with the College of Computer Science and Technology, Zhejiang University. His current research interests include digital library, unstructured data management, and distributed system.

Yi Yang received the Ph.D. degree in computer science from Zhejiang University, Hangzhou, China, in 2010.

He was a Post-Doctoral Fellow with the School of Computer Science, Carnegie Mellon University at Australia, Adelaide, SA, Australia, from 2011 to 2013. He is currently a Senior Lecturer with the Centre for Quantum Computation and Intelligent Systems, University of Technology Sydney, Ultimo, NSW, Australia. His current research interests include machine learning and its applications to multimedia content analysis and indexing.

Xi Li received the Ph.D. degree from the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, Beijing, China, in 2009.

He was a Senior Researcher with the University of Adelaide, Adelaide, SA, Australia. From 2009 to 2010, he was a Post-Doctoral Researcher with the Centre National de la Recherche Scientifique, Telecom ParisTech, Paris, France. He is currently a Full Professor with Zhejiang University, Hangzhou, China. His current research interests include visual tracking, motion analysis, face recognition, Web data mining, and image and video retrieval.

Xuelong Li (M'02–SM'07–F'12) is a full professor with the Center for OPTical IMagery Analysis and Learning (OPTIMAL), State Key Laboratory of Transient Optics and Photonics, Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China.

Yueting Zhuang received the B.Sc., M.Sc., and Ph.D. degrees in computer science from Zhejiang University, Hangzhou, China, in 1986, 1989, and 1998, respectively.

From 1997 to 1998, he was a Visiting Scholar with the Prof. T. Huang's Group, University of Illinois at Urbana–Champaign, Champaign, IL, USA. He is currently a Full Professor and the Dean of the College of Computer Science, Zhejiang University. His current research interests include artificial intelligence, multimedia retrieval, computer animation, and digital library.