

Boosted Multi-Modal Supervised Latent Dirichlet Allocation for Social Event Classification

Shengsheng Qian, Tianzhu Zhang, Changsheng Xu
 Institute of Automation, Chinese Academy of Sciences, P. R. China
 Email:{shengsheng.qian, tzzhang, csxu}@nlpr.ia.ac.cn

Abstract—With the rapidly increasing popularity of Social Media sites (e.g., Flickr, YouTube, and Facebook), it is convenient for users to share their own comments on many social events, which successfully facilitates social event generation, sharing and propagation and results in a large amount of user-contributed media data (e.g., images, videos, and texts) for a wide variety of real-world events of different types and scales. As a consequence, it has become more and more difficult to find exactly the interesting events from massive social media data, which is useful to browse, search and monitor social events by users or governments. To deal with these issues, we propose a novel boosted multi-modal supervised Latent Dirichlet Allocation (BMM-SLDA) for social event classification. Our BMM-SLDA has a number of advantages. (1) It can effectively exploit the multi-modality and the supervised information of social events jointly. (2) It is suitable to large-scale data analysis by utilizing boosting weighted sampling strategy to iteratively select a small subset data to efficiently train the corresponding topic models. (3) It effectively exploits boosting document weight distribution by classification error, and can iteratively learn new topic model to correct the previously misclassified documents. We evaluate our BMM-SLDA on a real-world dataset and show extensive results, which show that our model outperforms state-of-the-art methods.

I. INTRODUCTION

With the explosion of Internet bandwidth, there are more and more social media sites (e.g., Flickr, YouTube, Facebook, and Google News) for people to capture and share social media data online. As a result, a popular event that is happening around us and around the world can spread very fast, and there are substantial amounts of events with multi-modality (e.g., images, videos, and texts) in Internet. Most of these social events uploaded by users are related with some specific topics, and it is time-consuming to manually identify or cluster them. Therefore, automatically mining and identifying social events from massive social media data is important and helpful to better browse, search and monitor social events by users or governments. However, it is difficult to achieve this goal because: (1) The number of social event is substantial, and it is a big data problem; (2) Unlike normal textual documents, social events consist of both textual information and visual information, and have multi-modal property.

Recently, researchers have proposed many methods to automatically mine and monitor social events, such as social event classification [1], social event tracking [2], social event mining [3], and investigating event detection [4]. For most of the existing work, the major idea is to design efficient feature for social event modeling. As in [5], [6], textual features are used. Moreover, visual information, such as, images and videos, is

also useful for social event classification, because different users have different textual descriptions (comments, tags, etc.) to represent the same social event while they can have similar visual information to represent them. Therefore, how to adopt multi-modal features becomes more and more important for social event analysis [1]. In many existing methods [1], [7], different multi-modal features, such as tag, time, location or visual features are exploited, and as a result encouraging performance is achieved. However, most of these methods ignore the semantic relationship among multiple modalities.

To address these issues, we attempt to adopt topic model to exploit the semantic relationship among multiple modalities for social event modeling. We assume that social events represented with different modalities but describing the same concept are quite related in their hidden topic. Therefore, it is suitable to adopt topic model based methods to mine multi-modal topics of social events. For simplicity, we take Flickr, one of the most popular photo sharing websites, as the social media platform in our study of social event analysis. We exploit the rich context associated with social media contents for social event modeling, including user-provided annotations (e.g., title, tags) and visual image information, which sufficiently considers the multi-modality of social media data. Each image and its corresponding textual data (e.g., title and tags) are considered as one social media document. Under this setup, we can extend the traditional Latent Dirichlet Allocation (LDA) with multi-modal property. However, the multi-modal Latent Dirichlet Allocation (mmLDA) is unsupervised topic model that cannot use the supervised category labels, which are able to boost the classification performance. To overcome this problem, we propose a novel multi-modal supervised Latent Dirichlet Allocation (mm-SLDA) to consider the supervised information. Moreover, all the above topic models ignore the document weight distribution and their training process is time-consuming on a big dataset. Therefore, we utilize boosting weighted sampling algorithm to reduce the number of training documents by considering the document weight distribution and iteratively improve the model.

Inspired by the above discussions, we propose a novel boosted multi-modal supervised latent dirichlet allocation (BMM-SLDA) algorithm to iteratively obtain multiple classifiers for social event classification. The basic idea is to integrate a supervised topic model process in the boosting framework. Each iteration of boosting begins to select a small part of documents from large-scale training data according to

their weights assigned by the previous boosting step. Based on the sampled small subset data, the proposed mm-SLDA is applied to learn the corresponding topic model. The resulting topic model is then applied to learn a new classifier. Based on the learned classifier, the documents in the training data are classified to obtain the classification scores. Finally, the new weights of documents in training data are updated by the use of the classification scores. Based on the above procedure, it is clear that the documents are iteratively reused with different weights to learn multiple topic models to build a strong social event classifier. In a word, our algorithm has two steps. In the training step, we iteratively learn multiple topic models for selected documents from the training data, and obtain the corresponding weak classifiers. In the test step, the documents are described by the learned topic models and classified with the combination of the corresponding weak classifiers to determine their final class labels. Compared with existing methods, the contributions of this work are four-fold.

- Our BMM-SLDA algorithm is suitable to large-scale data analysis by utilizing boosting weighted sampling strategy to iteratively select a small subset data to efficiently train the corresponding topic models.
- Our BMM-SLDA effectively exploits boosting document weight distribution by classification error, and can iteratively learn new topic model to correct the previously misclassified documents.
- Our proposed mm-SLDA can effectively exploit the multi-modal property and the supervised information of social event jointly.
- We collect a big dataset for research on social event classification with multi-modality information, and will release it for academic use.

II. RELATED WORK

Due to the limited space, we briefly review previous methods which are most related to our work including event classification methods and existing topic model methods.

Event Classification: With the massive growth of social events in Internet, efficient organization and monitoring of social events becomes a challenge. Researchers have been working on social event analysis and proposed many different methods [1], [7], which are based on single-modality (e.g., text, images) information or multi-modality information. About the single-modality analysis, many existing methods adopt textual information (e.g., names, time references, locations, title, tags, and description) or visual information (e.g., images and videos) [5], [6], [8] to model social event. In [8], the authors studied the correlation between manually annotated visual concepts and topic annotations in story clustering. However, the single-modality based methods ignore the multi-modal property of social event and cannot outperform the multi-modality based methods generally. To address this problem, many researchers attempt to adopt multiple different features (e.g., time, tag, location feature, images, and videos) to calculate the similarity of social event documents [1], [7]. While the above methods focus on feature design to

improve experimental performance, the semantic relationship and importance of those features have not been studied in details. Different from the above methods, we make use of the rich multi-modal contents associated with social events, including user-provided textual information (e.g., title, tags) and visual information (e.g., images), and propose a novel mm-SLDA to model the multi-modality and the supervised category labels jointly.

Topic Model: Topic model such as Latent Dirichlet Allocation (LDA) [9] has been widely applied to various applications and has many extensions, such as supervised Latent Dirichlet Allocation (SLDA) [10], [11], multi-modal Latent Dirichlet Allocation (mmLDA) [12]. The traditional LDA [9] and SLDA [11] mainly focus on how to apply the model to textual corpora, while the SLDA model using continuous response values via a liner regression cannot be used for multi-class classification problem [13] and it also does not consider multi-modal corpora. The mmLDA [12] considers multi-modal information, such as users' textual annotations and visual images, and is proposed for social relation mining. Our proposed topic model is different from the previous models. Compared with [12], our proposed model focuses on social event classification. Furthermore, the traditional multi-modal LDA [14] does not utilize the supervised category labels. We extend multi-modal LDA to a supervised topic model (mm-SLDA) with softmax regression function, which is used because the social events have multi-class property and can be classified into multiple classes directly. Moreover, all the above topic models ignore the document weight distribution and it is time-consuming to train the models on a big dataset. Different from these models, we utilize boosting weighted sampling strategy to iteratively select only a small part of training documents by considering the document weight distribution to train the corresponding topic models. As a result, our algorithm is much more efficient.

III. THE PROPOSED ALGORITHM

In this section, we first overview our algorithm for multi-modal social event classification. We then introduce our proposed model and its learning algorithm. Finally, we show how to use our proposed model for social event classification.

A. Overview

Given a set of social media documents, the problem that we address in this paper is how to identify events (e.g., Syrian civil war, US presidential election) that are reflected in the documents, as well as the documents that correspond to each event. A multimedia document consisting of an image and the corresponding text information (such as title, description, tags, etc.) is thus summarized as a pair of vectors of word counts. An image word is denoted as a unit-basis vector v of size D_v with exactly one non-zero entry representing the membership to only one word in a dictionary of D_v words. A text word w_n is similarly defined for a dictionary of size D_w . We cast our task as a classification problem over social media documents (e.g., images, title, description, tags). Let $\mathbf{E}_S = \{(e_1, y_1), (e_2, y_2), \dots, (e_M, y_M)\}$ denote a

Algorithm 1: Boosted Multi-Modal Supervised Latent Dirichlet Allocation for Social Event Classification.

input : Training Data: $\mathbf{E}_S = \{\mathbf{e}_i\}_{i=1}^M$, $\mathbf{Y}_S = \{y_i\}_{i=1}^M$. K and T . $d_1^i = 1/M$, $\forall i = 1, \dots, M$
output: weak classifiers $\{\mathbf{h}_t(e)\}_{t=1}^T$, coefficients $\{\alpha_t\}_{t=1}^T$, topic models $\{\mathbf{F}_t(e)\}_{t=1}^T$.

- 1 **for** $t = 1$ **to** T **do**
- 2 Sample \mathbf{E}_{S_t} from \mathbf{E}_S according to \mathbf{d}_t , and learn the topic models $\mathbf{F}_t(e)$ on \mathbf{E}_{S_t} as in Section III-B1.
- 3 Design a weak classifier $\mathbf{h}_t(e)$ as in Section III-B2.
- 4 Compute the error ϵ_t and α_t according to (2) and (3), respectively.
- 5 Update instance weight distribution \mathbf{d}_{t+1} as in (4) in Section III-B3.
- 6 **end**

training dataset of M documents, where $e_m = [\mathbf{v}_m, \mathbf{w}_m]$ is the m^{th} image-text pair and \mathbf{v}_m and \mathbf{w}_m denote the visual and textual description of the m^{th} image-text pair, respectively, and $y_m \in \{1, 2, \dots, C\}$ is the class label of the document e_m . Here, C is the number of event class labels. Let $\mathbf{E}_{\mathcal{T}} = \{e_i\}$ be the test documents. Our aim is to learn a classifier $\mathbf{H}(e)$ for $\forall e \in \mathbf{E}_{\mathcal{T}}$ with the assistance of the labeled set \mathbf{E}_S .

To achieve this goal, we propose a boosted topic model learning method to iteratively obtain multiple topic model classifiers for social event classification. As shown in Algorithm 1, it gives the details of the training step of the proposed BMM-SLDA algorithm. In the training step, multiple topic models are learned inside a boosting procedure with different training documents. In each iteration t , we sample a subset \mathbf{E}_{S_t} from the whole training document set \mathbf{E}_S according to their weights \mathbf{d}_t assigned by the previous boosting iteration. This subset \mathbf{E}_{S_t} is then used as a guide to learn the corresponding topic model parameters $\mathbf{F}_t(e)$ as introduced in Section III-B1. Then, the learned parameters are applied to learn a new weak learner $\mathbf{h}_t(e)$. Based on the classifier, the documents $e \in \mathbf{E}_S$ are classified to obtain their classification errors ϵ_t . Finally, the new weights of documents are updated by using the classification error ϵ_t . In this way, documents in the training set with large weights are more likely to be selected for training a new topic model $\mathbf{F}_{t+1}(e)$ in the next iteration. Once this procedure converges, we obtain a set of topic models $\{\mathbf{F}_t(e)\}$, and a set of weak learners $\{\mathbf{h}_t(e)\}$ and their corresponding combination coefficients $\{\alpha_t\}$ to get the final strong classifier $\mathbf{H}(e)$ as shown in (5). After training, each document $e \in \mathbf{E}_{\mathcal{T}}$ is mapped with topic models $\{\mathbf{F}_t(x)\}$ to obtain their descriptions. Then, each mapped description is classified by its corresponding weak classifier $\mathbf{h}_t(e)$. The predicted results of all T weak classifiers are combined to decide the final class label $\mathbf{H}(e)$. The detail is introduced next.

B. Our Boosting Model

For each iteration t , our algorithm has three major components, topic model learning, weak learner, and document weight update, described as follows.

1) *Topic Model Learning:* In each iteration t , a subset \mathbf{E}_{S_t} is sampled from the whole training set \mathbf{E}_S according to their weights \mathbf{d}_t . Then, on this subset, we apply our mm-SLDA to learn the corresponding topic model $\mathbf{F}_t(e) = \{\Phi^w, \Phi^v, \eta\}$, and the details will be introduced in Section IV. Once obtain the topic model $\mathbf{F}_t(e)$, we can predict the label of a new event document. Given a new social event document e_{new} , which is composed of many textual words w_{new} and associated visual words v_{new} , we first sample the topic assignments of all tokens including text words and visual words. Then, we can obtain the empirical topic proportion of various topics \bar{e}_{new} and adopt the learned class coefficients η for prediction.

2) *Weak Learner:* Once we obtain the topic model $\mathbf{F}_t(e_i)$, we need to design an effective weak learner $\mathbf{h}_t(e_i)$. Specifically, we adopt an effective softmax regression function. The the weak classifier is defined as in Eq.(1).

$$\begin{aligned} \mathbf{h}_t(e_{new}) &= \arg \max_{y_{e_{new}} \in \{1, 2, \dots, C\}} (p(y_{e_{new}} = c | \bar{e}_{new}, \eta)) \\ &= \arg \max_{y_{e_{new}} \in \{1, 2, \dots, C\}} (\eta_c^T \bar{e}_{new}) \end{aligned} \quad (1)$$

where $p(y_{e_{new}} = c | \bar{e}_{new}, \eta) = \exp(\eta_c^T \bar{e}_{new}) / \sum_{l=1}^C \exp(\eta_l^T \bar{e}_{new})$.

After constructing the weak learner, similar to the conventional multi-class AdaBoost scheme [15], we compute its classification error ϵ_t and assign a weight α_t for the weak learner $\mathbf{h}_t(e)$ as shown in (2) and (3), respectively. Here, $\mathbb{I}(\cdot)$ is an indicator function, so that $\mathbb{I}(a \text{ true statement}) = 1$, and $\mathbb{I}(a \text{ false statement}) = 0$.

$$\epsilon_t = \frac{1}{\sum_{i=1}^M d_t^i} \sum_{i=1}^M d_t^i \cdot \mathbb{I}(y_i \neq \mathbf{h}_t(e_i)) \quad (2)$$

$$\alpha_t = \ln \left((1 - \epsilon_t) / \epsilon_t \right) + \ln(C - 1) \quad (3)$$

3) *Document Weight Update:* Our document weight distribution update scheme is shown in (4). In each iteration of the conventional AdaBoost algorithm, weights of the misclassified documents are increased while weights of the correctly classified documents are decreased. It is defined by

$$d_{t+1}^i = d_t^i \cdot \exp \left(\alpha_t \cdot \mathbb{I}(y_i \neq \mathbf{h}_t(e_i)) \right), \forall i = 1, \dots, M, \quad (4)$$

where $\mathbf{d}_t = [d_t^1, d_t^2, \dots, d_t^M]^T$ and d_t^i is the weight of the i^{th} document in the \mathbf{E}_S .

C. Social Event Classifier

Once the boosting procedure converges, we obtain a set of topic models $\{\mathbf{F}_t(e)\}$, and a set of weak learners $\{\mathbf{h}_t(e)\}$ and their corresponding combination coefficients $\{\alpha_t\}$. Then, the learned social event classifier $\mathbf{H}(e)$ is

$$\mathbf{H}(e) = \arg \max_c \sum_{t=1}^T \alpha_t \cdot \mathbb{I}(\mathbf{h}_t(e) = c), c \in \{1, \dots, C\}. \quad (5)$$

IV. OUR MM-SLDA MODEL

Based on the sampled subset \mathbf{E}_{S_t} at the t^{th} iteration, we formulate the social event classification problem as mm-SLDA model, which can make use of the event multi-modal property and the event category information jointly to learn an effective and discriminative event model. The proposed model has the graphical representation as shown in Fig. 1. From the figure,

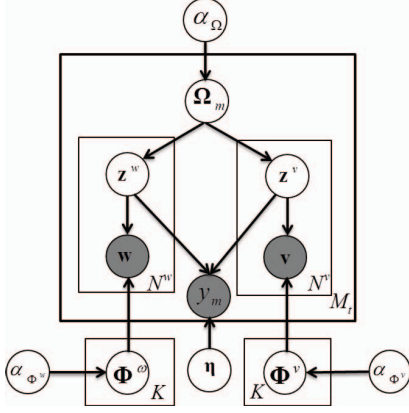


Fig. 1. The proposed mm-SLDA topic model for social event classification. For details, please refer to the corresponding text.

we can see that our model can mine the visual and textual topics of different social events together by considering the supervised label information. Input $M_t = |\mathbf{E}_{S_t}|$ documents with their labels y_m , our aim is to infer the event document distribution Ω_m , a set of C class coefficients $\eta_{1:C}$, and the K text and image topics Φ^w and Φ^v . Here, the K is the number of topics. The Ω_m represents that many tags and associated image in a social event document share the same document-specific distribution over topics. The inferred each coefficient η_c is a K dimensional vector, and represents the parameter values of softmax regression in the c^{th} class. The generative process of mm-SLDA for an image-text pair document m with N^v visual words, N^w text words and its label is given as follows:

1. For each visual topic $k \in \{1, 2, \dots, K\}$, Draw $\Phi^v | \alpha_{\Phi^v} \sim \text{Dir}(\alpha_{\Phi^v})$
2. For each textual topic $k \in \{1, 2, \dots, K\}$, Draw $\Phi^w | \alpha_{\Phi^w} \sim \text{Dir}(\alpha_{\Phi^w})$
3. Draw topic proportions $\Omega_m | \alpha \sim \text{Dir}(\alpha_{\Omega})$
4. For each visual word v_n , $n \in \{1, 2, \dots, N^v\}$
 - (a) Draw a topic assignment $z_n^v | \Omega_m \sim \text{Mult}(\Omega_m)$
 - (b) Draw a visual patch $v_n | z_n^v, \Phi^v \sim \text{Mult}(\Phi^v_{z_n^v})$
5. For each textual word w_n , $n \in \{1, 2, \dots, N^w\}$
 - (a) Draw a topic assignment $z_n^w | \Omega_m \sim \text{Mult}(\Omega_m)$
 - (b) Draw a word $w_n | z_n^w, \Phi^w \sim \text{Mult}(\Phi^w_{z_n^w})$
6. Draw class label $y_m | z_m \sim \text{softmax}(\bar{z}_m, \eta)$, where $\bar{z}_m = (\sum_v z_m^v + \sum_w z_m^w) / (N^v + N^w)$.

Here, the softmax function provides the following distribution $p(y_m | \bar{z}_m, \eta) = \exp(\eta_{y_m}^T \bar{z}_m) / \sum_{l=1}^C \exp(\eta_l^T \bar{z}_m)$, where K -dimensional vectors $\eta_{1:C}$ and \bar{z}_m represent a set of class coefficients in our mm-SLDA and the empirical proportion of textual and visual topics occurred in event document m , respectively. During the model learning process, we assume that the priors distributions follow symmetric Dirichlet, which are conjugate priors for multinomial. Exact inference is often intractable in many topic models and appropriate methods must be used, such as variational inference [9] and Gibbs sampling [16]. Gibbs sampling is a type of Markov chain Monte Carlo algorithm and is involved into an EM strategy for parameter inference in this paper. In EM terminology, we

sample the value of z by Gibbs sampling method given the parameters $\eta_{1:C}$ in E-step, and update $\eta_{1:C}$ by maximizing the joint likelihood of variables in M-step.

E-step: In the E-step, we adopt collapsed Gibbs sampling to sample from the distribution conditioned on the previous state. Under our model, the hidden variables z^w and z^v need to be assigned. The conditional posterior distribution of the latent topic indicators in text document can be written as

$$\begin{aligned}
 p(z_{m,i}^w = k | \mathbf{z}_{-(m,i),w}, \mathbf{w}, \mathbf{v}, \mathbf{y}, \alpha_{\Omega}, \alpha_{\Phi^w}, \alpha_{\Phi^v}, \eta) \\
 \propto \frac{(n_{m,k}^{-(i)} + \alpha_{\Omega})}{\sum_{k=1}^K (n_{m,k}^w + \alpha_{\Omega}) - 1} \frac{n_{q,k}^{-(m,i),w} + \alpha_{\Phi^w}}{\sum_{p=1}^{D_w} (n_{p,k}^w + \alpha_{\Phi^w}) - 1} \\
 \prod_{l=1}^C \{ \exp(\eta_l^T \bar{\mathbf{z}}) / \sum_{j=1}^C \exp(\eta_j^T \bar{\mathbf{z}}) \}^{1\{y^{(m)}=l\}} \quad (6)
 \end{aligned}$$

Here, $\mathbf{z}_{-(m,i),w}$ denotes the vectors of topic assignment except the considered word at position i in the textual information w of event document m , $n_{q,k}^{-(m,i),w}$ denotes the number of times of word q assigned to topic k except the current assignment in the text document w , $\sum_{p=1}^{D_w} (n_{p,k}^w + \alpha_{\Phi^w}) - 1$ denotes the total number of words assigned to topic k except the current assignment in the text document w , $n_{m,k}^{-(i)}$ denotes the number of text words and image patches in event document m assigned to topic k except the current assignment, $\sum_{k=1}^K (n_{m,k}^w + \alpha_{\Omega}) - 1$ denotes the total number of text words and image patches in event document m assigned to topic k except the current assignment, α_{Ω} , α_{Φ^w} , α_{Φ^v} are symmetric hyperparameters controlling the corresponding Dirichlet prior distributions, η denotes class coefficients and each class coefficient η_c is a K dimensional vector. The descriptions of parameters in images v are similar.

After Gibbs sampling, we can estimate Φ^w , Φ^v , Ω :

$$\Phi_{k,q}^w = \frac{n_{q,k}^w + \alpha_{\Phi^w}}{\sum_{p=1}^{D_w} (n_{p,k}^w + \alpha_{\Phi^w})} \quad (7)$$

$$\Phi_{k,q}^v = \frac{n_{q,k}^v + \alpha_{\Phi^v}}{\sum_{p=1}^{D_v} (n_{p,k}^v + \alpha_{\Phi^v})} \quad (8)$$

$$\Omega_{m,k} = \frac{(n_{m,k}^w + \alpha_{\Omega})}{\sum_{k=1}^K (n_{m,k}^w + \alpha_{\Omega})} \quad (9)$$

M-step: In the M-step, we update the class coefficients η by maximizing the joint likelihood. Because we fix parameters obtained in the E-step, it is equivalent to maximizing $p(\mathbf{y} | \bar{\mathbf{z}}, \eta)$ where each event document m is represented by $\bar{\mathbf{z}}$ newly updated in the E-step. Specifically, we learn L_2 -regularized softmax regression model which solves the following unconstrained optimization problem:

$$\min_{\eta} \left(-\frac{1}{M_T} \sum_{m=1}^{M_T} \sum_{c=1}^C 1\{y^{(m)} = c\} \log \frac{e^{\eta_c^T \bar{\mathbf{z}}_m}}{\sum_{l=1}^C e^{\eta_l^T \bar{\mathbf{z}}_m}} + \frac{\lambda}{2} \sum_{i=1}^C \eta_i^T \eta_i \right)$$

where λ is a regularization parameter and is set to be 1.0, and we apply a trust region Newton method [17] for optimization.

V. EXPERIMENTAL RESULTS

In this Section, we show extensive experimental results on our collected dataset to demonstrate the effectiveness of our model. We first introduce dataset construction and then show feature extraction. Finally, we give results and analysis.

TABLE I
THE STATISTICS OF OUR SOCIAL EVENT DATASET.

Event ID	Event Name	# of Documents
1	the dispute of the South China Sea	3600
2	cannabis legalization in the U.S.	4118
3	2011 England riots	4563
4	Japan earthquake and tsunami	5670
5	Putin's return	4017
6	Death of Michael Jackson	5385
7	Greek protests	5698
8	Mars Reconnaissance Orbiter	5762
9	the 2011 Norway attacks	5546
10	Syrian civil war	5812
11	the development of Microsoft Windows	4454
12	Apple Jobs	4908

A. Dataset Collection

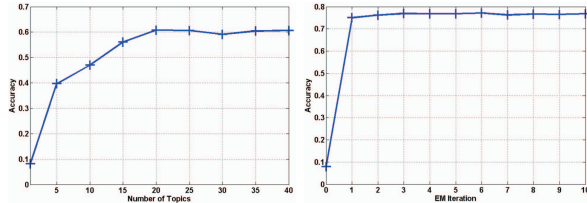
To our best knowledge, there is no multi-modality social event dataset available for classification. Therefore, we collect the dataset by ourselves from the photo-sharing website Flickr. The dataset contains 12 different social events happened in the past few years as shown in Table I. For each social event, we use keywords and the site's public API to crawl related images and text information. Each image and the associated text information (tags, title, and description) are considered as a social event document. The collected 12 social events cover a wide range of topics including politics, economics, entertainment, military, society, and so on. For each social event, there are about 3000 to 6000 documents, and totally, there are about 59,500 social media documents on this dataset. When we do our experiments, 50% documents of each social event category are used for training and the rest for test.

B. Feature Extraction

For textual description, we use stemming method and stop words elimination and remove words with a corpus frequency less than 15, and there are 35,565 unique words left. For visual description, the words are based on image patches, which are obtained by SLIC segmentation method [18]. To obtain the description for image patch, we densely sample SIFT points, and adopt the popular sparse coding based method [19], [20] to encode each SIFT point. Then, based on the feature codes of all SIFT points of each patch, we adopt max pooling to obtain its description. Once obtaining all image patch descriptions, we adopt K-means to build a codebook (5000 words). By hard assignment coding of each patch, each image can be described as the counts of the words in the codebook.

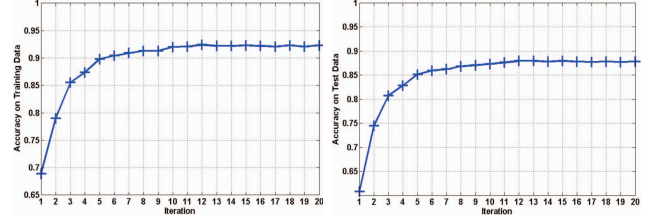
C. Results and Analysis

In this Section, we show more results and analysis. In the subsection V-C1, we show the parameter analysis of our



(a) Accuracy vs number of topics. (b) The iteration process of EM.

Fig. 2. The parameter analysis of our proposed topic model mm-SLDA.



(a) On training data. (b) On test data.

Fig. 3. The accuracy with the boosting procedure of our BMM-SLDA.






model. In V-C2, we give the qualitative evaluation of the mined visual and textual topics. In the subsection V-C3, we show the quantitative results compared with the existing methods.

1) *Parameter Analysis: Topic Number K* : In topic modeling, how to set the topic number K is not trivial. We evaluate the accuracy of our proposed BMM-SLDA with different numbers of topics. For simplicity, we evaluate this parameter at the first iteration of boosting and fix it for other iterations. The result is shown in Fig.2(a). From the Fig.2(a), we can see that the number of topics is changed from 1 to 40 and the accuracy of our model is quite stable when the number of topics is changed from 20 to 40. Note that the value of K depends on the social event dataset. In our dataset, we set K as 20, which achieves the best at the first iteration of boosting.

EM Convergence: We adopt EM algorithm to solve our mm-SLDA model, and it is important to guarantee its convergence. The iteration process of our optimization is shown in Fig.2(b). From the Fig.2(b), we can see that the performance of our mm-SLDA increases quickly during the first 3 iterations and tends to converge after that. The result shows that EM algorithm can guarantee the convergence of our mm-SLDA.

Boosting Convergence: In Fig.3, we show that the classification accuracies increase with the iteration of boosting procedure on training data and test data, respectively. From the results, we can see that, when we iteratively learn more and more useful topic models, the accuracies can be improved. Based on the iteration process as shown in Fig.3, we can see that our BMM-SLDA can converge well and the reasonably accurate solutions are available after 12 iterations.

2) *Qualitative Evaluation:* We demonstrate the effectiveness of BMM-SLDA model on the mining of social events and show the discovered topics in Fig.4. Due to the limited space, we only visualize the topics mined in the first iteration of our boosting procedure. Here, we show 2 of the discovered 20 topics with their top five textual words and the five most related images, respectively. In text and image visualization, the textual words are sorted by the probability $p(w|z)$, while the images are sorted by counting the number of visual descriptors and textual words with the corresponding topic in different event documents $p(z_k|w_d, v_d) = (n_{d,k}^v + n_{d,k}^w) / \sum_{k=1}^K (n_{d,k}^v + n_{d,k}^w)$, where $n_{d,k}^v = \sum_i 1(z_{d,i}^v = k)$ and $n_{d,k}^w = \sum_i 1(z_{d,i}^w = k)$ represent the numbers of topic k assigned to the textual words and the visual descriptors of event document d , respectively. As can be seen, the results are impressive and satisfy our expectation, where each extracted event topic is meaningful and textual

Topic #4				
protest	Greek	demonstrator	solider	police
0.09474	0.05813	0.02705	0.01131	0.00932
				
0.60352	0.59714	0.58071	0.56512	0.48159






Topic #15				
riot	England	police	Tottenham	conflict
0.07143	0.04702	0.03056	0.01879	0.00789
				
0.47541	0.42165	0.41576	0.41033	0.40857

Fig. 4. The discovered topics by BMM-SLDA. See text for more details.

words are well aligned with the corresponding visual image content. Based on the results, we can confirm that our proposed model can effectively mine the topics of social events.

3) *Quantitative Evaluation:* We compare our models (mm-SLDA and BMM-SLDA) with 4 baseline methods (mmLDA+SG, mmLDA+SVM, SLDA(Visual), and SLDA(Text)), which are the most related to our work. For the two standard classification algorithms mmLDA+SG and mmLDA+SVM, we firstly use the mmLDA model, an unsupervised model, to represent each event document to a K -dimensional vector using textual and visual information in train dataset. Then, we train classifier using softmax regression method and Support Vector Machine (SVM), respectively. Finally, we use the trained model to predict class labels of test data. The SLDA(Visual) and SLDA(Text) [13] are the supervised model with only visual feature and textual feature, respectively. The mm-SLDA is our proposed supervised model trained with all train data, which ignores the document weight distribution compared with the BMM-SLDA.

The quantitative results are shown in Table II. Because the dataset is quite difficult, no methods can achieve 100% accuracy performance. SLDA (Text) is better than mmLDA (SG), which shows supervised information is useful. SLDA (Text) is better than SLDA (Visual), which shows the textual information is much more helpful than the visual information for social event classification. This can be explained that the images are very diverse. From the results, we can see that our model can outperform all other four models. This is because mmLDA only adopts the multi-modality information and SLDA only uses the supervised information. Different from these methods, our mm-SLDA can exploit the multi-modal property and the multi-class property jointly for social event modeling and boost the classification performance. Compared with mm-SLDA, BMM-SLDA is much more efficient and effective. Due to the boosting weighted sampling strategy, our BMM-SLDA works on a small part of the data and is much more efficient. Therefore, the model can work on large-scale dataset. In our experiments, at each iteration of the boosting procedure, the numbers of documents from each category based on their weight distribution are empirically set to 500. Moreover, the BMM-SLDA effectively models the document weight distribution with classification error, and iteratively

TABLE II
THE ACCURACY COMPARED WITH OTHER EXISTING METHODS.

Methods	Accuracy	Methods	Accuracy
mmLDA+SG	0.671	mmLDA+SVM	0.715
SLDA(Visual)	0.401	SLDA(Text)	0.717
mm-SLDA	0.766	BMM-SLDA	0.877

learn new topic models to correct the previously misclassified social event documents. As a result, our BMM-SLDA considers the data structure and can boost the performance. Based on the results, we can confirm that it is important to iteratively learn multiple topic models and weak classifiers with a small sampled subset data by considering the document weight distribution for event modeling.

ACKNOWLEDGMENT

This work is supported in part by the National Program on Key Basic Research Project (973 Program, Project No. 2012CB316304), and National Natural Science Foundation of China (61225009, 61303173), also by the Singapore National Research Foundation under the IDM Programme Office.

REFERENCES

- [1] L. Xueliang and H. Benoit, "Heterogeneous features and model selection for event-based media classification," in *ICMR*, 2013.
- [2] T. Zhang and C. Xu, "Cross-domain multi-event tracking via co-pmht," in *ACM Transactions on Multimedia Computing, Communications and Applications*, 2014.
- [3] K. Radinsky and E. Horvitz, "Mining the web to predict future events," in *WSDM*, 2013.
- [4] G. Kumaran and J. Allan, "Text classification and named entities for new event detection," in *SIGIR*, 2004.
- [5] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine, "Diamonds in the rough: Social media visual analytics for journalistic inquiry," in *IEEE VAST*, 2010, pp. 115–122.
- [6] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi, "Simple semantics in topic detection and tracking," *Inf. Retr.*, vol. 7, no. 3-4, pp. 347–368, 2004.
- [7] C. S. Firan, M. Georgescu, W. Nejdl, and R. Paiu, "Bringing order to your photos: event-driven classification of flickr images based on social knowledge," in *CIKM*, 2010.
- [8] J. R. Kender and M. R. Naphade, "Visual concepts for news story tracking: Analyzing and exploiting the nist trecvid video annotation experiment," in *CVPR*, 2005.
- [9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *JMLR*, vol. 3, pp. 993–1022, Mar. 2003.
- [10] B. Yang, C. Nigell, and D. Anindya, "A partially supervised cross-collection topic model for cross-domain text classification," in *CIKM*, 2013.
- [11] D. Blei and J. McAuliffe, "Supervised topic models," in *NIPS*, 2008.
- [12] S. Jitao and X. Changsheng, "Right buddy makes the difference: An early exploration of social relation analysis in multimedia applications," in *ACM MM*, 2012.
- [13] C. Wang, D. M. Blei, and F.-F. Li, "Simultaneous image classification and annotation," in *CVPR*, 2009.
- [14] D. Putthividhy, H. Attias, and S. Nagarajan, "Topic regression multi-modal latent dirichlet allocation for image annotation," in *CVPR*, 2010.
- [15] J. Zhu, H. Zou, S. Rosset, and T. Hastie, "Multi-class adaboost," 2009.
- [16] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5228–5235, 2004.
- [17] C.-J. Lin, R. C. Weng, and S. S. Keerthi, "Trust region newton method for logistic regression," *JMLR*, vol. 9, pp. 627–650, 2008.
- [18] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels," 2010, in Technical report, EPFL.
- [19] L. Liu, L. Wang, and X. Liu, "In defense of softassignment coding," 2011, in *ICCV*.
- [20] T. Zhang, B. Ghanem, S. Liu, C. Xu, and N. Ahuja, "Low-rank sparse coding for image classification," in *ICCV*, 2013.