Multi-Modal Supervised Latent Dirichlet Allocation for Event Classification in Social Media

Shengsheng Qian
National Lab of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
Beijing, China
shengsheng.qian@nlpr.ia.ac.cn

Tianzhu Zhang
National Lab of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
Beijing, China
tzzhang@nlpr.ia.ac.cn

Changsheng Xu
National Lab of Pattern Recognition
Institute of Automation
Chinese Academy of Sciences
Beijing, China
csxu@nlpr.ia.ac.cn

ABSTRACT

In social media, many existing websites (e.g., Flickr, YouTube, and Facebook) are for users to share their own interests and opinions of many popular events, and successfully facilitate the event generation, sharing and propagation. As a result, there are substantial amounts of user-contributed media data (e.g., images, videos, and textual content) for a wide variety of real-world events of different types and scales. The aim of this paper is to automatically identify the interesting events from massive social media data, which are useful to browse, search and monitor social events by users or governments. To achieve this goal, we propose a novel multi-modal supervised latent dirichlet allocation (mm-SLDA) for social event classification. Our proposed mm-SLDA has a number of advantages. (1) It can effectively exploit the multi-modality and the multi-class property of social events jointly. (2) It makes use of the supervised social event category label information and is able to classify multi-class social event directly. We evaluate our proposed mm-SLDA on a real world dataset and show extensive experimental results, which demonstrate that our model outperforms state-of-the-art methods.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.4.10 [**Image Processing and Computer Vision**]: Image Representation; I.5.4 [**Pattern Recognition**]: Applications—*Computer Vision*

General Terms

Algorithms, Experimentation, Performance

Keywords

Event Classification, Multi-Modal, Supervised LDA, Social Media.

ICIMCS'14, July. 10–12, 2014, Xiamen, Fujian, China

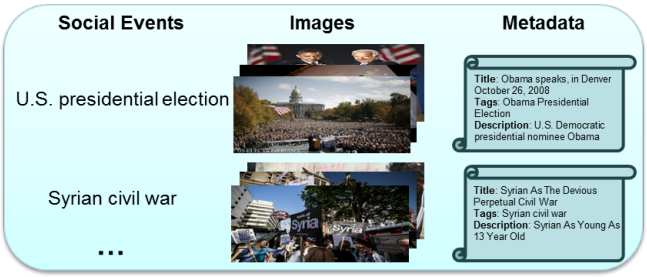

Figure 1: Social events in Flickr have multi-modal property (e.g., images, title, tags, description) and multi-class property (e.g., Syrian civil war, US presidential election).

1. INTRODUCTION

With the rapid development of Internet, more and more social media sites (e.g., Flickr, YouTube, and Facebook) spring up and vast amounts of social events with overwhelming metadata (comments, tags, titles, etc.) and various images generated by users are being shared in these popular sites. Most of these social events uploaded by users are connected with some specific topics, and it is time-consuming to manually cluster or identify them. Therefore, automatically mining and summarizing hotspot topics of 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 there are substantial amounts of events with multi-modal property (e.g., images, videos, and textual content) and multi-class property (different categories, e.g., Syrian civil war, Occupy wall street) on the Internet. An example is shown in Fig.1.

Recently, how to address these challenges for automatically mining and monitoring social events has drawn much attention in the multimedia research community, such as social event classification [17], social event tracking [20], social event mining [14], and investigating event detection [8, 15]. Most of the existing work focuses on feature design for social event modeling. In [4, 11], the textual features are adopted. While social media events have rich visual information, such as, images and videos, which are helpful for

the analysis and mining of social events. For the same social events, they may have different textual descriptions (comments, tags, etc.) due to different users, but their visual information may be similar. Therefore, multi-modal feature fusion becomes more and more popular for social event analysis. In [17], the similarity between different social events is conducted based on time, location or text features. In above work, many multi-modal features, such as tag, time, location or visual features are exploited, and as a result encouraging performance is achieved. However, these methods ignore the multi-class property of social events, and do not exploit the discriminative category information to improve the event classification performance.

Different from the previous work, we attempt to exploit the multi-modal and multi-class property jointly for social event modeling. 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. In Fig.1, we show an example of two different social events with multi-modality including user-provided metadata (e.g., title and tags) and images. Here, each image and its corresponding metadata are considered as one social media document. 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. To make this come true, 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 utilize the supervised category labels, which are able to boost the classification performance.

Inspired by the above discussions, we propose a multi-modal supervised latent dirichlet allocation (mm-SLDA) for social event classification. Firstly, the proposed model can capture the visual and textual topics across multi-modal social event data jointly. Secondly, the proposed model can directly utilize the supervised event category information for social event classification modeling, which is implemented by considering the class label response to variable drawn from a softmax regression function. Finally, the proposed model is solved by an EM iterative learning algorithm where we seek to optimize the marginal distribution in the E-step and the conditional distribution in the M-step, respectively. Compared with existing methods, the contributions are threefold.

- We propose a novel multi-modal supervised latent dirichlet allocation (mm-SLDA) topic model for social event classification, which can effectively exploit the multi-modal property and the multi-class property of social event jointly.
- To make the proposed topic model suitable for multi-class social event classification, we adopt an effective softmax regression method, which can classify multi-class social event directly.
- We collect a dataset for research on social event classification with multi-modality information, and will release it for academic use. We evaluate our proposed model and demonstrate that it achieves much better performance than existing methods.

The rest of the paper is organized as follows. In Section 2, the related work is reviewed. The proposed mm-SLDA and its optimization process are presented in Section 3. In Section 4, we report and analyze extensive experimental results. Finally, we conclude the paper with future work in Section 5.

2. RELATED WORK

We review previous methods which are most related to our work. Due to the limited space, we provide a brief overview of event classification methods and existing topic models.

Event Classification: With the massive growth of social events in Internet, efficient organization and monitoring of social events becomes a challenge. To address this problem, many researchers are working on social event analysis [17, 5, 11, 13]. Existing approaches on event analysis are based on single-modality (e.g., text, images) information or multi-modality information. For the single-modality analysis, there are many existing methods using textual information (e.g., names, time references, locations, title, tags, and description) or visual information (e.g., images and videos) [4, 11] to model social event. However, the single-modality based methods ignore the multi-modal property of social event and cannot outperform the multi-modality based methods generally. To deal with this problem, many researchers investigate the different features between events and social media data and adopt multiple features to calculate the similarity of the social documents, such as time, tag and location feature [17, 5]. While the above methods focus on feature design to improve experimental performance, the importance and effectiveness of those features have not been studied in details. Different from the above existing methods, we sufficiently exploit 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 event classification algorithm, which is able to model multi-modality and supervised category label information jointly.

Topic Model: There are many topic models based on Latent Dirichlet Allocation (LDA) [3], such as supervised Latent Dirichlet Allocation (SLDA) [18], multi-modal Latent Dirichlet Allocation (mmLDA) [7]. In [18], a supervised topic model called partially supervised cross-collection LDA is proposed for cross-domain learning in a unified way. However, traditional LDA and SLDA [2] 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 [16] and they do not consider multi-modal corpora. In the multi-modal topic model [7], the authors consider multi-modal information, such as users' textual annotations and visual images, and propose a multi-modal topic-sensitive inference model for social relation mining. Our proposed topic model is different from the previous models. Compared with [7], our proposed model focuses on social event classification. Furthermore, the traditional multi-modal LDA [12] does not utilize the supervised category labels. We extend multi-modal LDA to a supervised topic model with softmax regression function, which is used because the social events have multi-class property and can be classified into multiple classes directly.

3. THE PROPOSED ALGORITHM

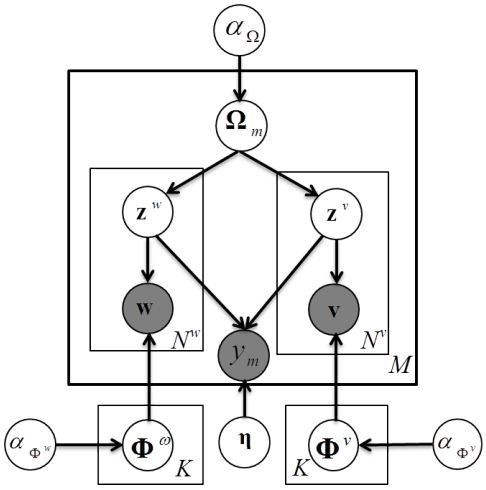

Figure 2: The proposed multi-modal supervised Latent Dirichlet Allocation topic model for social event classification. For details, please refer to the corresponding text.

In this section, we first formally define our problem of 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.

3.1 Problem Definition

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 problem as a classification problem over social media documents (e.g., images, title, description, tags). Let $E = \{(e_1, y_1), (e_2, y_2), \ldots, (e_M, y_M)\}$ denote a training dataset of M examples, where $e_m = [\mathbf{v}_m, \mathbf{w}_m]$ is the m^{th} image-text pair and $y_m \in \{1, 2, \ldots, C\}$ is the class label of the sample e_m. Here, C is the number of event class labels. The goal is to generate a learner to classify a new image-text pair based on its instance $\mathbf{v}$ and text information $\mathbf{w}$. To achieve this goal, we propose a novel multi-modal supervised LDA. The details are introduced in the next section.

3.2 Our Model

This section formalizes the problem of social event classification under the 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. 2. From the figure, 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 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 $k \in \{1, 2, \ldots, K\}$, Draw $\Phi^v | \alpha_{\Phi^v} \sim \text{Dir}(\alpha_{\Phi^v})$
2. For each textual topic $k \in \{1, 2, \ldots, K\}$, 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, $n \in \{1, 2, ..., N^v\}$
   (a) Draw a topic assignment $\mathrm{z}_n^v | \Omega_m \sim \text{Mult}(\Omega_m)$
   (b) Draw a visual patch $v_n | z_n^v, \Phi^v \sim \text{Mult}(\Phi_{z_n}^v)$
5. For each textual word w_n, $n \in \{1, 2, ..., N^w\}$
   (a) Draw a topic assignment $\mathrm{z}_n^w | \Omega_m \sim \text{Mult}(\Omega_m)$
   (b) Draw a word $w_n | z_n^w, \Phi^w \sim \text{Mult}(\Phi_{z_n}^w)$
6. Draw class label $y_m | z_m \sim \text{softmax}(\overline{z}_m, \eta)$

Here, the softmax function provides the following distribution $p(y_m | \overline{z}_m, \eta) = \exp(\eta_{y_m}^T \overline{z}_m) / \sum_{l=1}^{C} \exp(\eta_l^T \overline{z}_m)$, where K-dimensional vectors $\eta_{1:C}$ and $\overline{z}_m$ 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.

3.3 Parameter Inference

In our mm-SLDA, the posterior joint probability can be factorized as:

$$
\begin{aligned}
&p(\mathbf{z}^w, \mathbf{z}^v, \mathbf{w}, \mathbf{v}, \mathbf{y} | \alpha_\Omega, \alpha_{\Phi^w}, \alpha_{\Phi^v}, \eta) \\
&\propto p(\mathbf{z}^w, \mathbf{z}^v | \alpha_\Omega) \times p(\mathbf{w} | \mathbf{z}^w, \alpha_{\Phi^w}) \times p(\mathbf{v} | \mathbf{z}^v, \alpha_{\Phi^v}) \times p(\mathbf{y} | \overline{\mathbf{z}}, \eta) \\
&\propto \int p(\mathbf{z}^w | \boldsymbol{\Omega}) p(\mathbf{z}^v | \boldsymbol{\Omega}) p(\boldsymbol{\Omega} | \alpha_\Omega) d\boldsymbol{\Omega} \int p(\mathbf{w} | \mathbf{z}^w, \boldsymbol{\Phi}^w) p(\boldsymbol{\Phi}^w | \alpha_{\Phi^w}) d\boldsymbol{\Phi}^w \\
&\quad \int p(\mathbf{v} | \mathbf{z}^v, \boldsymbol{\Phi}^v) p(\boldsymbol{\Phi}^v | \alpha_{\Phi^v}) d\boldsymbol{\Phi}^v \times p(\mathbf{y} | \overline{\mathbf{z}}, \eta) \\
&\propto \prod_{m=1}^{M} \frac{B(\mathbf{n}_{m,\cdot}^w + \mathbf{n}_{m,\cdot}^v + \alpha_\Omega)}{B(\alpha_\Omega)} \prod_k \frac{B(\mathbf{n}_{\cdot,k}^w + \alpha_{\Phi^w})}{B(\alpha_{\Phi^w})} \prod_k \frac{B(\mathbf{n}_{\cdot,k}^v + \alpha_{\Phi^v})}{B(\alpha_{\Phi^v})} \\
&\prod_{m=1}^{M} \prod_{l=1}^{C} \{\exp(\eta_l^T \overline{\mathbf{z}}_m) / \sum_{j=1}^{C} \exp(\eta_j^T \overline{\mathbf{z}}_m)\}^{1\{y^{(m)}=l\}}, \qquad (1)
\end{aligned}
$$

where $B(\alpha) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma(\sum_k \alpha_k)}$ is a normalizing constant, $\mathbf{n}_{m,\cdot}^w =< n_{m,1}^w, \ldots, n_{m,k}^w, \ldots, n_{m,K}^w >$ and $\mathbf{n}_{m,\cdot}^v =< n_{m,1}^v, \ldots, n_{m,k}^v, \ldots, n_{m,K}^v >$, where $n_{m,k}^w = \sum_i 1(z_{m,i}^w = k)$ and $n_{m,k}^v = \sum_i 1(z_{m,i}^v = k)$ are the numbers of topic k assigned to the textual information and visual information of event document m respectively. Similarly, $\mathbf{n}_{\cdot,k}^w =< n_{1,k}^w, \ldots, n_{i,k}^w, \ldots, n_{D_w,K}^w >$ and $\mathbf{n}_{\cdot,k}^v =< n_{1,k}^v, \ldots, n_{i,k}^v, \ldots, n_{D_v,K}^v >$, where $n_{i,k}^w$ and $n_{i,k}^v$ are the numbers of word i assigned to textual topic Φ_k^w and visual topic

Φ_k^v in all event documents, respectively. $1\{.\}$ is the indicator function, so that $1\{\text{a true statement}\} = 1$, and $1\{\text{a false statement}\} = 0$.

Exact inference is often intractable in many topic models and appropriate methods must be used, such as variational inference [3]and Gibbs sampling [6]. 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

$$\begin{aligned} &p(z_{m,i}^w = k|\mathbf{z}_{\neg(m,i),w}, \mathbf{w}, \mathbf{v}, \mathbf{y}, \alpha_\Omega, \alpha_{\Phi^w}, \alpha_{\Phi^v}, \eta) \\ &\propto \frac{(n_{m,k}^{\neg(i)} + \alpha_\Omega)}{\sum_{k=1}^K (n_{m,k} + \alpha_\Omega) - 1} \frac{n_{t,k}^{\neg(m,i),w} + \alpha_{\Phi^w}}{\sum_{p=1}^{D_w} (n_{p,k}^w + \alpha_{\Phi^w}) - 1} \\ &\prod_{l=1}^C \{\exp(\eta_l^T \overline{\mathbf{z}}) / \sum_{j=1}^C \exp(\eta_j^T \overline{\mathbf{z}})\}^{1\{y^{(m)}=l\}} \end{aligned} \tag{2}$$

Here, $\mathbf{z}_{\neg(m,i),w}$ denotes the vectors of topic assignment except the considered word at position i in the textual information w of event document m, $n_{t,k}^{\neg(m,i),w}$ denotes the number of times of word t 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, $n_{m,k}^{\neg(i)}$ denotes the number of text words and image patches in event document m assigned to topic k except the current assignment, $\sum_{k=1}^K (n_{m,k} + \alpha_\Omega) - 1$ 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, and the conditional posterior distribution of the latent topic indicators is:

$$\begin{aligned} &p(z_{m,i}^v = k|\mathbf{z}_{\neg(m,i),v}, \mathbf{w}, \mathbf{v}, \mathbf{y}, \alpha_\Omega, \alpha_{\Phi^w}, \alpha_{\Phi^v}, \eta) \\ &\propto \frac{(n_{m,k}^{\neg(i)} + \alpha_\Omega)}{\sum_{k=1}^K (n_{m,k} + \alpha_\Omega) - 1} \frac{n_{t,k}^{\neg(m,i),v} + \alpha_{\Phi^v}}{\sum_{p=1}^{D_v} (n_{p,k}^v + \alpha_{\Phi^v}) - 1} \\ &\prod_{l=1}^C \{\exp(\eta_l^T \overline{\mathbf{z}}) / \sum_{j=1}^C \exp(\eta_j^T \overline{\mathbf{z}})\}^{1\{y^{(m)}=l\}} \end{aligned} \tag{3}$$

After finishing Gibbs sampling, we can estimate Φ^w, Φ^v, Ω as [6].

$$\Phi_{k,t}^w = \frac{n_{t,k}^w + \alpha_{\Phi^w}}{\sum_{p=1}^{D_w} (n_{p,k}^w + \alpha_{\Phi^w})} \tag{4}$$

$$\Phi_{k,t}^v = \frac{n_{t,k}^v + \alpha_{\Phi^v}}{\sum_{p=1}^{D_v} (n_{p,k}^v + \alpha_{\Phi^v})} \tag{5}$$

$$\Omega_{m,k} = \frac{(n_{m,k} + \alpha_\Omega)}{\sum_{k=1}^K (n_{m,k} + \alpha_\Omega)} \tag{6}$$

M-step: In the M-step, we update the class coefficients η by maximizing the joint likelihood in Eq.(1). Because we fix parameters obtained in the E-step, it is equivalent to maximizing $p(\mathbf{y}|\overline{\mathbf{z}}, \eta)$ where each event document m is represented by $\overline{\mathbf{z}}$ newly updated in the E-step. Specifically, we learn L_2-regularized softmax regression model which solves the following unconstrained optimization problem:

$$\min_\eta (-\frac{1}{M} \sum_{m=1}^M \sum_{c=1}^C 1\{y^{(m)} = c\} \log \frac{e^{\eta_c^T \overline{\mathbf{z}}_m}}{\sum_{l=1}^C e^{\eta_l^T \overline{\mathbf{z}}_m}} + \frac{\lambda}{2} \sum_{i=1}^C \eta_i^T \eta) \tag{7}$$

where λ is a regularization parameter and is set to be 1.0, and we apply a trust region Newton method [9] for optimization.

3.4 Social Event Classification Model

After the E-step and the M-step, we obtain the updated parameters Φ^w, Φ^v and η, and can predict the label of a new event document. Given a new social event document e_{new} which is composed of many text information 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 $\overline{\mathbf{z}}_{new}$ and adopt class coefficients η for prediction. Specifically, we predict the class label of a new event document e_{new} according to Eq.(8).

$$\begin{aligned} y_{e_{new}} &= \arg\max_{y_{e_{new}} \in \{1,2\cdots,C\}} (p(y_{e_{new}} = c|\overline{\mathbf{z}}_{new}, \eta)) \\ &= \arg\max_{y_{e_{new}} \in \{1,2\cdots,C\}} (\eta_c^T \overline{\mathbf{z}}_{new}) \end{aligned} \tag{8}$$

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

4. EXPERIMENTAL RESULTS

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

4.1 Dataset Collection

To the best of our knowledge, there is no multi-modality social event dataset available for classification in the multimedia community. Therefore, we collect the dataset by ourselves from the photo-sharing website Flickr. The dataset contains 10 different social events happened in the past few years as shown in Table 1. 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 10 social events cover a wide range of topics including politics, economics, entertainment, military, society, and so on. For each social event, there are about 2500 to 5000 documents, and totally, there are about $36,000$ social media documents on this dataset. When we do our experiments, 60% of each social event documents are used for training and the rest for test.

4.2 Feature Extraction

For textual description, we use stemming method and stop words elimination and remove words with a corpus frequency

Event ID	Event Name	# of Documents
1	Senkaku Islands Dispute	2764
2	Occupy Wall Street	5503
3	U.S. presidential election	4219
4	the War in Afghanistan	3341
5	Gangnam Style	2990
6	North Korea nuclear program	3346
7	Greek protests	3683
8	Mars Reconnaissance Orbiter	2482
9	the 2011 Norway attacks	4001
10	Syrian civil war	4179

Table 1: The statistics of our collected social event dataset.

less than 10. There are 22913 unique words finally. For visual description, the words are based on image patches, which are obtained by SLIC segmentation method [1]. To obtain the description for image patch, we densely sample SIFT points, and adopt the popular sparse coding based method [10, 19] 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.

4.3 Parameters Turning

In topic modeling, the selection of topic number K is not trivial. In Fig.3, we show that how the classification accuracy is changed with the number of topics. In our experiments, the number of topics is changed from 1 to 50. From the Fig.3, we can see that the accuracy of our proposed mm-SLDA is quite stable when the number of topics is changed from 15 to 50. Note that the value of K depends on the social event dataset. In our dataset, we set K as 20, which achieves the best.

Our mm-SLDA is solved by EM algorithm, and it is important to guarantee the convergence of the optimization. The iteration process of our optimization is shown in Fig.4. From the Fig.4, we can see that the performance of our mm-SLDA increases quickly during the first 5 iterations and tends to converge after that. The result shows that our model can ensure the convergence.

4.4 Results and Analysis

In this Section, we show more results and analysis. In the subsection 4.4.1, we give the qualitative evaluation of the mined discriminative visual and textual topics. In the subsection 4.4.2, we show the quantitative results compared with the existing methods.

4.4.1 *Qualitative Evaluation*

Due to the limited space, in Fig.5, we only visualize 2 of

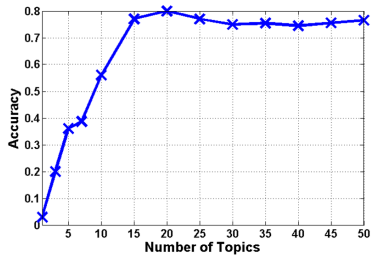

Figure 3: The accuracy of classification vs the number of topics.

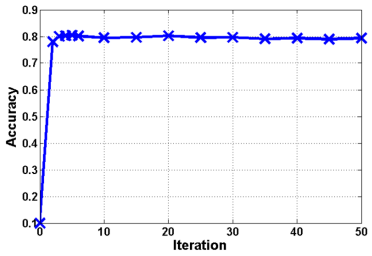

Figure 4: The iteration process of our optimization.

Figure 5: Illustration of discovered topics by mm-SLDA. More details please see the text.

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 $p(z_k|w_d, v_d)$:

$$p(z_k|w_d, v_d) = \frac{n^v_{d,k} + n^w_{d,k}}{\sum_{k=1}^{K}(n^v_{d,k} + n^w_{d,k})}$$

where $n^v_{d,k} = \sum_i 1(z^v_{d,i} = k)$ and $n^w_{d,k} = \sum_i 1(z^w_{d,i} = k)$ represent the numbers of topic k assigned to the textual words and the visual descriptors of event document d, respectively.

By providing a multi-modal information of the representative textual and visual words, it is very intuitive to interpret the social events with each associated topic. As can be seen, the results are impressive and satisfy our expectation, where each extracted event topic is meaningful and textual words are well aligned with the corresponding visual image content. Based on the results, we can confirm that our proposed mm-SLDA can effectively mine the topics of social events.

4.4.2 *Quantitative Evaluation*

We compare our approach 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, firstly, we use the mmLDA model, an unsupervised model, to represent each event document to a K-dimensional vector using textual and visual information. Then, we train classifier using softmax regression method

Methods	Accuracy
mmLDA+SG	0.699
mmLDA+SVM	0.755
SLDA(Visual)	0.359
SLDA(Text)	0.758
mm-SLDA	**0.803**

Table 2: The social event classification accuracy compared with other existing methods.

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)[16] are the supervised model with only visual feature and textual feature, respectively.

The quantitative results are shown in Table 2. 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.

5. CONCLUSIONS

In this paper, we have presented a multi-modal supervised latent dirichlet allocation for event classification in social media. Our proposed model can exploit the multi-modal property and the multi-class property of social event jointly, and can predict the label of a new social event document directly. We have conducted experiments on our collected dataset and extensive results have demonstrated that our model outperforms all other existing models. For the future work, we will extend our model to large-scale event data with more classes.

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 International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

6. REFERENCES

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels, 2010. In Technical report, EPFL.

[2] D. Blei and J. McAuliffe. Supervised topic models. In *NIPS*. 2008.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, Mar. 2003.

[4] N. Diakopoulos, M. Naaman, and F. Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *IEEE VAST*, pages 115–122, 2010.

[5] 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.

[6] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101:5228–5235, 2004.

[7] S. Jitao and X. Changsheng. Right buddy makes the difference: An early exploration of social relation analysis in multimedia applications. In *ACM MM*, 2012.

[8] G. Kumaran and J. Allan. Text classification and named entities for new event detection. In *SIGIR*, 2004.

[9] C.-J. Lin, R. C. Weng, and S. S. Keerthi. Trust region newton method for logistic regression. *JMLR*, 9:627–650, 2008.

[10] L. Liu, L. Wang, and X. Liu. In defense of softassignment coding, 2011. In ICCV.

[11] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi. Simple semantics in topic detection and tracking. *Inf. Retr.*, 7(3-4):347–368, 2004.

[12] D. Putthividhy, H. Attias, and S. Nagarajan. Topic regression multi-modal latent dirichlet allocation for image annotation. In *CVPR*, 2010.

[13] S. Qian, T. Zhang, and C. Xu. Boosted multi-modal supervised latent dirichlet allocation for social event classification. In *ICPR*, 2014.

[14] K. Radinsky and E. Horvitz. Mining the web to predict future events. In *WSDM*, 2013.

[15] Z. Tianzhu, X. Changsheng, Z. Guangyu, L. Si, and L. Hanqing. A generic framework for video annotation via semi-supervised learning. *IEEE Transactions on Multimedia*, 14(4):1206–1219, 2012.

[16] C. Wang, D. M. Blei, and F.-F. Li. Simultaneous image classification and annotation. In *CVPR*, 2009.

[17] L. Xueliang and H. Benoit. Heterogeneous features and model selection for event-based media classification. In *ICMR*, 2013.

[18] B. Yang, C. Nigel, and D. Anindya. A partially supervised cross-collection topic model for cross-domain text classification. In *CIKM*, 2013.

[19] T. Zhang, B. Ghanem, S. Liu, C. Xu, and N. Ahuja. Low-rank sparse coding for image classification. In *ICCV*, 2013.

[20] T. Zhang and C. Xu. Cross-domain multi-event tracking via co-pmht. In *ACM Transactions on Multimedia Computing, Communications and Applications*, 2014.