Decision Support Systems xxx (xxxx) xxx

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

A novel label-based multimodal topic model for social media analysis

Hao Li [a], Yang Qian [a,b,*], Yuanchun Jiang [a,c], Yezheng Liu [a,d], Fan Zhou [a]

[a] School of Management, Hefei University of Technology, Hefei, Anhui 230009, China
[b] Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei, Anhui 230009, China
[c] Key Laboratory of Philosophy and Social Sciences for Cyberspace Behaviour and Management, Anhui Province
[d] National Engineering Laboratory for Big Data Distribution and Exchange Technologies, Shanghai 200436, China

ARTICLE INFO

Keywords:
Multimodal data
Topic modeling
Label data
Supervised model
Image representation

ABSTRACT

Extracting useful knowledge from multimodal data is the core of many multimedia applications, such as recommendation systems, and cross-modal retrieval. In this paper, we propose a label-based multimodal topic (LB-MMT) model to jointly model text and image data tagged with multiple labels. Specifically, we use the labels as supervised information to generate the text and image data. In the LB-MMT model, we assume that the textual words and visual words related to each text and image are drawn from a mixture of latent topics, where each topic is represented as a group of textual words and visual words. Moreover, we introduce multiple topics for each label, to build the top-down relationship from label to text and image. To investigate the effectiveness of the proposed approach, we conduct extensive experiments on a real-world multimodal dataset with labels. The results show the proposed approach obtains superior performances on topic coherence and label prediction compared with previous competitors. In addition, we show that our model yields interesting insights about multimodal topics. The proposed model provides important practical implications, e.g., designing more attractive multimodal contents for marketers.

1. Introduction

With the development of information technology, multimodal data in online social networks (e.g., Facebook) and e-commerce (e.g., Amazon) platforms, is becoming more prevalent, such as texts, images, and video clips. Mining valuable information from these multimodal data plays an important role in many applications, e.g., recommendation systems [1,2], cross-modal information retrieval [3], and online advertising [4].

Texts and their associated images are typical multimodal data. Recently, many previous studies have focused on modeling text and image modalities, simultaneously. Most of these works mainly can be divided into two categories. The first one based on deep learning models attempts to learn the joint representations of texts and images for tasks of image annotation and tag recommendation, etc. [5–8]. The second category of work based on probabilistic topic models tends to extract the interpretable information by modeling the semantic correlations between textual contents and visual features of images [9]. For instance, Blei and Jordan [10] first proposed a multimodal topic model, namely Corr-LDA, to learn the relationship between images and text modalities. Despite the success of the existing studies, one limitation is that they ignore some other important information related to the texts and images, i.e., labels.

On many online platforms, the texts and their associated images are often labeled with multiple human-offered tags. For example, in Taobao.com (see Fig. 1), marketers often use textual descriptions and product images to promote products and also customize some tags or keywords related to these contents. Note that the label set conveys the key information and reveals a summary of the semantic content of texts and images. In terms of platforms, the label set also plays an important role in helping users quickly retrieve the information. We conjecture that considering the label information could greatly enhance the performance for understanding the textual contents and images.

In our paper, our goal is to extend multimodal topic models to detect valuable and interpretable topics by jointly leveraging the texts and associated images with labels. Due to the nature of multi-modal data, there are several challenges we have to solve. The first challenge is how to characterize the top-down relationship from label to text and image. Intuitively, the label set on a webpage (see Fig. 1) denotes higher-level information that can clearly account for the words and images (lower-

* Corresponding author at: School of Management, Hefei University of Technology, Hefei, Anhui 230009, China.
E-mail addresses: handsomeli@mail.hfut.edu.cn (H. Li), soberqian@hfut.edu.cn (Y. Qian), ycjiang@hfut.edu.cn (Y. Jiang), liuyezheng@hfut.edu.cn (Y. Liu), zhoufan@mail.hfut.edu.cn (F. Zhou).

https://doi.org/10.1016/j.dss.2022.113863
Received 4 March 2022; Received in revised form 11 July 2022; Accepted 24 August 2022
Available online 1 September 2022
0167-9236/© 2022 Elsevier B.V. All rights reserved.

level) used on this webpage. The second challenge is the one-to-many relationship between label and topic: each label in the label set may have multiple sub-topics. For example, one topic related the appearance of the product (representative words with "crew neck," "loose," "turtleneck," and "slim") and one topic related to fashion compatibility for the product (representative words with "casual pants," "matching," "shoes" and "bag") belong to the same label, sweater. The last challenge is how to represent the visual and textual information under the constraint of their associated labels. In Fig. 1, it is evident that labels can be viewed as summaries of the semantic content of textual contents and images.

The principle mechanism of the supervised learning methods offers a promising way to address the first two challenges. Unlike Latent Dirichlet Allocation (LDA) [11] that is a popular unsupervised topic model, supervised topic models can incorporate a target variable into the learning process, to discover the latent topics. Representative attempts contain supervised Latent Dirichlet Allocation (sLDA) [12], discriminative LDA (DiscLDA) [13], maximum entropy discrimination LDA (MedLDA) [14]. However, these works assume that each document is endowed with a single response label (or rating), thus are inapplicable to a multi-label setting. Labeled LDA (L-LDA) proposed by Ramage et al. [15], is a typical multi-label probabilistic generative model. Similar to LDA, L-LDA assumes that each document is modeled as a mixture of latent topics, whereas the topic prior is constrained by the space of the label set of this document. Although L-LDA can model the top-down relationship from label to text, it can not capture latent sub-topics within a given label. In contrast, Partially Labeled Dirichlet Allocation (PLDA) [16] is proposed to solve this issue by introducing per-label latent topics to construct the one-to-many relationship between label and topic. Successfully addressing the third challenge requires the multimodal topic model as we mentioned earlier.

Therefore, we propose a method for multimodal and multi-label corpora, called **l**abel-**b**ased **m**ulti**m**odal **t**opic (LB-MMT) model. The LB-MMT model is built upon the PLDA by introducing visual information and integrating word embedding. In particular, based on the idea of the bag of words, we use scale-invariant feature transform (SIFT) to represent each image to visual words. We use the labels as supervised information to generate the text and image data. We assume that the textual words and visual words related to each text and image are drawn from a mixture of latent topics, where each topic is represented as a group of textual words and visual words that are semantically related to each other. For the textual data, continuous space word embeddings can efficiently capture the semantic relationships among words and thus be widely applied to improve the performance of topic models [17,18]. In the proposed model, to enhance the topic learning, we use a pre-trained BERT model to obtain the embedding vectors of textual words and then integrate these word embedding into our multimodal topic model. In addition, the LB-MMT model introduces multiple topics for each label, which allows us to better examine the relationship between labels and topics. Finally, we design a novel collapsed Gibbs sampling algorithm for LB-MMT inference.

We evaluate our proposed model using a field dataset crawled from Taobao. We collect a large number of marketer-generated contents including texts and images, tagged with multiple labels. The empirical experiments demonstrate superior performances of our model when compared with other benchmark models. Note that integrating the label

Fig. 1. Two examples of textual descriptions and product images labeled with multiple tags.

and visual information into the model can enhance topic learning.

The main contributions of this paper can be summarized as follows:

(1) To our best knowledge, this paper is the first attempt that jointly analyzes text and image data tagged with multiple labels. Although previous works have explored how to model text and image modalities simultaneously, integrating the label as supervised information has not been explored yet.
(2) We propose a label-based multimodal topic model to discover more interpretable topics from the multimodal data with labels. And this model strives to address the three key challenges of our problem: the top-down relationship from label to text and image; the one-to-many relationship between label and topic; and the representation of the visual and textual information in the model.
(3) To measure the performance of our model, we collect a real-world multimodal dataset with labels. Experimental results show that our model can uncover more interpretable topics than benchmark models. In addition, the learned topics can be used as input for subsequent label classification tasks.

The rest of the article is organized as follows. In Section 2, we briefly review relevant studies to our work. Section 3 mainly introduces our model framework. We present the experimental results in Section 4. Finally, we make the conclusions in the last section.

2. Related work

Our work is related to the studies of multimodal learning, unsupervised topic model, label-based supervised topic model. In this section, we review the related literature.

2.1. Multimodal learning

Multimodal learning is an active research field in computer science and other applications, e.g., recommendation systems [19–21] and popularity prediction [22]. Multimodal learning methods are mainly divided into two categories: traditional multimodal representation methods and deep learning frameworks.

For traditional multimodal representation learning, Li and Xie [23] first used linguistic content category variables (e.g., affect words, and social words) to encode textual contents, and then applied some open tools to extract some interpretable features (e.g., colorfulness and image quality) to encode images. Kan et al. [24] extended the discriminant analysis and proposed a multi-view discriminant analysis (MvDA) method to learn common discriminative features from multimodal data by maximizing inter-class variation and minimizing intra-class variation at the same time. By introducing Joint Representation Learning (JRL) to extract multimodal data features, Zhai et al. [25] jointly explored the relevance and semantic information in a unified optimization framework. Although these methods have achieved good results in multimodal feature learning, they may face several issues. First, some methods [23] usually rely on feature engineering to extract useful features, which consumes a lot of human efforts. If mistaken or imperfect features are obtained from multimodal data, the subsequent empirical studies and predictive analytics may be affected. Second, traditional machine learning methods are not able to capture the advanced nonlinear information in real-world multimodal data.

For deep multimodal representation learning, Deng et al. [26] proposed a novel deep network framework to capture more general semantic relevance between cross-modal retrieval. Hu et al. [27] employed multiple feedforward neural networks and a novel eigenvalue-based multi-view objective function to capture the complex cross-modal correlation with high nonlinearity. Qian et al. [28] proposed a text-guided attention neural network model to jointly extract features from textual and visual contents. Although the deep learning models can obtain more complex relationships between multimodal data, they often face the black-box problem, which lacks the ability to interprete the underlying processes of some phenomenon. In addition, deep learning methods rely heavily on large-scale datasets to achieve a good performance.

Our work belongs to the probabilistic graph model, which extends the topic model to the context of multimodal data. One of the major advantages of using topic models to learn multimodal representations is their generative nature, which allows a simple way to understand the model's architecture and has good interpretability.

2.2. Unsupervised topic model

Probabilistic topic models are typical natural language processing techniques that are widely used in many tasks, e,g., recommendation systems and information retrieval [29]. Popular topic models include variants of LDA, and they are unsupervised topic models. For example, Slof et al. [30] used LDA to extract topics from this textual data as variables for predicting churn reasons. Cheng et al. [31] proposed a Biterm topic model (BTM) for extracting topics from short texts.

For multimodal data, Corr-LDA is one of the representative multimodal topic models, which assumes the topics of the text and image modalities have a one-to-one correspondence [10]. And these are a series of variants of this model. For example, Cao and Li [32] proposed a spatially coherent latent topic model (Spatial-LTM) to segment and classify objects in images. To explain correlations between different modalities, Multimodal LDA [33] introduced a regression module to connect two sets of topics, thus can obtain general forms of association and allow the number of topics in the two data modalities to be different. In terms of application scenarios, KGE-MMSLDA [34], a multi-modal topic model, combined knowledge entity priors to jointly use text descriptions and images to discover interpretable topics for public social event analysis.

Our work also belongs to this line of work, extending topic models to the context for multimodal data. Different from these models, we introduce the label information to enhance the performance of topic learning in multimodal data.

2.3. Label-based supervised topic model

Supervised topic models mainly concentrate on extending unsupervised topic models (e.g., LDA) to deal with the supervised tasks. Recent efforts are often used to analyze the textual contents and responses, which each response can be viewed as a label or review rating. The supervised topic models mainly treat labels or review ratings as valuable supervision signals to improve topic learning [35]. The sLDA [12] was first proposed to model each document associated with a response variable. In detail, sLDA generates the numerical response variable (e.g., rating) using the generalized linear model and categorical response variable (e.g., label) using the softmax function. Another supervised topic model for label classification, DiscLDA [13], added a label-dependent linear transformation to the topic mixture proportions, where the transformation matrix was trained by maximizing the conditional marginal likelihood of the text labels. The MedLDA [14] incorporated the mechanism behind the support vector machines (SVMs) and topic models (e.g., LDA) into a unified framework. Specifically, MedLDA needs to balance two objective functions: the log-likelihood of the topic learning and prediction error on training data. However, these methods belong to downstream supervised topic models, which focus on the label prediction based on the topic representation of the document. In addition, these works assume that each document is endowed with a single response label (or rating), thus are inapplicable to a multi-label setting.

Recently, there has been a stream of methods that are designed from multi-label settings. L-LDA [15] made a strong assumption that each document can be only represented by limited topics and these topics are only sampled from the label set of this document. However, L-LDA

builds a one-to-many relationship between the label and topic, which tightly restricts topic assignment for each word. To break this defect of L-LDA, PLDA [16] used a topic class associated with the document labels, to infer the latent topics within each label, as well as unlabeled. Besides, based on the attention mechanism of the deep neural network, Wang and Yang [36] proposed Topic Attention Model (TAM) for label classification and optimized the parameters by variational inference. In TAM, labels in the document are used to improve the latent topic structure learned by the variational autoencoder. Although these models introduce the document-level label information, it can be used only in topic extraction from a text collection, ignoring the important visual information.

Our work is built on label-based supervised topic models. We incorporate the visual information to extend these models for supervised learning. We design the collapsed Gibbs sampling algorithm to infer model parameters. To the best of our knowledge, we are the first to develop such a supervised topic model to jointly leverage the texts and associated images with multiple labels.

3. Proposed method

The overall framework of our proposed LB-MMT model is illustrated in Fig. 2. The input to LB-MMT includes textual contents, image contents, and associated labels. Before we start the modeling process, we first apply an advanced language representation model, namely pre-trained BERT model [37], to obtain word embedding from textual contents. Then, we use SIFT algorithm to extract invariant feature points from images as visual words, construct a vocabulary, and apply the visual words in the vocabulary to represent each image. Next, we design the supervised topic model, LB-MMT, which uses labels to supervise visual and textual information generation. This model can easily mine latent topics from multimodal data, and use the discovered distribution of topic probability to interpret the meaning of labels.

Before we continue to specify the model, we first formally give our problem definition. Then, we describe our proposed model in detail. Finally, we describe the inference of the proposed model.

3.1. Problem definition

We use $M = \{d_1, d_2, \ldots, d_M\}$ to denote a set of multimodal documents, where each d contains three kinds of information $d = \{\boldsymbol{w}_d, \boldsymbol{v}_d, \boldsymbol{\Lambda}_d\}$. Let $\boldsymbol{w}_d = \{w_{d,1}, w_{d,2}, \ldots, w_{d,i}, \ldots, w_{d,N_d}\}$ denote a multi-set of textual words in document d from a textual vocabulary $\mathbb{W}$, where $w_{d,i}$ is the i^{th} word token and N_d denotes the total number of words in document d. For the word token $w_{d,i}$, we denote $e_{d,i} \in \mathbb{R}^{\boldsymbol{E}}$ as the embedding of this textual word, where $\boldsymbol{E}$ is the dimension of word embedding. Let $\boldsymbol{\Lambda}_d$ denote a set of labels from a space of labels $\mathbb{L}$. Let $\boldsymbol{v}_d = \{v_{d,1}, v_{d,2}, \ldots, v_{d,i}, \ldots, v_{d,C_d}\}$ denote visual words in document d from a visual vocabulary $\mathbb{V}$, where $v_{d,i}$ is the i^{th} visual word in the document d and C_d denote the total number of visual words in document d. In Appendix A, we list all the notations used in this paper.

In our model, we define some number of topics $\mathbb{K}_l$ (indexed by 1, 2, ⋯, K_l) for each label. For each topic $k \in \{1, 2, \ldots, K_j\}$, we define a multivariate Gaussian distribution with mean $\boldsymbol{\mu_k}$ and covariance $\boldsymbol{\Sigma_k}$, which is used to model the word embedding in the continuous space. In addition, we define $\boldsymbol{\varphi}_k$ as a $\mathbb{V}$-dimensional multinomial distribution over the visual words. The traditional topic model (e.g., LDA) assumes that each document has a K-dimensional multinomial distribution $\boldsymbol{\theta}_d$ over all topics. However, our model restricts $\boldsymbol{\theta}_d$ to be defined only over all topics $\sum_{l \in \Lambda_d} K_l$ that correspond to its labels $\boldsymbol{\Lambda}_d$.

Based on the above notations, we formally define our problem: Given all the documents that contain texts, images, and labels, the target is to learn topic-related parameters $\boldsymbol{\mu_k}$, $\boldsymbol{\Sigma_k}$ and $\boldsymbol{\varphi}_k$ related to each label, and infer document topic distribution $\boldsymbol{\theta}_d$.

3.2. Model description

3.2.1. BERT model for word representation

The traditional topic model (e.g., LDA) naturally represents documents via bag-of-words methods, while ignoring the sequential and semantic relationships among words in the documents. In our proposed model, we use an advanced embedding method, namely pre-trained BERT model [37] to obtain each word representation in our textual contents. This BERT model can transform each word as a continuous vector into a low-dimensional Euclidian space. Due to the learned word representation revealing semantic and syntactic relations, it is widely used in enhancing topic learning [17,18].

In this paper, we independently capture word embeddings offline to reduce the computational load time during topic inference. Specifically, for each textual content, we first prepend a start token [CLS], then use sequential word tokens as input for the BERT model. Finally, the contextualized representation of each word token can be computed. Due to the Chinese data of our context, we select the pre-trained BERT-base-chinese model[1] with 12-layer, 768-hidden, and 12-heads. We set the learning rate to $1e^{-5}$, batch size to 100 and epoch to 50 for the task of fine-tuning in BERT. However, this tool regards each Chinese character as a word, and each Chinese character is represented as a vector with 768-dimension. To obtain the word embedding, we use an average strategy to integrate all the character embedding in each word [38].

3.2.2. BOVW model for image representation

The bag of visual words (BOVW) model is a traditional and basic method for automatic image representation, which has been widely used in image retrieval and classification [39,40]. Its concept is adapted from the bag of words (BOW) in natural language processing (NLP). Instead of textual words, BOVW attempts to encode the image with the statistical frequency of the visual words [41]. As shown in Fig. 3, BOVW consists of three steps: the SIFT feature extraction, K-means clustering, and representation.

We first apply the Scale Invariant Feature Transform (SIFT) [42] to identify local features in images. The SIFT is one of the most efficient and robust computer vision algorithm that has been applied to different fields of management research [43,44]. The first step of SIFT is to extract a set of keypoints that reveal the most important and distinct information from local regions of the image. To detect the stable keypoints, Lowe [42] applied a difference-of-Gaussian function to compute extrema location in scale-space. Then, we extract more distinctive feature descriptors for the keypoints by sampling the image gradient magnitudes and orientations around the keypoints. The descriptor is a vector in continuous space, which is used to encode the salient aspects of keypoint. Specifically, we first divide each image into several patches. Then, for each image patch, we extract its SIFT descriptor that is presented as a promising 128- dimension vector. For this step, we use an existing software library that is implemented by Python language, namely opencv-python.[2] This library has been widely used in image processing [45–47]. Fig. 4 gives three examples to show the keypoints extracted by SIFT on our dataset. The identified regions are displayed by lower images, with keypoints displayed as circles with color.

Second, using the SIFT descriptors as input, we employ the K-means clustering method to capture the discrete set of visual terms for constructing the visual vocabulary. Each cluster is regarded as a distinct visual word in the vocabulary. We select the number of k in K-means by varying the number of k and using the label classification task to evaluate the clustering performance. According to the results (see Appendix B), the number of k in K-means is set to 900. Finally, each image can be represented by 900 visual words. With the cluster assignments of each visual word in an image, we can obtain the visual word distribution, and

[1] https://huggingface.co/bert-base-chinese
[2] https://github.com/opencv/opencv-python

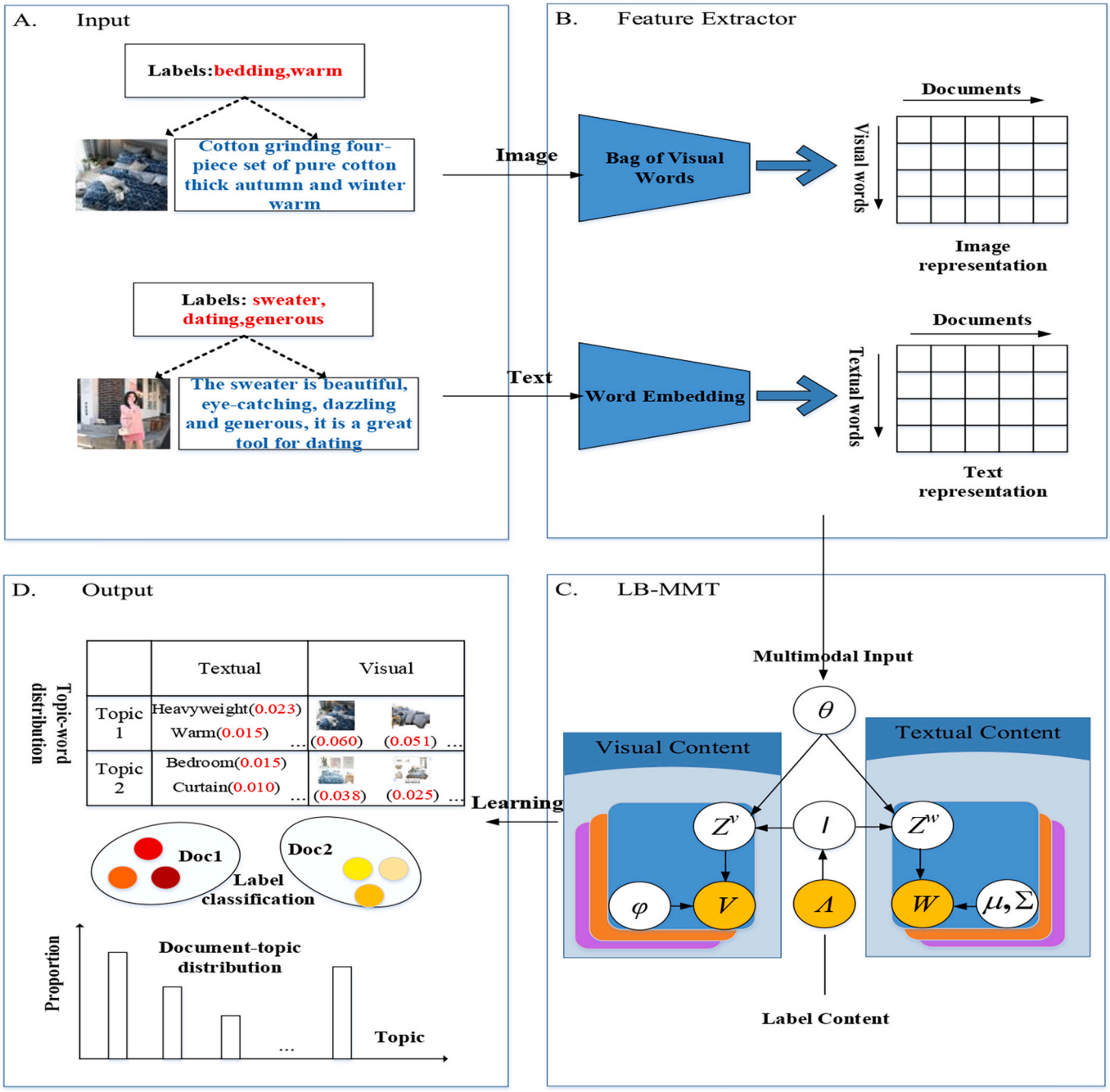

Fig. 2. The framework of the proposed method.

create a frequency histogram as a feature vector to represent this image.

3.2.3. LB-MMT model

In this part, we propose a label-based multimodal topic (LB-MMT) model that is used to mine latent topics from multimodal data, and use the discovered distribution of topic probability to detect valuable topics. In the LB-MMT model, labels, textual and visual information are integrated into a unified framework. All textual and visual words are generated in the same topic space.

Generally, many documents consist of texts and associated visual content, and these documents will carry multiple human-provided tags. To make better use of the multimodal information of documents, the proposed model integrates labels into multimodal topic modeling by using labels to supervise topic generation of text and visual contents. By doing this, the LB-MMT model improves the performance of topic learning in multimodal data. Fig. 5 shows the graphical representation of the LB-MMT, where the shaded nodes are the observed variable and unshaded nodes are the latent variable. Next, we describe the LB-MMT in detail.

Each multimodal document d consists of N_d textual words, C_d visual words and Λ_d labels. The LB-MMT model can be considered as a three-layer model with document layer, label layer, and topic layer. In the document layer, unlike the traditional topic model, the proposed model assumes that each multimodal document is a mixture of only those topics that are in a topic class related to one or more of the labels in this document. In the label layer, we argue that labels are generally available, and they can serve as useful supervision signals to improve topic modeling learning. Several supervised topic models have integrated the label information related to documents [35,48]. The proposed model assumes that there are a set of $\mathbb{L}$ labels (indexed by 1,2,…,L). We assign a

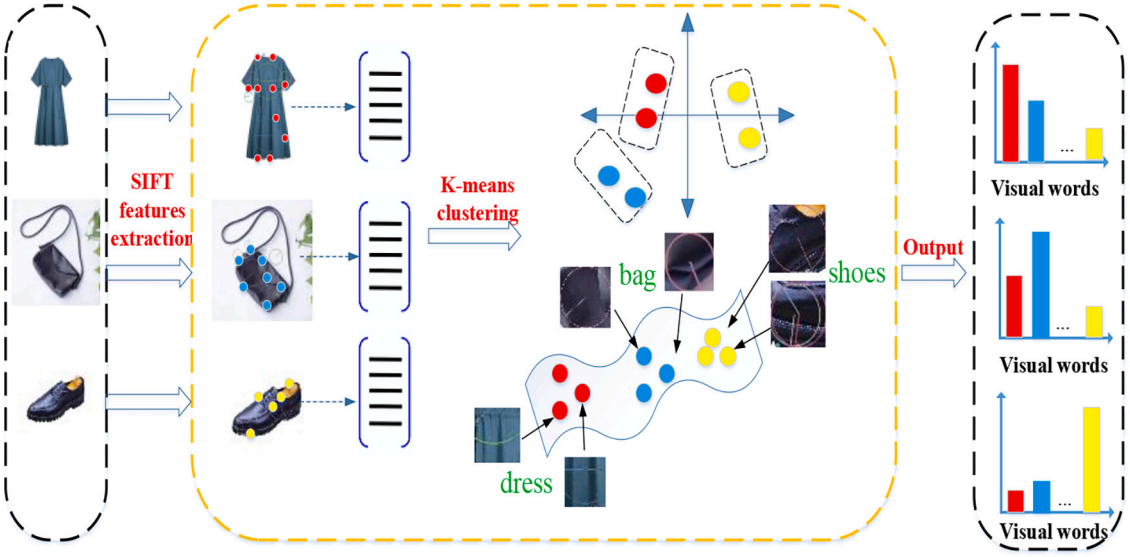

Fig. 3. The process of BOVW representation.

Fig. 4. Visualization of the keypoints extracted by SITF.

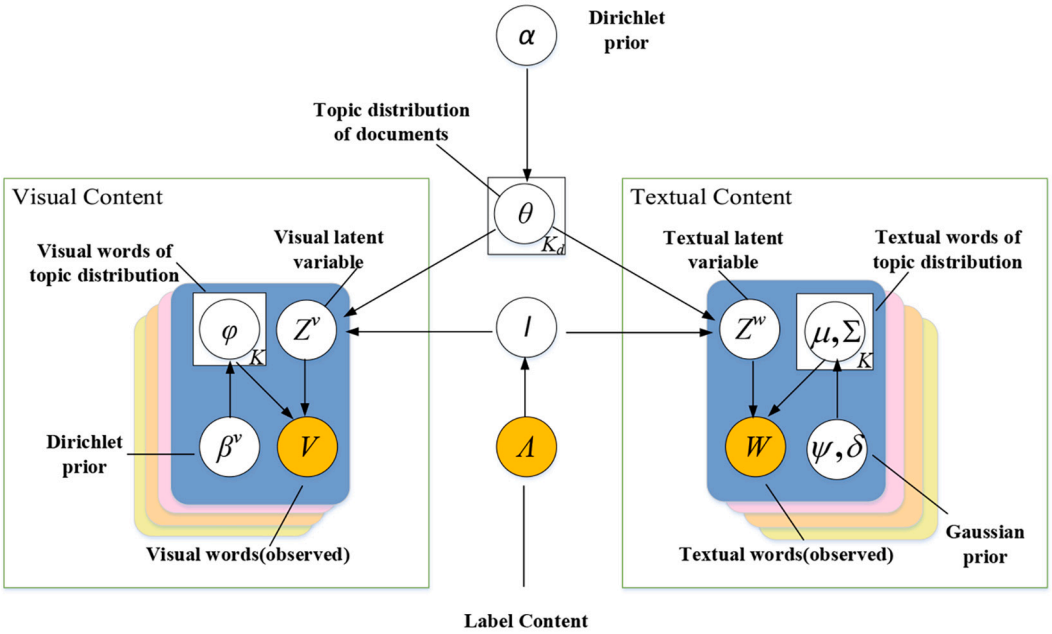

Fig. 5. The graphical representation of LB-MMT.

number of topics $\mathbb{K}_l$ for each label. Therefore, a one-to-many relationship is formed between the label and the topics. Based on the mechanisms of supervised learning, we constrain the topic assignments so that the topic for the textual word or visual word is only related to the labels in the document. Because these are general or noise information that is irrelevant to labels of documents, we also define a background label to generate some background topics, to filter out these irrelevant information. In the topic layer, each topic k is modeled as two types of distributions over the textual vocabulary $\mathbb{W}$ and over the visual vocabulary $\mathbb{V}$. Since the textual words are transformed into continuous vectors by BERT, we model the word embedding using a multivariate Gaussian distribution [49]. Formally, we give the generative process of the proposed LB-MMT model in Fig. 6.

As shown in Fig. 6, textual topic z^w and visual topic z^v are generated from the same topic distribution θ. In the generative process of a multimodal document, textual and visual information are processed independently. Similar to LDA, each multimodal document is considered as a mixture of latent topics. However, these topics are shared across documents' labels. To be more specific, for the document d, we first pick up a label $j \in \Lambda_d$ and generate the per-document topic distribution $\theta_{d,j}$ over topics $1 \ldots K_j$. The $\theta_{d,j}$ is drawn from a Dirichlet prior with parameter $\alpha_{\theta_{d,j}}$. Then, to generate the i^{th} textual word of document d, a topic $z^w_{d,i}$ is first generated from $\theta_{d,j}$. Finally, the textual word embedding $e_{d,i}$ is chosen from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_{j,z^w_{d,i}}$ and covariance $\boldsymbol{\Sigma}_{j,z^w_{d,i}}$. To improve the computational efficiency, we place conjugate priors on the parameters of the multivariate Gaussian distribution. Thus, the covariance for each topic is drawn from an inverse Wishart distribution with parameters ψ and δ. And the mean for each topic is drawn from a multivariate Gaussian distribution with the parameters ρ and κ. On the other hand, to generate the i^{th} visual word of document d, a topic $z^v_{d,i}$ is first generated from $\theta_{d,j}$. Finally, the visual word $v_{d,i}$ is chosen from the per-topic visual word distribution $\varphi_{j,z^v_{d,i}}$.

We can learn various probability distributions from the LB-MMT model to describe the distribution of labels, textual and visual content of multimodal documents. $\theta_{d,j,k}$ represents the degree of preference that topic k belongs to label j in the multimodal document d. For the multimodal documents, $\boldsymbol{\mu}_{j,k}$ and $\boldsymbol{\Sigma}_{j,k}$ is used to model the probability of textual word belonging to topic k of label j, $\varphi_{j,k,v}$ captures the probability of visual word v belonging to topic k of label j.

3.3. Model inference

To learn the latent variables (i.e., $\mu_{j,k}$, $\Sigma_{j,k}$, $\theta_{d,j,k}$, $\varphi_{j,k,v}$), we propose a collapsed Gibbs sampling algorithm. Deriving the joint probability distribution is the key to implementing the collapsed Gibbs sampling algorithm. During the sampling process, the new value of the latent variables is iteratively updated according to the state before the conditional distribution. The solution processes of the latent variables are shown as follows.

The joint probability distribution of the observed textual word embedding e and visual word v with the unobserved label, unobserved textual word topic assignment z^w and visual word topic assignment z^v can be written as follows:

$$p(e, v, z^w, z^v, l|\alpha, \beta^w, \beta^v) = p(e|z^w, l, \xi)p(v|z^v, l, \beta^v)p(z^w, z^v, l|\alpha) \tag{1}$$

where the tuple $\xi = (\psi, \delta, \rho, \kappa)$ denotes the parameters of the prior distribution for the multivariate Gaussian distribution. We note that the first part of Eq. (1) can be derived as follows:

$$p(e|z^w, l, \xi) = \int\int \mathcal{N}(e|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\mathcal{N}\left(\boldsymbol{\mu}_k|\rho, \frac{1}{\kappa}\boldsymbol{\Sigma}_k\right)\mathscr{W}^{-1}(\psi, \delta)d\boldsymbol{\mu}_k d\boldsymbol{\Sigma}_k = \prod_{j\in\Lambda}\prod_{k=1}^{K_j} \curlywedge_{\delta_{j,k}-E+1}\left(e|\mu_{j,k}, \frac{\kappa_{j,k}+1}{\kappa_{j,k}}\Sigma_{j,k}\right) \tag{2}$$

Algorithm 1: Generative process of LB-MMT

```
for each topic k ∈ {1, 2, ··· , K_j} do
    Generate textual topic covariance Σ_{j,k} ~ W^{-1}(ψ, δ)
    Generate textual topic mean μ_{j,k} ~ N(ρ, (1/κ)Σ_{j,k})
    Generate φ^v_{j,k} ~ Dir(β^v)
end
for each document d ∈ {1, 2, ··· , M} do
    for each label j ∈ Λ_d, where Λ_d ∈ {1, 2, ··· , L} do
        Generate θ_{d,j} ~ Dir(α_{θ_{d,j}})
        for i^th textual word w_{d,i} of document d, where i ∈ {1, 2, ··· , N_d}
        do
            Generate z^w_{d,i} ~ Multinominal(θ_{d,j})
            Generate e^w_{d,i} ~ N(μ_{j,z^w_{d,i}}, Σ_{j,z^w_{d,i}})
        end
        for i^th visual word v_{d,i} of document d, where i ∈ {1, 2, ··· , L_d}
        do
            Generate z^v_{d,i} ~ Multinominal(θ_{d,j})
            Generate v_{d,i} ~ Multinominal(φ^v_{j,z^v_{d,i}})
        end
    end
end
```

Fig. 6. The generative process of the LB-MMT model.

where $\curlywedge_f(\boldsymbol{X}|\boldsymbol{\mu}', \boldsymbol{\Sigma}')$ denotes the multivariate t-distribution with the degree of freedom f and parameters $\boldsymbol{\mu}'$ and $\boldsymbol{\Sigma}'$. The posterior parameters in this t-distribution are shown in Appendix C. The second part of Eq. (1) can be derived as follows:

$$p(v|z^v, l, \beta^v) = \int p(v|z^v, l, \varphi)p(l, \varphi|\beta^v)d\varphi = \prod_{j\in\Lambda}\prod_{k=1}^{K_j}\frac{\Delta\left(m_{*,j,k,v}+\beta^v\right)}{\Delta(\beta^v)} \tag{3}$$

where the notation $m_{*,j,k,v} = \sum_{d=1}^{M} m_{d,j,k,v}$ denotes the number of the visual word v generated by the label j topic $k \in \{1, 2, \ldots, K_j\}$. The third part of Eq. (1) can be derived as follows:

$$p(z^w, z^v, l|\alpha) = \int p(z^w|l, \theta)p(z^v|l, \theta)p(\theta|\alpha)d\theta \tag{4}$$

Eq. (4) can be considered the joint likelihood of the topic assignments and labels. We not that:

$$p(z^w|l,\theta) = \prod_{d=1}^{M}\prod_{i=1}^{N_d} p\left(z_{d,i}^w|l_{d,i}, \theta_{d,l_{d,i}}\right) = \prod_{d=1}^{M}\prod_{i=1}^{N_d}\theta_{d,l_{d,i},z_{d,i}^w} = \prod_{d=1}^{M}\prod_{j\in\Lambda_d}\prod_{k=1}^{K_j}\left(\theta_{d,j,k}\right)^{n_{d,j,k,*}} \tag{5}$$

where $n_{d,\,j,\,k,\,*}$ represents the number of textual words corresponding to the label $j \in \Lambda_d$ topic $k \in \{1, 2, \ldots, K_j\}$ in the document d and $n_{d,j,k,*} = \sum_{w=1}^{\mathbb{W}} n_{d,j,k,w}$. $n_{d,\,j,\,k,\,w}$ denotes the number of the textual word w generated by the label j's topic $k \in \{1, 2, \ldots, K_j\}$. The second part of Eq. (4) can be written as:

$$p(z^v|l,\theta) = \prod_{d=1}^{M}\prod_{i=1}^{L_d} p\left(z_{d,i}^v|l_{d,i}, \theta_{d,l_{d,i}}\right) = \prod_{d=1}^{M}\prod_{i=1}^{L_d}\theta_{d,l_{d,i},z_{d,i}^v} = \prod_{d=1}^{M}\prod_{j\in\Lambda_d}\prod_{k=1}^{K_j}\left(\theta_{d,j,k}\right)^{m_{d,j,k,*}} \tag{6}$$

where $m_{d,\,j,\,k,\,*}$ represents the number of visual words corresponding to the label $j \in \Lambda_d$ topic $k \in \{1, 2, \ldots, K_j\}$ in the document d and $m_{d,j,k,*} = \sum_{v=1}^{\mathbb{V}} m_{d,j,k,v}$. Thus, can rewrite Eq. (4) into the following form:

$$p(z^w, z^v, l|\alpha) = \prod_{d=1}^{M}\prod_{j\in\Lambda_d}\frac{\Delta\left(n_{d,j,k,*}+m_{d,j,k,*}+\alpha\right)}{\Delta(\alpha)} \tag{7}$$

The assignment of textual words in document d is restricted to the set of observed label Λ_d. Using the above Equations, collapsed Gibbs sampling formula for textual word w can be obtained:

$$p\left(l_{d,i}=j, z_{d,i}^w=k|l_{-d,i}, z_{-d,i}^w, z^v, e_{d,i}=e, v, \alpha, \xi\right) \propto \mathbb{I}\left(j\in\Lambda_d \wedge k = 1, 2, \ldots, K_j\right)\left(n_{d,j,k,*}^{-d,i}+m_{d,j,k,*}+\alpha\right)\curlywedge_{\delta_{j,k}-E+1}\left(e|\mu_{j,k}, \frac{\kappa_{j,k}+1}{\kappa_{j,k}}\Sigma_{j,k}\right) \tag{8}$$

where $\mathbb{I}(\bullet)$ is the indicator function. The notation $n^{-d,\,i}$ denotes the corresponding count except for the current position i of the document d. And $n_{d,\,j,\,k,\,*}^{-d,\,i}$ represents the number of textual words corresponding to the label $j \in \Lambda_d$ topic $k \in \{1, 2, \ldots, K_j\}$ in the document d, except for the current textual word at the position i of the document d.

Similarly, collapsed Gibbs sampling formula for visual word v can be obtained:

$$p\left(l_{d,i}=j, z_{d,i}^v=k|l_{-d,i}, z_{-d,i}^v, z^w, v_{\mathrm{d},i}=v, e, \alpha, \beta^v\right) \propto \mathbb{I}\left(j\in\Lambda_d \wedge k = 1, 2, \ldots, K_j\right)\frac{m_{*,j,k,v}^{-d,i}+\beta^v}{\sum_{v=1}^{\mathbb{V}}\left(m_{*,j,k,v}^{-d,i}+\beta^v\right)}\left(m_{d,j,k,*}^{-d,i}+n_{d,j,k,*}+\alpha\right) \tag{9}$$

where $m^{-d,\,i}$ denotes the corresponding count except for the current position i of the document d. The $m_{*,j,k,v}^{-d,i}$ represents the number of the visual word v generated by the label j topic $k \in \{1, 2, \ldots, K_j\}$, except for the current visual word at the position i of the document d. And $m_{d,j,k,*}^{-d,i}$ represents the number of visual words corresponding to the label $j \in \Lambda_d$ topic $k \in \{1, 2, \ldots, K_j\}$ in the document d, except for the current visual word at the position i of the document d.

After performing the collapsed Gibbs sampling algorithm until convergence, we can estimate the latent variable $\theta_{d,\,j,\,k}$, and $\varphi_{j,\,k,\,v}$:

$$\theta_{d,j,k} = \frac{n_{d,j,k,*}+m_{d,j,k,*}+\alpha}{\sum_{k=1}^{K_j}\left(n_{d,j,k,*}+m_{d,j,k,*}+\alpha\right)} \tag{10}$$

$$\varphi_{j,k,v} = \frac{m_{*,j,k,v}+\beta^v}{\sum_{v=1}^{\mathbb{V}}\left(m_{*,j,k,v}+\beta^v\right)} \tag{11}$$

In addition, based on Eq. (8), we can determine topic assignments for each textual word. We define a parameter $\phi_{j,\,k,\,w}$, which denotes the probability of textual word w belongs to topic k of label j. To simplify the calculation, we apply empirical estimation to measure $\phi_{j,\,k,\,w}$.

$$\phi_{j,k,w} = \frac{n_{*,j,k,w}}{\sum_{w=1}^{\mathbb{W}}\left(n_{*,j,k,w}\right)} \tag{12}$$

where the notation $n_{*,j,k,w} = \sum_{d=1}^{M} n_{d,j,k,w}$. As formulated above, we can obtain updated latent variables $\theta_{d,\,j,\,k}$, $\varphi_{j,\,k,\,v}$ and $\phi_{j,\,k,\,w}$. We noted that the update rules of this sampler are similar to the Latent Dirichlet Allocation, with the constraint that only the topics corresponding to the document labels can be sampled. We can easily track the topic assignments of the visual words and textual words via the Gibbs sampling algorithm, since every topic only takes part in a single label. We don't need to allocate resources to record which label is assigned to each visual word or textual word. Thus, the proposed model has substantial computational advantages, especially on the dataset with more shared labels across documents. We give the process of the collapsed Gibbs sampling algorithm in Appendix D.

Table 1
Statistics of the dataset.

	Mean	Max	Min	Standard deviation
Number of words per document	168.01	1292	27	160.85
Number of images per document	5.09	44	2	5.67
Number of labels per document	1.94	6	1	1.04

4. Experimental results

In this section, we validate the performance of our proposed model on a real-world dataset. We first describe our dataset and the preprocessing process for the raw data. Then, we give the selected competitors for model comparison. Finally, we analyze the experimental results through quantitative evaluation and qualitative evaluation.

4.1. Data

We collect the dataset from the *E*-commerce platform Taobao to

evaluate the performance of the proposed model. This dataset is related to marketer-generated contents. To collect the texts and their associated images, we implement a web crawler via the Java language. Due to the unstructured raw data, we perform a series of preprocessing steps to transform it into a structured format as input of the proposed model. First, we filter out the labels whose frequencies on the whole corpus are very low (<10) or very high (>500). Second, we remove the multimodal data with less than two images, one label, and 20 words in the textual content. Third, because there is no segmentation symbol in Chinese, we use a word segmentation tool (i.e., Jieba[3]) to process textual contents. Finally, we remove the noise information (e.g., stopwords) in textual contents to improve the performance of topic learning. We give the summary of our preprocessed dataset in Table 1. The dataset contains 12,157 multimodal documents, 59,084 images, and 122 unique labels. On average, each multimodal document consists of 168.01 words, 5.09 images, and 1.94 labels.

4.2. Baselines for comparison

To evaluate the performance, we compare the proposed model with several carefully selected alternative baselines. We organize these baselines into two categories: unsupervised and supervised topic models. The unsupervised topic models are the following:

LDA: LDA [11] model is one of the most classic methods in the field of topic learning. We apply the standard LDA model to extract the topics from the text descriptions in our dataset. And we use Gibbs sampling to infer the latent parameters, e.g., document-topic distribution.

G-LDA: G-LDA [49] is a variant of LDA, which assumes each document is a collection of word embeddings, and each word embedding is drawn from multivariate Gaussian distributions. We use Gibbs sampling to infer the latent parameters.

Link-LDA: Similar to PLDA, Link-LDA [50] is a mixed-membership model for documents consisting of abstracts and references. We use this model for the task of multi-modal data analysis, which treats the text descriptions and visual information as words and references in the framework of Link-LDA. In addition, we apply the Gibbs sampling to learn the latent parameters.

Corr-LDA: Corr-LDA [10] is a multimodal topic model that can build the correspondences between images and text modalities. We use text descriptions and visual information as input of Corr-LDA to detect topics. Different from Link-LDA, this model first generates the visual words and subsequently generates textual words. We apply variational inference to learn the latent parameters.

The supervised topic models are the following:

L-LDA: L-LDA [15] can build the one-to-one relationship between label and topic. The input of this model is text descriptions and their corresponding labels. We use Gibbs sampling to learn the latent parameters of this model.

sLDA: The sLDA [12] introduces a response variable for each document and uses the generalized linear model to generate this response variable. This model assumes that each document is endowed with a single response label (or rating), which is inapplicable to a multi-label setting. To run this model, we thus select only one label for each document based on the frequency of this label in the whole corpus. We use variational inference to learn the latent parameters.

PLDA: PLDA [16] is used to model documents and their related label information, simultaneously. This model can infer latent topic structure within the scope of observed labels. We use Gibbs sampling to learn the latent parameters of this model.

LB-MMT-BOW: This model is a variant of LB-MMT, which replaces the word embedding in the LB-MMT model with the bag of words (BOW). Similar to LDA, LB-MMT-BOW assumes that a topic is a multinomial distribution over a fixed vocabulary of words. We use Gibbs sampling to learn the latent parameters of this model.

To achieve a fair comparison, we tune the hyperparameter values of all models. We empirically set the hyperparameters $\beta^v = 0.01$, and $\alpha = 50/K$ for our proposed model. The hyperparameter $\boldsymbol{\rho}$ is set to the mean of all the word embeddings, δ to the number of dimensions of the word embeddings, and $\boldsymbol{\psi}$ to an identity matrix. Because our model supports latent sub-topics within a given label, we thus choose the hyperparameters $K_l = 2$ for labels. For LDA and L-LDA, we set the hyperparameters $\alpha = 50/K$, and $\beta = 0.01$. For G-LDA, we set the hyperparameters $\alpha = 50/K$, and other hyperparameters for multivariate Gaussian distribution are the same as our model settings. For Link-LDA and Corr-LDA, we set the hyperparameters $\alpha = 50/K$, $\beta^w = \beta^v = 0.01$. For sLDA, we set $\alpha = 50/K$. For PLDA, we set the hyperparameters $\alpha = 50/K$, $\beta = 0.01$, and the size of the sub-topics under each label $K_l = 2$. For LB-MMT-BOW, we set $\beta^w = 0.01$ and other hyperparameters are the same as our model settings. We implement the proposed model by Java language and run it on an Ubuntu 16.04 (GNU/Linux) server.

4.3. Topic coherence

The objective of this research is to learn the topic by using the method of probability topic model in the given data of associated text and images with labels. To measure the quality of topics learned from the topic model, we use a metric, namely topic coherence [51,52]. Topic coherence is considered to be an important measure of human understanding of the latent topics learned. Significantly, the superiority of this metric is that its value is evaluated over the original corpus used to train the topic models, rather than an external corpus. We apply the following equation to calculate topic coherence scores.

$$C\left(t, W^{(t)}\right) = \sum_{m=2}^{M} \sum_{l=1}^{m-1} log \frac{D\left(w_m^{(t)}, w_l^{(t)}\right) + \epsilon}{D\left(w_l^{(t)}\right)} \tag{13}$$

where $D(w)$ denotes the document frequency of the word type w, $D(w, w')$ denotes the number of times w and w' appearing together in a document. $W^{(t)} = (w_1^{(t)}, w_2^{(t)}, \ldots, w_M^{(t)})$ denotes the top M words with the highest frequency under the topic t. In order to guarantee that the score yields a real number, we introduce a smoothing factor ϵ and set it to 0.1. We use the average coherence scores over all topics to evaluate the overall quality of topic models:

$$Avg(c) = \frac{\sum_{t=1}^{K} C\left(t, w^{(t)}\right)}{K} \tag{14}$$

Better topic quality is indicated by a higher value of $Avg(c)$. To measure $Avg(c)$, we run five experiments for each topic model on our dataset by fixing the number of words as 5, 10, 15, and 20, respectively. Table 2 displays the results. As shown in Table 2, we can see that G-LDA performs better than LDA, which shows using word embedding can improve topic learning. Corr-LDA and Link-LDA significantly outperform LDA ($p < 10^{-3}$), which means considering text descriptions and

Table 2
The average coherence scores compared with other models.

	Number of words	M = 5	M = 10	M = 15	M = 20
Unsupervised topic model	LDA	−27.6	−134.0	−336.5	−622.7
	G-LDA	−19.0	−102.4	−260.3	−489.6
	Link-LDA	−18.8	−99.2	−242.6	−496.7
	Corr-LDA	−19.1	−102.8	−257.9	−504.3
Supervised topic model	L-LDA	−18.5	−97.6	−243.5	−468.7
	sLDA	−18.7	−98.5	−246.7	−475.3
	PLDA	−18.4	−96.5	−231.7	−463.3
	LB-MMT-BOW	−11.7	−62.1	−165.9	−335.8
Our approach	**LB-MMT**	**−10.3**	**−56.0**	**−156.4**	**−315.6**

[3] https://github.com/fxsjy/jieba

visual information jointly can significantly improve topic learning. Compared to LDA and G-LDA, we note that the supervised topic models (i.e., L-LDA, sLDA, PLDA, and LB-MMT-BOW) improve a lot, which demonstrates the joint analysis of text and labels can help the topic learning. LB-MMT-BOW achieves the second-best performance, which shows leveraging the texts and associated images with labels can detect more valuable topics. The proposed model LB-MMT significantly outperforms the state-of-the-art baseline methods, which means our model can identify more prominent and coherent topics. Compared to LB-MMT-BOW, the proposed LB-MMT model introduces word embeddings to ehance topic learning.

4.4. Label classification

Following the prior work [53], we first use the proposed model and the baseline methods to extract latent topics and represent document-topic distributions to construct new variables, and subsequently, these variables are as input into a deep learning method for the multi-label classification task. Specifically, we randomly select a set of multi-modal documents to train the LB-MMT model. In this process, the label information serves as supervision signals to control the topic assignments. Then, based on the observable data (i.e., visual words and textual words), we can easily infer document-topic distributions for new multimodal documents, similar to previous studies [54,55].

For the classification task, we select the Deep & Cross network (DCN) method [56], which can jointly train feed-forward neural networks with document-topic distributions and linear models with feature transformations for label classification. The superiority of this deep learning model is that it can capture feature interactions, thus bringing additional improvement for label classification. Fig. 7 gives the framework of the DCN method. In addition, we randomly split the dataset into a training set (80%), a test set (10%), and a validation set (10%).

To evaluate the performance of the classification with different topic models, we adopt two different evaluation metrics: average *Precision@N* and *Recall@N*. For the robust comparison, we run the DCN method five times based on the document-topic distributions of our proposed model and each baseline. The higher the values of *Precision@N* and *Recall@N* indicate better classification results. Fig. 8 shows the comparison results. We find consistently, the proposed model (LB-MMT) achieves a better performance than other comparison methods over these two metrics. This suggests that our model is reasonable to explicitly model the mapping between multimodal topics and labels. In contrast with LB-MMT, LB-MMT-BOW achieves the second-best performance, which reveals that introducing word embedding learned by BERT can obtain more discriminated topics for label classification. In addition, we note that the improvements of LB-MMT and LB-MMT-BOW over PLDA are significant ($p < 0.01$) by using a paired *t*-test.

To analyze the impact of introducing images into the topic model, Table 3 gives the case studies of label predictions by LB-MMT and PLDA. As shown in this table, we give the ground-truth labels as references and list the top-3 highest-scoring predicted labels for the LB-MMT and PLDA. And we mark the correct predictions in blue and incorrect predictions in red. It is clear that our approach produces higher quality results than PLDA.

4.5. Robustness analysis

To evaluate the stability of the proposed model, we conduct robustness analysis. To this end, we first verify the impact of different word embedding methods on the model performance. We compare the experimental results with GloVe [57], Word2Vec [58] and BERT. For a fair comparison, the word embedding size of these three methods is all set to 768. The performance of each model is assessed by topic coherence scores. Table 4 shows the performance comparisons on these word embedding methods. We find that LB-MMT with GloVe performs similarly to LB-MMT (Word2Vec). Although the overall performances of the proposed model with BERT are better than that with GloVe and Word2Vec, the improvement is not very significant ($p < 0.1$) by using a paired *t*-test.

Although SIFT is among the most prevalent and influential image representation methods, it still needs to be compared with alternative algorithms to investigate whether this method is suitable for our image data. To achieve this goal, we employ two robust algorithms, namely ORB [59], and SURF [60] as the alternatives, to extract feature (or descriptor) for each image patch. Based on these features, we use the K-Means to construct visual vocabulary and then obtain representations for all images. We set the number of visual vocabulary to 900. Table 5 reports the comparative results with ORB, SURF, and SIFT. We find that LB-MMT with SIFT performs better than that with SURF and ORB. We

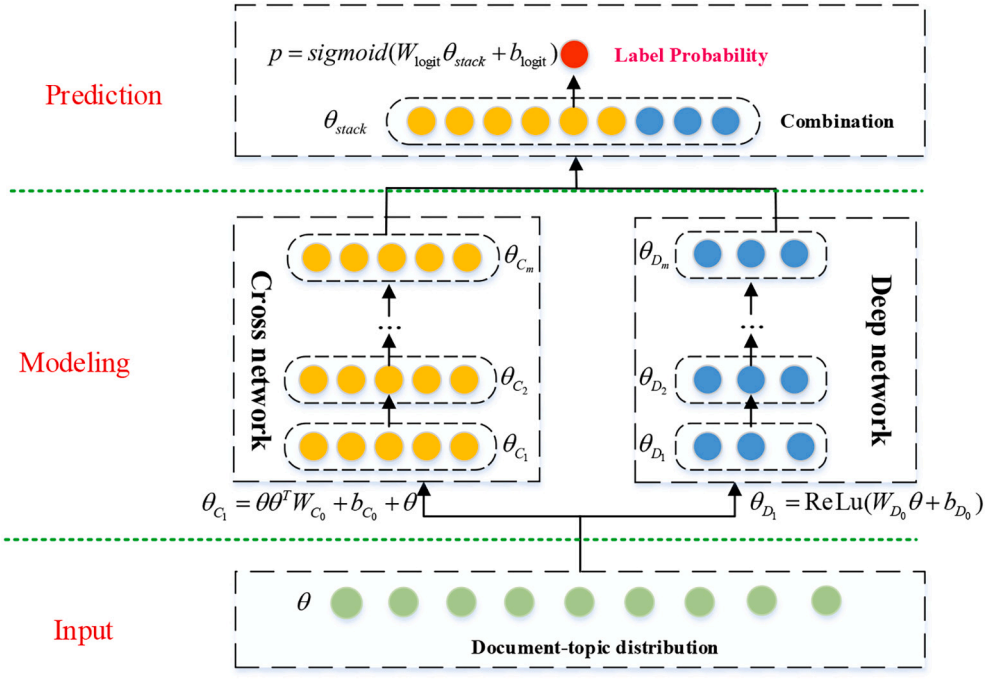

Fig. 7. The Deep & Cross network method for label prediction.

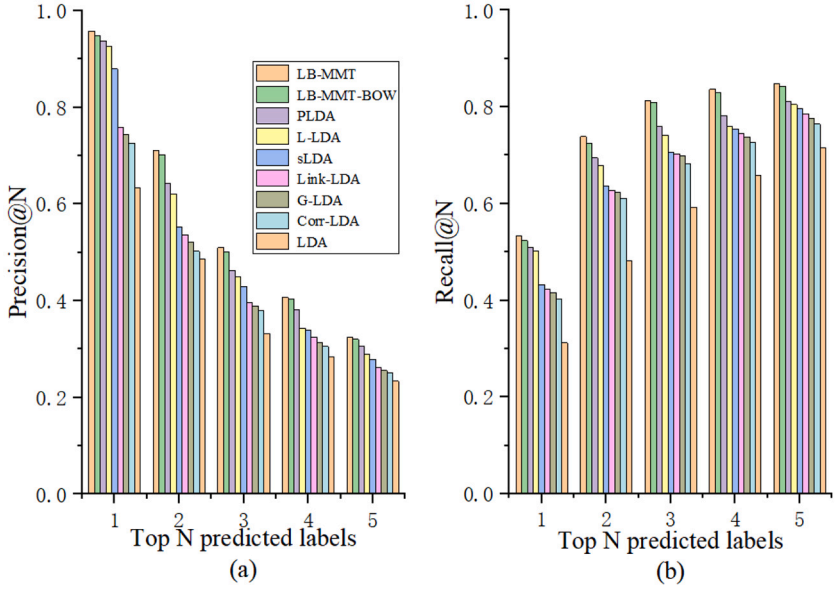

Fig. 8. The results of the Label classification.

Table 3
Prediction results with six randomly selected samples.

Index	True label	PLDA	LB-MMT	Image
1	Retro/Japanese/ Autumn-winter	Japanese/ Literary/ Autumn-winter	Autumn-winter/ Retro/Japanese	
2	Board shoes/ Men's shoe	Board shoes/ Men's shoe/Jeans	Board shoes/ Men's shoes/Manly	
3	Original/Loose/Coat/	Original/Coat/ Casuals	Coat/Loose/Original	
4	Sweet/Fresh/Playful	Fresh /Sexy /Sweet	Fresh/Playful/ Sweet	
5	Shoulder-bag/ Sweet/ Delicate	Delicate/Shoulder-bag/ Generous	Delicate/Sweet/Shoulder-bag	
6	Sweater	Sweater/Sweatpants/ Autumn-winter/	Sweater/ Leisure/ Autumn-winter/	

Table 4
Performance comparisons on different word embedding methods.

Number of words	M = 5	M = 10	M = 15	M = 20
LB-MMT (GloVe)	−10.1	−59.2	−165.7	−320.5
LB-MMT (Word2Vec)	−10.2	−58.3	−162.4	−325.8
LB-MMT (BERT)	−10.3	−56.0	−156.4	−315.6

Table 5
Performance comparisons for different image features.

Number of words	M = 5	M = 10	M = 15	M = 20
LB-MMT (ORB)	−11.8	−65.9	−175.6	−350.4
LB-MMT (SURF)	−11.3	−63.2	−168.7	−330.6
LB-MMT (SIFT)	−10.3	−56.0	−156.4	−315.6

suggest that SIFT has stable and good performance in keypoints detection, compared with the other two alternatives. This is possible because that SURF and ORB are not good at processing images with varying intensity and color composition values [61].

4.6. Qualitative analysis of multimodal topics

We further study the content of the multimodal topics for qualitative analysis. Because of the space limitation, we randomly select four labels with eight topics in our results for visualization. The LB-MMT model

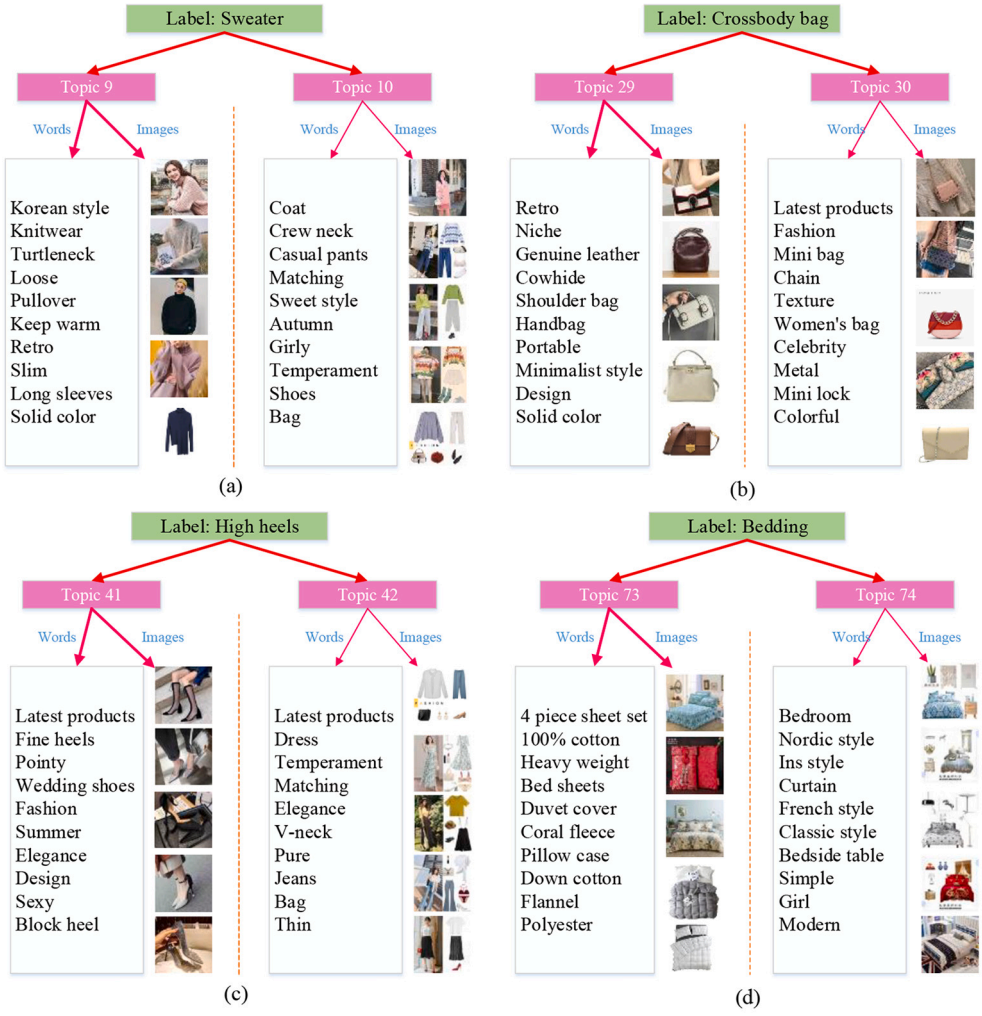

Fig. 9. Examples of multimodal topics.

yields distributions of textual words and visual words for each topic. Based on these probabilities, we can easily obtain how likely a given textual word or visual word assigns to that topic. We use these probabilities to rank the words and images in each topic. For each topic of labels, we present its 10 most probable words and 5 images in Fig. 9.

From Fig. 9, we note that our model can yield a one-to-many relationship between labels and multimodal topics. Under the constraints of labels, it is very interesting to see how certain textual words and images congregate to form meaningful and interpretable multimodal topics. For example, in Fig. 9 (a), topic 9 is about the appearance of the sweater, since textual words "turtleneck," "loose," "pullover," and "slim" allow people to quickly sketch what a sweater looks like. This is also confirmed by the images on topic 9. Interestingly, topic 10 talks about fashion compatibility for the sweater, as evidenced by the use of words such as "casual pants," "matching," "shoes" and "bag." We found that although these two topics belong to the same label (i.e., sweater), there are obvious differences in the content described. In Fig. 9 (b), we note topic 29 contains the words "retro," "niche" and "minimalist style," and topic 29 contains the words "latest products," "fashion" and "celebrity." While the textual words and images in topic 29 and topic 30 both focus on describing the style of the crossbody bag, the former is related to the niche retro style, and the latter is related to the popular style of celebrities. These results not only reflect the public opinion trend of market products, but also help marketers to analyze the style of their products. In Fig. 9 (c), topic 41 is concerned with stilettos, as evidenced by the use of words such as "fine-heels" and "pointy." From the textual words and images in topic 42, we find this topic is associated with fashion compatibility for stilettos. Similarly, topic 73 is about the material of the bedding (e.g., "100% cotton," "coral fleece," and "down cotton") and topic 74 is about the style of the bedding (e.g., "Nordic style," "Ins style" and "Curtain"). According to the above analysis, we conclude that our model can generate deep qualitative insights into how multimodal data is organized under the constraints of labels.

4.7. Practical implications

The spread of the Internet and mobile applications have led to large-scale multimodal information posted by users online through media, e.g., Taobao and Amazon. This paper contributes to providing a powerful framework for discovering interesting patterns from these multimodal data. The results of this study provide several important practical implications.

First, marketers in online platforms (e.g., Taobao) may take advantage of our results to design more attractive multimodal contents that can influence consumer attention and promote consumers' purchase intentions. For example, the proposed framework provides a reliable and reasonable way to extract meaningful features (i.e., multimodal topics) from the large-scale multimodal data. For a product to be promoted, marketers can easily retrieve the related tags, textual topics, and images through the model results. This information makes it possible for marketers to quickly track market dynamics, and further determine which labels, textual descriptions, and pictures may improve their campaign

performance.

Second, online platforms can apply our results to actively monitor the multimodal contents, which is better gain the business points and optimize their search engine. For example, for sweaters, the platform can have close to perfect insight into which styles and designs are the most popular at present, and which clothes can match with products. This enables the platform to provide consumers with better services, not only recommending popular products for consumers, but also giving advice on outfit matching. In addition, our method can clearly build the relationship between tags and their textual contents and images. The online platforms take into account these results to optimize search engines to match target consumers' search interests.

5. Conclusion

In this study, we propose a new probabilistic topic model called the label-based multimodal topic (LB-MMT) model for the analysis of multimodal data in social media. The LB-MMT model integrates labels, textual and visual information into a unified framework to explore latent topics in data space. Different from other works, we use the labels as supervised signals to generate the text and image data. And we assume that the textual words and visual words related to each text and image are drawn from a mixture of latent multimodal topics. We conduct extensive experiments using a real-world dataset to validate the effectiveness of the proposed model. The results of the experiment demonstrate that the LB-MMT model outperforms all the baselines in quantitative evaluations, and yields several interesting insights regarding multimodal topics.

We close by highlighting broad areas for future research. First, future research may use deep learning methods (e.g., pre-trained VGG-16 CNNs) to extract the image features, and design a new model for fitting image data. Second, our proposed modeling framework does not consider the label irrelevant information in textual contents and images. As a future direction, we can introduce new latent variables to filter out the irrelevant information related to labels. Third, in this study, we focus on detecting multimodal topics based on social media data. Future research would be to build a new model that uses the multimodal topics learned by our model in other areas, e.g., predicting the popularity of contents on social media.

CRediT authorship contribution statement

Hao Li: Idea, Experiment. **Yang Qian**: Idea, Writing the manuscript. **Yuanchun Jiang**: Design of the study, Final proofreading. **Yezheng Liu**: Design of the study, Final proofreading. **Fan Zhou**: Experiment, Result analysis.

Declaration of Competing Interest

The authors certify that there is no conflict of interest in the subject matter discussed in the manuscript.

Data availability

Data will be made available on request.

Acknowledgment

We appreciate the constructive comments from the anonymous reviewers. This work is supported by the National Natural Science Foundation of China (72101072, 72171071, 91846201), the China Postdoctoral Science Foundation (2021M690852), the Fundamental Research Funds for the Central Universities (JZ2022HGTB0282, JZ2021HGQB0272), and the National Engineering Laboratory for Big Data Distribution and Exchange Technologies.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.dss.2022.113863.

References

[1] Y. Ma, J. Jia, S. Zhou, J. Fu, Y. Liu, Z. Tong, Towards better understanding the clothing fashion styles: A multimodal deep learning approach, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[2] A. Salah, Q.-T. Truong, H.W. Lauw, Cornac: a comparative framework for multimodal recommender systems, J. Mach. Learn. Res. 21 (2020) 91–95.
[3] Y. Zeng, D. Cao, X. Wei, M. Liu, Z. Zhao, Z. Qin, Multi-modal relational graph for cross-modal video moment retrieval, in, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 2215–2224.
[4] W. Zhang, W. Wang, J. Wang, H. Zha, User-guided hierarchical attention network for multi-modal social image popularity prediction, in: Proceedings of the 2018 World Wide Web Conference, 2018, pp. 1277–1286.
[5] Y. Ma, J. Jia, S. Zhou, J. Fu, Y. Liu, Z. Tong, Towards Better Understanding the Clothing Fashion Styles: A Multimodal Deep Learning Approach, 2017.
[6] K. Sohn, W. Shang, H. Lee, Improved multimodal deep learning with variation of information, in: Proceedings of the 27th International Conference on Neural Information Processing Systems Volume 2, 2014, pp. 2141–2149.
[7] N. Srivastava, R. Salakhutdinov, Learning representations for multimodal data with deep belief nets, in: International Conference on Machine Learning Workshop, 2012.
[8] Y.C. Chen, K.T. Lai, D. Liu, M.S. Chen, TAGNet: triplet-attention graph networks for hashtag recommendation, in: IEEE Transactions on Circuits and Systems for Video Technology, 2021.
[9] Y. Zheng, Y. Zhang, H. Larochelle, Topic modeling of multimodal data: An autoregressive approach, in: IEEE Conference on Computer Vision and Pattern Recognition 2014, 2014, pp. 1370–1377.
[10] D.M. Blei, M.I. Jordan, Modeling annotated data, in: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval, 2003, pp. 127–134.
[11] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learning Res. 3 (2003) 993–1022.
[12] D.M. Blei, J.D. McAuliffe, Supervised topic models, in: Proceedings of the 20th International Conference on Neural Information Processing Systems, 2007, pp. 121–128.
[13] S. Lacoste-Julien, F. Sha, M.I. Jordan, DiscLDA: discriminative learning for dimensionality reduction and classification, in: Proceedings of the 21st International Conference on Neural Information Processing Systems, 2008, pp. 897–904.
[14] J. Zhu, A. Ahmed, E.P. Xing, MedLDA: Maximum Margin Supervised Topic Models 13, 2012, pp. 2237–2278.
[15] D. Ramage, D. Hall, R. Nallapati, C.D. Manning, Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 Volume 1, 2009, pp. 248–256.
[16] D. Ramage, C.D. Manning, S. Dumais, Partially labeled topic models for interpretable text mining, in, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2011, pp. 457–465.
[17] P. Zhang, S. Wang, D. Li, X. Li, Z. Xu, Combine topic modeling with semantic embedding: embedding enhanced topic model, IEEE Trans. Knowl. Data Eng. 32 (2019) 2322–2335.
[18] H. Xu, W. Wang, W. Liu, L. Carin, Distilled wasserstein learning for word embedding and topic modeling, Adv. Neural Inf. Proces. Syst. 31 (2018).
[19] X. Chen, H. Chen, H. Xu, Y. Zhang, Y. Cao, Z. Qin, H. Zha, Personalized fashion recommendation with visual explanations based on multimodal attention network: Towards visually explainable recommendation, in: Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 2019, pp. 765–774.
[20] R. Ma, X. Qiu, Q. Zhang, X. Hu, Y.-G. Jiang, X. Huang, Co-attention memory network for multimodal microblog's hashtag recommendation, IEEE Trans. Knowl. Data Eng. 33 (2019) 388–400.
[21] J. Ni, Z. Huang, Y. Hu, C. Lin, A two-stage embedding model for recommendation with multimodal auxiliary information, Inf. Sci. 582 (2022) 22–37.
[22] J. Lv, W. Liu, M. Zhang, H. Gong, B. Wu, H. Ma, Multi-feature fusion for predicting social media popularity, in: Proceedings of the 25th ACM international conference on Multimedia, 2017, pp. 1883–1888.
[23] Y. Li, Y. Xie, Is a picture worth a thousand words? An empirical study of image content and social media engagement, J. Mark. Res. 57 (2020) 1–19.
[24] M. Kan, S. Shan, H. Zhang, S. Lao, X. Chen, Multi-view discriminant analysis, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2016) 188–194.
[25] X. Zhai, Y. Peng, J. Xiao, Learning cross-media joint representation with sparse and Semisupervised regularization, IEEE Transact. Circ. Syst. Video Technol. 24 (2014) 965–978.
[26] C. Deng, Z. Chen, X. Liu, X. Gao, D. Tao, Triplet-based deep hashing network for cross-modal retrieval, IEEE Trans. Image Process. 27 (2018) 3893–3903.
[27] P. Hu, D. Peng, Y. Sang, Y. Xiang, Multi-view linear discriminant analysis network, IEEE Trans. Image Process. 28 (2019) 5352–5365.

[28] Y. Qian, W. Xu, X. Liu, H. Ling, Y. Jiang, Y. Chai, Y. Liu, Popularity prediction for marketer-generated content: a text-guided attention neural network for multi-modal feature fusion, Inf. Process. Manag. 59 (2022), 102984.
[29] M. Zihayat, A. Ayanso, X. Zhao, H. Davoudi, A. An, A utility-based news recommendation system, Decis. Support. Syst. 117 (2019) 14–27.
[30] D. Slof, F. Frasincar, V. Matsiiako, A competing risks model based on latent Dirichlet allocation for predicting churn reasons, Decis. Support. Syst. 146 (2021), 113541.
[31] X. Cheng, X. Yan, Y. Lan, J. Guo, Btm: topic modeling over short texts, IEEE Trans. Knowl. Data Eng. 26 (2014) 2928–2941.
[32] L. Cao, L. Fei-Fei, Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes, in: 2007 IEEE 11th International Conference on Computer Vision, IEEE, 2007, pp. 1–8.
[33] D. Putthividhy, H.T. Attias, S.S. Nagarajan, Topic regression multi-modal Latent Dirichlet Allocation for image annotation, in: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2010, pp. 3408–3415.
[34] F. Xue, R. Hong, X. He, J. Wang, S. Qian, C. Xu, Knowledge-based topic model for multi-modal social event analysis, IEEE Transact. Multimedia 22 (2020) 2098–2110.
[35] Y. Yang, K. Zhang, Y. Fan, sDTM: a supervised Bayesian deep topic model for text analytics, Inf. Syst. Res. (2022) 1–20.
[36] X. Wang, Y. Yang, Neural topic model with attention for supervised learning, in: Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, 2020, pp. 1147–1156.
[37] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, Proceedings of NAACL-HLT (2019) 4171–4186.
[38] W. Liu, T. Xu, Q. Xu, J. Song, Y. Zu, An encoding strategy based word-character LSTM for Chinese NER, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 2379–2389.
[39] H. Kato, T. Harada, Image reconstruction from bag-of-visual-words, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 955–962.
[40] J. Yang, Y.-G. Jiang, A.G. Hauptmann, C.-W. Ngo, Evaluating bag-of-visual-words representations in scene classification, in: Proceedings of the International Workshop on Workshop on Multimedia Information Retrieval, 2007, pp. 197–206.
[41] B. Fernando, E. Fromont, D. Muselet, M. Sebban, Discriminative feature fusion for image classification, in: IEEE Conference on Computer Vision and Pattern Recognition 2012, 2012, pp. 3434–3441.
[42] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[43] Q. Wang, B. Li, P.V. Singh, Copycats vs. original mobile apps: a machine learning copycat-detection method and empirical analysis, Inf. Syst. Res. 29 (2018) 273–291.
[44] Z. Jiang, Y. Huang, D.R. Beil, The role of feedback in dynamic crowdsourcing contests: a structural empirical analysis, Manag. Sci. 68 (7) (2022) 4858–4877.
[45] X. Yang, L. Hou, Y. Zhou, W. Wang, J. Yan, Dense label encoding for boundary discontinuity free rotation detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15819–15829.
[46] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, A. Bovik, From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3575–3585.
[47] P. Majumdar, S. Mittal, R. Singh, M. Vatsa, Unravelling the effect of image distortions for biased prediction of pre-trained face recognition models, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 3786–3795.
[48] D. Ramage, D. Hall, R. Nallapati, C.D. Manning, Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora, in: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, 2009, pp. 248–256.
[49] R. Das, M. Zaheer, C. Dyer, Gaussian LDA for topic models with word embeddings, in, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015, pp. 795–804.
[50] E. Erosheva, S. Fienberg, J. Lafferty, Mixed-membership models of scientific publications, Proc. Natl. Acad. Sci. 101 (2004) 5220–5227.
[51] Z. Chen, B. Liu, Mining topics in documents: Standing on the shoulders of big data, in: Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2014, pp. 1116–1125.
[52] D. Mimno, H. Wallach, E. Talley, M. Leenders, A. McCallum, Optimizing semantic coherence in topic models, in: Proceedings of the 2011 Conference on Empirical Methods In Natural Language Processing, 2011, pp. 262–272.
[53] J. Zhu, A. Ahmed, E.P. Xing, MedLDA: maximum margin supervised topic models, J. Mach. Learning Res. 13 (2012) 2237–2278.
[54] Y. Feng, M. Lapata, Topic models for image annotation and text illustration, in: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 2010, pp. 831–839.
[55] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[56] R. Wang, B. Fu, G. Fu, M. Wang, Deep & cross network for ad click predictions, in: Proceedings of the ADKDD'17, 2017. Article 12.
[57] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014, pp. 1532–1543.
[58] T. Mikolov, I. Sutskever, K. Chen, G.S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, Adv. Neural Inf. Proces. Syst. 26 (2013).
[59] E. Rublee, V. Rabaud, K. Konolige, G. Bradski, ORB: An efficient alternative to SIFT or SURF, in: 2011 International Conference on Computer Vision, 2022.
[60] H. Bay, T. Tuytelaars, L.V. Gool, Surf: Speeded up robust features, in: European Conference on Computer Vision, Springer, 2006, pp. 404–417.
[61] E. Karami, S. Prasad, M. Shehata, Image matching using SIFT, SURF, BRIEF and ORB: performance comparison for distorted images, The 24th Annual Newfoundland Electrical and Computer Engineering Conference, NECEC (2015) arXiv:1710.02726.

Hao Li is working toward a Master's degree in Management Science and Engineering from the Hefei University of Technology. He received his bachelor's degree from Hefei University of Technology. His research interests include machine learning and data mining, and visual marketing.

Yang Qian is a postdoctoral fellow at the School of Management, Hefei University of Technology. He received his Ph.D. in Management Science and Engineering from Hefei University of Technology in 2020. His research interests include electronic commerce, online marketing, and machine learning. His work has appeared in journals including *European Journal of Operational Research*, *ACM Transactions on Knowledge Discovery from Data, Information Processing & Management*, and *World Wide Web*.

Yuanchun Jiang is a professor at School of Management, Hefei University of Technology, China. He received his Ph.D. in Management Science and Engineering from Hefei University of Technology, Hefei, China. He teaches electronic commerce, business intelligence and business research methods. His research interests include online marketing, electronic commerce and data mining. He has published papers in journals such as *Marketing Science*, *European Journal of Operational Research*, *Decision Support Systems*, and *IEEE Transactions on Dependable and Secure Computing*.

Yezheng Liu is a professor of Electronic Commerce at Hefei University of Technology, China. He received his Ph.D. in Management Science and Engineering from Hefei University of Technology in 2001. His main research interests include decision science, electronic commerce, intelligent decision support systems and data mining. His work has appeared in journals including *Marketing Science*, *IEEE Transaction on Software Engineering*, *Information Sciences*, and *ACM Transactions on Information Systems*.

Fan Zhou is working toward a Master's degree in Management Science and Engineering from the Hefei University of Technology. She received her bachelor's degree from Hefei University of Technology. Her research interests include machine learning and text mining.