Knowledge-Based Systems 195 (2020) 105695

Contents lists available at ScienceDirect

Knowledge-Based Systems

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

A survey on multi-modal social event detection[☆]

Han Zhou [a,b], Hongpeng Yin [a,b,*], Hengyi Zheng [c], Yanxia Li [a,b]

[a] College of Automation, Chongqing University, Chongqing, China
[b] Key Laboratory of Complex System Safety and Control, Ministry of Education, Chongqing, China
[c] College of Mechanical Engineering, Chongqing University, Chongqing, China

ARTICLE INFO

Article history:
Received 24 October 2019
Received in revised form 21 February 2020
Accepted 22 February 2020
Available online 27 February 2020

Keywords:
Multi-modality
Social event detection
Feature learning
Event inference

ABSTRACT

Due to the prevalence of social media sites, users are allowed to conveniently share their ideas and activities anytime and anywhere. Therefore, these sites hold substantial real-world event related data. Different from traditional social event detection methods which mainly focus on single-media, multi-modal social event detection aims at discovering events in vast heterogeneous data such as texts, images and video clips. These data denote real-world events from multiple dimensions simultaneously so that they can provide comprehensive and complementary understanding of social event. In recent years, multi-modal social event detection has attracted intensive attentions. This paper concentrates on conducting a comprehensive survey of extant works. Two current attempts in this field are firstly reviewed: event feature learning and event inference. Particularly, event feature learning is a prerequisite because of its ability on translating social media data into computer-friendly numerical form. Event inference aims at deciding whether a sample belongs to a social event. Then, several public datasets in the community are introduced and the comparison results are also provided. At the end of this paper, a general discussion of the insights is delivered to promote the development of multi-modal social event detection.

1. Introduction

Social media sites, such as Twitter,[1] Facebook[2] and Sina Weibo,[3] are broadcast platforms that users are allowed to conveniently share texts, images or videos anytime and anywhere. People witnessing or involved in any social events can disseminate information on these platforms. Consequently, social media sites hold amount of valuable real-life data. Meanwhile, vast social data can be easily accessed via the web-clawers and public API. These facts motivate researchers to develop a wide range of impressive social applications, such as business marketing [1], sentiment analysis [2,3], advertisements recommendation [4] and social event detection [5,6] etc. In the above applications, social event detection task attracts many attentions because it can yield valuable information for helping the Internet users capture and understand what is happening around the world.

Social event detection focuses on mining real-world occurrences in unprecedentedly vast social media data. It may date back to the topic detection and tracking (TDT) project [7], which tries to discover events in continuous well-formed texts stream. Although these single-media focused works have achieved satisfied results in this project, they are difficult to handle current situation because social media platforms usually contain considerable multi-modal data. As illustrated in Fig. 1, a post usually includes texts, images, audio as well as other meta-data. Compared with the limited information capacity of single-media, multi-modal media can show real-world events from multiple dimensions (textual, visual audio) and thus provide more comprehensive and complementary information. Therefore, social event detection research began to shift from single-media type to multi-modal types.

Multi-modal social event detection is challenging because there is the "media gap" among different modalities, which means that the descriptions of different media types are inconsistent and cannot be measured directly. In recent years, researchers have devoted many efforts on this topic and many methods have been proposed. After reviewed several references, we classified existing efforts into two aspects i.e. event feature learning and event inference. Feature learning is the foundation of multi-modal

☆ No author associated with this paper has disclosed any potential or pertinent conflicts which may be perceived to have impending conflict with this work. For full disclosure statements refer to https://doi.org/10.1016/j.knosys.2020.105695.

* Corresponding author at: College of Automation, Chongqing University, Chongqing, China.
E-mail address: yinhongpeng@gmail.com (H. Yin).

1 http://www.twitter.com
2 https://www.facebook.com
3 https://weibo.com

https://doi.org/10.1016/j.knosys.2020.105695
0950-7051/

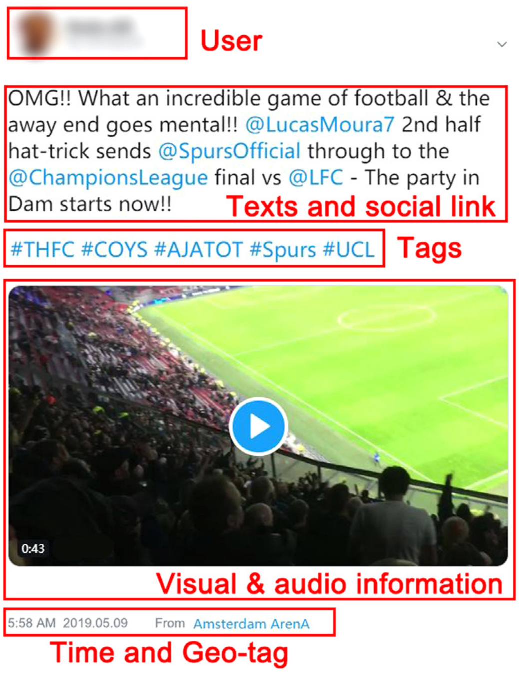

Fig. 1. An example of multi-modal social data.

social event detection, which aims at extracting distinguishing features from media sequences or collections. It is also the usual stage to solve the heterogeneity problem. Further, event inference is a key stage to decide whether a message or a post belongs to a social event or not. The effectiveness of event detection system depends heavily on this step.

In recent years, research community has conducted amount of related survey studies on social event detection, as summarized in Table 1. Some surveys focused relevant works on a specific social platform. For example, Atefeh et al. [8] mainly reviewed social event detection techniques for Twitter streams. Petkos et al. [9] mainly summarized several detecting approaches on Flickr images.[4] Goswami et al. [10] individually surveyed event detection techniques based on different online social platforms such as newswire, web forums. Meanwhile, many works only consider social event happened with single media type, especially the textual data [11–13]. Although Tzelepis et al. [14] reached the issue of event detection in different media types, they treated the media type individually. Therefore, their work cannot be considered as multi-modal event detection. As for current surveys about multi-modal event detection, Matthias et al. [15] provided extensive experimental results of social event classification tasks depending on textual, visual as well as multi-modal representations. However, the inference techniques are seldom discussed in their paper. Zhou et al. [16] mainly surveyed topic modeling based methods on topic analysis in both text and multi-modal corpora. Liu et al. [17] provided a comprehensive review on feature learning and event inference technologies. However, the type of social events is not well specified in their work so that they fail to provide clear thoughts to make the readers understand the motivation of categorization of current detection methods.

Different from the extant surveys, a comprehensive analysis about multi-modal social event detection is conducted in this paper. This paper starts with event features leaning technologies, where both single-media and multi-media based works are surveyed. Further, the event inference methods are grouped according to the social events attribute. Specifically, they are discussed by (1) specific/unspecific social events, (2) new/retrospective social events. Meanwhile, several public datasets are introduced and the results comparison is provided. At the end of this paper, a general discussion is delivered to promote the development of multi-modal social event detection. The rest of this paper is organized as follows: event feature learning methods are presented in Section 2. Section 3 discusses several event inference methods. The public datasets and results comparison are provided in Section 4. Section 5 gives conclusion and general discussion.

[4] https://blog.flickr.net

2. Event feature learning

Event feature learning is a pre-requisite for social event detection. It aims at extracting features from social media data, translating data into numerical form, which can be easily understood by computers. Good features can not only capture distinguishing characteristics of social data, improving event detection performance, but also economically reduce the computation costs.

2.1. Single-modality feature learning

Single-modality understanding is the foundation of multi-modality feature learning. Since many surveys have been conducted in this field [18–22], this section only presents a summary of widely used methods. Depending on the specific media type, different representations can be extracted utilizing various methodologies. Most of the social media contents belong to one or more of the following aspects: (a) textual information, including sentences and tags, (b) visual information, including images and videos, (c) audio information.

2.1.1. Textual information

Text information based methods, in the majority of related references, mainly adopt Natural Language Processing (NLP) techniques to learn features. A simple yet efficient model is Vector Space Model (VSM), which represents a document as a vector of terms and term weight. It intends to reflect how important a word is to a document [23]. However, VSM totally ignores words order since it extracts text features from the perspective of numerical statistics. Accordingly, VSM model may lack of contextual information. To avoid this problem, topic modeling can be adopted, such as Latent Semantic Indexing (LSI) [24], Probabilistic Latent Semantic Indexing (PLSI) [25], Latent Dirichlet Allocation (LDA) [26,27]. In topic modeling, assuming that topics are semantically related clusters of word, texts are explicitly represented from the semantic level. In recent years, deep learning also achieved satisfied learning performance where word embedding is the pre-requisite. Different from the conventional VSM model where each word corresponds to a feature space, word embedding tries to learn a uniform word feature space where words sharing similar meaning are as close as possible [28,29]. In this situation, it can avoid the problem of sparseness and dimensionality curse. Based on this idea, Convolutional neural network (CNN) [30,31], Deep Belief Networks (DBN) [32] are popular tools to learn the inherent features of textual documents. Recurrent Neural networks (RNN), Long Short Term Memory networks (LSTM) are more natural to handle the sequential social text stream because it can capture long-term dependencies [33,34].

Table 1
A summary of the related studies and their main ideas.

Ref.	Published time	Modality	Media type	Main ideas	Characteristics
[8]	2015	Single	Textual	• Reviewed social event detection techniques on Twitter streams and categorized these techniques by event type, detection task.	
[9]	2014	Multi	Textual, visual	• Summarized the set of social events detection approaches on MediaEval benchmark challenge which includes more than 100K Flickr images.	Platform-targeted
[10]	2016	Single	Textual	• Surveyed event detection techniques based on different online social platforms. Further, reviewed popular detection tools in online social networks.	
[11]	2016	Single	Textual	• Studied path-breaking approaches on detecting uncertain and outbreak social events (e.g. natural disasters, contagious disease spreading) in textual social media data.	
[12]	2016	Single	Textual	• Defined the type of social events at first and discussed the popular detection methods based on the defined event type.	
[13]	2016	Single	Textual	• Viewed the social event detection as clustering problem and discussed different clustering based event detection methods on textual data.	Single-modality targeted
[14]	2016	Single	Textual, visual, audio	• Reached the issue of event detection in different media types, i.e. audio-, video-, textual-based, but they treated the media type individually according to feature representations.	
[15]	2016	Both	Textual, visual	• Provided extensive experimental results of social event classification tasks depending on textual, visual as well as multi-modal representations.	
[16]	2017	Both	Textual, visual	• Surveyed the existing topic modeling based methods on event analysis in both text and multi-modal corpora.	Multi-modality targeted
[17]	2019	Multi	Textual, visual	• Summarized representative works in multimedia social event analysis from several aspects, including multi-modal feature learning and social event detection.	

2.1.2. Visual information

Visual information based methods rely on methodologies derived from computer vision field. Traditionally, images are usually represented in three perceptual categories, including color, texture as well as shape [21]. Color features, such as works [35,36], typically reflect the color distribution of an image so that they have the inherent ability on maintaining strong information of human perception. Texture features capture the surface information (structure or random) over homogeneous contents of an image. Popular tools for extracting texture features include filter based methods and statistical models [37,38]. Shape features usually describe the geometric characteristics in an image. The main concern in reliably obtaining shape features is the requirement of rigid interesting objects, such as pedestrian [39]. As for videos, one more factor, motion, needs to be considered since they can be viewed as image sequence. Although these traditional features have obtained satisfied performance in several fields, how to select features under different scenarios remains as a challenge. In practice, it usually depends on prior knowledge and it is chosen empirically.

Recently, deep neural networks have been widely applied due to their powerful ability on feature learning. Since they can be designed as end-to-end models, the essential and accurate image/video representations can be extracted without empirical feature selection. Popular methods include Auto-Encoder (AE) [40, 41], Convolutional Neural Network (CNN) [42], Generative Adversarial Networks (GAN) [43] etc. AE is a three-layer network including an encoder and a decoder. The encoder maps the input data into codes, from the original space to a latent space, and the decoder reconstructs the input data from the corresponding codes. It can learn good features from unlabeled data for higher-level tasks. For example, Yin et al. [40] constructed sparse AE to learn features of scene images and group them into semantic categories. Song et al. [41] designed a hierarchical binary AE to model the temporal dependencies in videos and achieved accurate video retrieval. Different from the simple architecture of AE, a CNN can have several layers where filter layers are applied to each training image and the output of each convolved image is used as the input to the next layer. The CNN usually start from simple features and gradually reach complex features that ultimately define the image. Krizhevsky et al. [42] early trained a deep convolutional neural network (AlexNet) to classify large-scale images. Other complicated models are further proposed, such as VGGnet [44], GoogleNet [45], and the performance is considerably better than the previous state-of-the-art. CNN is also widely adopted in the GAN. This results in a standardized GAN structure, called Deep Convolutional Generative Adversarial Network (DCGAN), a more stable model of GAN [46]. Most GANs today are at least loosely based on the DCGAN architecture. The idea of DCGAN is based on a game theoretic scenario in which the generator network must compete against an adversary. The generator network produces samples. Its adversary, the discriminator network, tries to distinguish samples drawn from real domain or the generator. The successful application shows its powerful ability to generate realistic data, most notably in images.

2.1.3. Audio information

General audio can be described in different perceptual categories. From low-level features such as acoustic, phonotactic, prosodic information, to high-level features such as morphology and syntax [22].

Acoustic features are usually considered as the first level of audio analysis because they can be directly extracted from raw signal. They distinguish different speech events according to amplitude and frequency components of speech waves. Zero-Crossing Rate (ZCR), Perceptual Linear Prediction (PLP), Prediction Cepstral Coefficient (PCC) and Mel Frequency Cepstral Coefficient (MFCC) are popular choices for acoustic features [47,48]. Kucukbay et al. [49] adopted MFCC to detect audio event detection in office environment. To distinguish noisy flat voice, they utilized different window and hop sizes by changing the number of Mel coefficients in the analyses. Xu et al. [50] detected whistling in soccer match by ZCR but they found that individual feature was not robust to noise. Therefore, they combined several acoustic features and achieved better detection results.

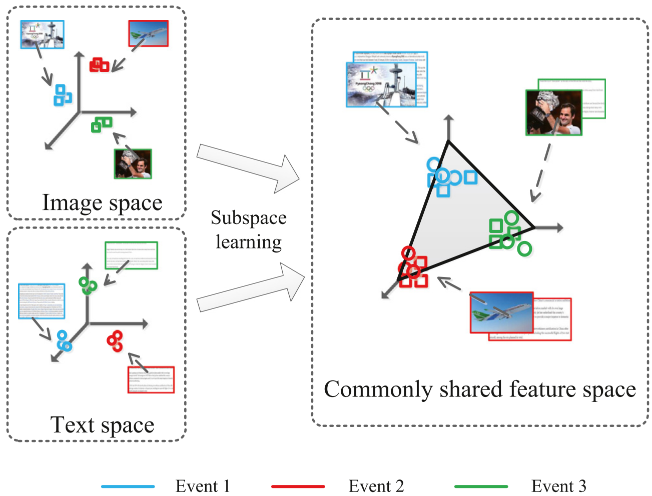

Fig. 2. A brief illustration of commonly shared feature space learning method for multi-modal data (considering the images and texts as examples).

Phonology deals with a set of physically produced sounds while Prosody is concerned with the study of tone, stress, rhythm in audio. They can be viewed as middle-level feature of audio information. Phonology and prosody carries language discriminative information and they are widely used in language identification. For example, work [51] pointed that the phoneme /st/ is very common in English, whereas it is not existed in Japanese. The pitch variations in tone are often used to identify Mandarin Chinese, where the tone of a word determines its meaning [52].

Works focusing on learning high-level features try to reflect the specific traits of communication such as the internal words relationship and sentence structure [53,54], also called morphology and syntax. In other words, they focus on the meaningful contents, leading to improvements in audio content understanding. Due to the non-vulnerability to acoustic variability of speech, the high-level features are more robust to noise. However, not many social event detection works make use of morphological and syntactic information. It is not essential that an audio event detection method uses all levels of features.

2.2. Multi-modality feature learning

Many works have proved that properly integrating multi-modal features can improve the accuracy of data analysis [55,56]. Similarly, effective multi-modal media analysis is able to boost comprehensive and meaningful knowledge discovery from social media data. On the contrary, if modalities are not aggregated appropriately, the learned representation may even degrade the social event detection performance. Since the characteristics of multi-modal data is different from that of single-modal data, feature learning methods within multi-modal data call for new approaches.

At present, there has been a great deal of works simultaneously learning the multiple media types for social event detection. As shown in Fig. 2, their basic idea aims at jointly projecting heterogeneous features of multiple types media into a commonly shared space by linear or nonlinear functions. This jointly projection strategy can fully leverage abundant information of multi-modal data and ensure that the similarity among different modality data can be measured directly. Through investigating existing multi-modal feature learning works, we divide the current methods into following major categories.

2.2.1. Correlation based methods

Correlation based methods are the basic paradigm and foundation of multi-modality representation learning. They usually project heterogeneous data by optimizing statistical correlation among different modality. Canonical Correlation Analysis (CCA) [57] is a classical approach introduced in [58,59]. Give a social media datasets $\mathbf{D}$ including N samples $\mathbf{D} = \{\mathbf{d}_1, \ldots, \mathbf{d}_i, \ldots, \mathbf{d}_N\}$. Each sample contains two modalities, usually texts and images $\mathbf{d}_i = \{(\mathbf{T}_i, \mathbf{I}_i)\}$, where $\mathbf{T} \in \Re^{d_T \times N}$ and $\mathbf{I} \in \Re^{d_I \times N}$. CCA learns two linear projection matrices $\boldsymbol{\omega}_\mathbf{T}$ and $\boldsymbol{\omega}_\mathbf{I}$, ensuring that the projected matrices $\boldsymbol{\omega}_\mathbf{T}^\mathbf{T}\mathbf{T}$ and $\boldsymbol{\omega}_\mathbf{I}^\mathbf{T}\mathbf{I}$ share the largest correlation:

$$\rho = \max_{\omega_T,\omega_I} corr\left(\boldsymbol{\omega}_\mathbf{T}^\mathbf{T}\mathbf{T}, \boldsymbol{\omega}_\mathbf{I}^\mathbf{T}\mathbf{I}\right) \tag{1}$$

Generally:

$$corr(\cdot) = \max_{\omega_T,\omega_I} \frac{\boldsymbol{\omega}_\mathbf{T}^\mathbf{T} \sum_\mathbf{TI} \boldsymbol{\omega}_\mathbf{I}^\mathbf{T}}{\sqrt{\boldsymbol{\omega}_\mathbf{T}^\mathbf{T} \sum_\mathbf{TT} \boldsymbol{\omega}_\mathbf{T}^\mathbf{T}} \sqrt{\boldsymbol{\omega}_\mathbf{I}^\mathbf{T} \sum_{II} \boldsymbol{\omega}_\mathbf{I}^\mathbf{T}}} \tag{2}$$

Where $\sum_\mathbf{TI}$ denotes the cross-covariance matrix between modality T and I while $\sum_\mathbf{TT}$, $\sum_\mathbf{II}$ represent the empirical covariance matrices for the two modalities respectively. The basis of projected matrices $\boldsymbol{\omega}_\mathbf{T}^\mathbf{T}\mathbf{T}$ and $\boldsymbol{\omega}_\mathbf{I}^\mathbf{T}\mathbf{I}$ can be applied to obtain the uniform multi-modality representation. Suppose that $\mathbf{T}_\mathbf{i}^\mathbf{p}$ and $\mathbf{I}_\mathbf{i}^\mathbf{p}$ are two projected representations for modality T and I respectively, simple operations for data fusion are as follow.

Sum operation: $\mathbf{H}^{sum} = \mathbf{T}_i^p + \mathbf{I}_i^p$.

Max operation: $\mathbf{H}^{max} = max\{\mathbf{T}_i^p, \mathbf{I}_i^p\}$.

Concatenation operation: $\mathbf{H}^{concat} = [\mathbf{T}_i^p; \mathbf{I}_i^p]$.

With adopting CCA, Wu et al. [58] analyzed the correlation between images and audios while Rasiwasia et al. [59] and Pereira et al. [60] focused on analyzing the correlation between texts and images. CCA is also extended into nonlinear manner to fully capture the complex correlations between two modalities, such as kernel Canonical Correlation Analysis (KCCA) [61] and Deep Canonical Correlation Analysis (DCCA) [62–64]. Work [61] transformed original data into the Hilbert space by kernel function and further find a feature that maximizes the correlation coefficients in the Hilbert spaces. Andrew et al. [62] proposed the DCCA where two separate deep encoders learned the maximum correlated encodings of two modalities. Huang et al. [64] also bridged textual and visual information of social media by nonlinear CCA. Particularly, the textual features are pre-encoded by LSTM and the visual features are pre-encoded by multi-layer perceptron with a residual link. Their work can be viewed as a variant of DCCA. In practice, DCCA is more flexible than KCCA since DCCA does not require the empirical selection on kernel functions. However, most of the existing CCA based approaches only focus on analyzing two kind of media type while event-related date are usually more than two media types in real life.

2.2.2. Matrix decomposition based methods

The basic idea of matrix decomposition is to decompose a matrix $\mathbf{X}$ into the production of two sub-matrices:

$$\mathbf{X} = \mathbf{U}\mathbf{V}^\mathbf{T} \tag{3}$$

This idea can be easily extended into multi-modal representation learning. Representative works include Principal Component Analysis (PCA) [65], Singular Value Decomposition (SVD) [66], Non-negative Matrix Factorization (NMF) [67–72], Dictionary Learning (DL) [73–76]. Taking work [72] as an example, for a dataset including M modalities $\mathbf{D} = \{\mathbf{X}^1, \ldots, \mathbf{X}^m, \ldots, \mathbf{X}^M\}$, there are N samples with d_m feature dimensions, $\mathbf{X}^m = \{\mathbf{x}_1^m, \ldots, \mathbf{x}_n^m, \ldots, \mathbf{x}_N^m\} \in \Re^{d_m \times N}$. The common approach is to jointly minimize the squared error loss of following objective function:

$$\min_{\mathbf{U},\mathbf{V}} \sum_{m=1}^{M} \left\| \mathbf{X}^m - \mathbf{U}^m (\mathbf{V})^T \right\|_F^2 + \Omega \tag{4}$$

Where $\mathbf{U}^m \in \Re^{d_m \times K}$ is the basis matrix of each modality and $\mathbf{V} \in \Re^{N \times K}$ is a uniform consensus matrix which encodes the consistency content of all modalities. K is an empirical parameter deciding the dimension of subspace. Ω is an optional regularization to regularize the correlation learning process, such as semi-supervised regularization [70], sparseness regularization [72]. From the above objective function, we can see that heterogeneous matrix $\mathbf{X}^m$ can be represented by a shared coefficient matrix $\mathbf{V}$. Namely, the same features can describe multiple types of social media so that media difference can be measured directly. Unlike CCA, this model can be easily extended into multi-modal scenarios regardless of the number of modality. For example, Huang et al. [68] introduced modality *time* to analyze sequential data and further discover events evolutionary trends. Since some side information could provide guidance for event detection, Xue et al. [70] extended this model into a semi-supervised manner by imposing an additional label embedding constraint. The constraints encourage that data labeled to the same event are made tight and the unlabeled data should be close to the labeled data in subspace. However, above works only promised the multi-modality consistency because features of different modality are encouraged to be same. Some works also believe that the complementarity should be also jointly investigated in the multi-modal learning problem since each modality may compensate the missing information that other modalities do not have. Therefore, Gupta et al. [67,69] exploited both the consistency and complementary by dividing basis matrix into common part and specific part. With tuning the number of shared basis vectors, their methods explicitly controlled the level of subspace sharing from none to full.

2.2.3. Graph learning based methods

Graph learning based methods usually utilize the basic idea of graph Laplacian, where the vertices denote media and the edge weights are determined by their similarity. Content similarities, relationships and semantic category labels also can be adopted for graph construction. For example, Chu et al. [77] considered the both temporal similarity and content similarity as the weights since time attribute plays an important role in social event detection. Generally, graph $\mathbf{G^m}$ for single type media is constructed individually at first:

$$\mathbf{G}_{ij}^m = \begin{cases} s_{ij}^m & x_j^m \in \mathbf{N}_i^m \\ 0 & otherwise \end{cases} \tag{5}$$

Where s_{ij}^m is the similarity between sample x_i^m and x_j^m in modality m. $\mathbf{N}_i^m$ is the set of nearest neighbors of sample x_i^m. Popular choices for the similarity measurement include 0–1 strategy, heat kernel strategy and dot product strategy.

• 0–1 strategy: $s_{ij}^m = 1$ if x_j^m belongs to the set of nearest neighbors of sample x_i^m.

• Heat kernel strategy: $s_{ij}^m = e^{(\|x_i^m - x_j^m\|^2)/\sigma^2}$ if sample x_i^m and x_j^m are connected in the graph.

• Dot product strategy: $s_{ij}^m = x_i^m x_j^{mT}$ if sample x_i^m and x_j^m are connected in the graph.

After obtaining the single-modal graph, a feasible strategy to learn the multi-modal features is co-training [78,79], where the optimization problem can be simplified as follow:

$$\max_{\mathbf{U}^1,\ldots,\mathbf{U}^M} \sum_{m=1}^{M} (\mathbf{U^m}^T \mathbf{L^m} \mathbf{U^m}) + \Omega \tag{6}$$

Where $\mathbf{L^m} = \mathbf{D^m}^{-1/2} \mathbf{G^m} \mathbf{D^m}^{-1/2}$ is the graph Laplacian of each modality. $\mathbf{D^m}$ is a diagonal matrix whose element is the sum of the ith row of $\mathbf{G}$, i.e. $\mathbf{D^m}(i,i) = \sum_j \mathbf{G^m}(i,j)$. $\mathbf{U^m}$ is the eigenvector matrix corresponding to graph Laplacian $\mathbf{L^m}$. Generally, an optional co-regularization term Ω will be imposed to enforce $\mathbf{U}^m$ of different modalities be close, such as centroid-based regularization [78], pairwise regularization [79].

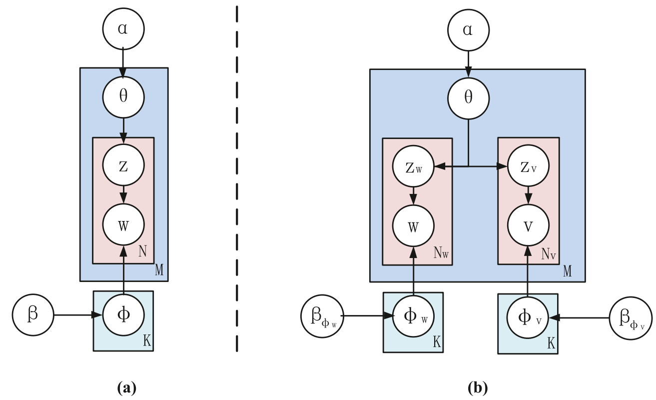

Fig. 3. The graphical model representation of (a) LDA; (b) multi-modal LDA.

Different from the co-training strategy which treats single-modal graphs individually, some works try graph fusion operation to obtain a uniform multi-modal graph $\mathbf{G}$. Popular strategy is fusing single-modal graphs by means of a weighted summation [80]. Tong et al. [81] also proposed linear fusion and sequential fusion strategies. Ma et al. [66] adopted logical OR strategy on single-modal graphs. The OR logical operation can appropriately compensate the missing correspondence when a social media misses a correspondence modality. After obtaining the multi-modal graph, matrix reduction methods can be adopted to obtain the uniform representation of multi-modal data, such as SVD [66], Laplacian Eigenmaps (LE) [82]. These methods are usually applicable when all the social media data are available.

2.2.4. Topic model based methods

Classical probabilistic models [26,83], also can be called topic models, mainly focus on detecting events in textual documents where word occurrences are utilized to model topics as a mixture of words and documents as a mixture of topics. The model can be depicted in Fig. 3a. α and β is the parameters of Dirichlet distribution. z is the topic assignments for text. w is the textual word. θ is the distribution of topic specific to a document. ϕ is the distribution of word specific to a topic. K is the number of topics; M is the number of documents. N is the number of words in a document. The model is usually tackled as a problem of Bayesian inference solved by Gibbs sampling. However, the basic assumption of topic model is that a document contains more than one topic so that it is possible to mine many-to-many correlation between long documents and events. As for social media sites, such as Twitter and Weibo, traditional topic models may not be a perfect choice because they limit their posts in 150 words. Meanwhile, due the fact that the content of social media data is not limited to texts, topic models need to take into account multi-modal content (mmLDA).

To tackle these problems, several methods try to extend conventional topic models to be aligned with the social media data [84–87]. In this situation, both words and other modalities are considered as observable variables. Take works [84,85] as example, the model is depicted in Fig. 3b where w and v are textual word and visual word respectively. Cai et al. [87] only assigned each tweet to one topic and each topic has a mixture of four Twitter features distributions: a multinomial hashtag distribution, a multinomial distribution over specific words, a Beta timestamp distribution, and a Dirichlet distribution over specific

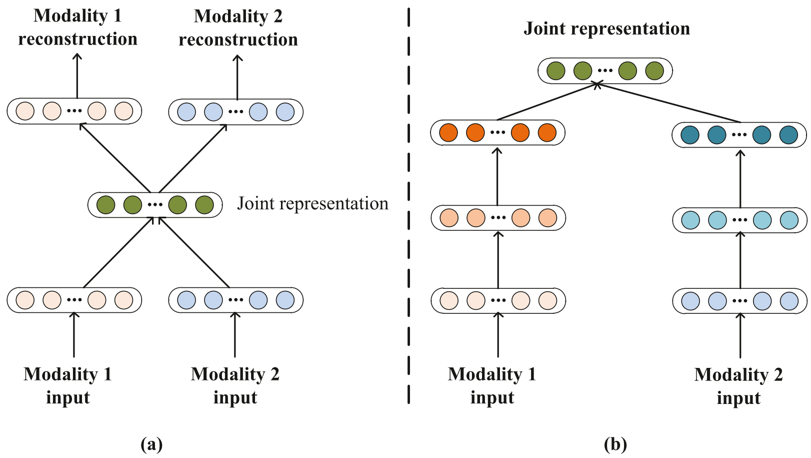

Fig. 4. The layer-shared architecture of neural network based methods.

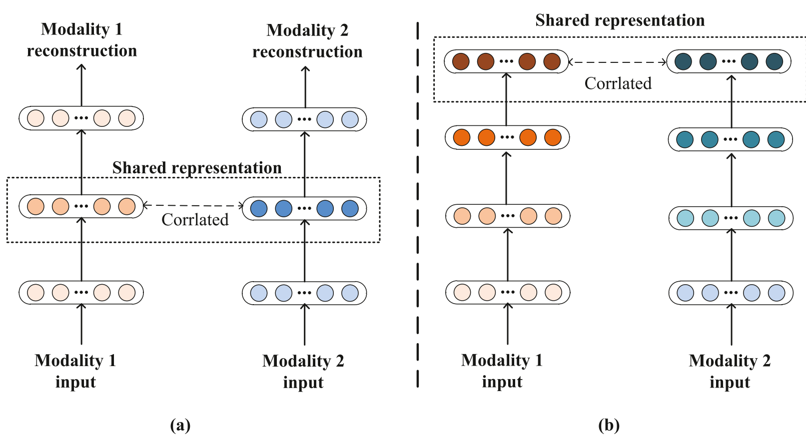

Fig. 5. The correlation regularized architecture of neural network based methods.

images. To boost the event detection performance, Qian et al. [88–90] extended the mmLDA into supervised manner by leveraged category labels. Qian et al. [91] also proposed multimodal event topic model (mmETM) to incrementally model multi-modal information. Particularly, the multi-modal topic model is sequentially updated at each epoch using the newly obtained event documents and the parameters of the previous epoch. In this situation, their model can work in an online mode to obtain evolutionary trends of social events.

2.2.5. *Neural networks based methods*

A. Unsupervised manner

Inspired by recent advance of neural networks, a variety of neural networks based methods have been proposed to learn multi-modal features. Their basic idea is to adopt structure-shared multiple networks to capture nonlinear correlation of multi-modality data. Generally speaking, according to the sharing strategy, the overall structures can be divided into two categories. Each sub-network can be associated by a sharing layer (Fig. 4) or correlation regularization (Fig. 5).

Typical examples of the first strategy include works [92–94]. Srivastava et al. [92] developed two layer-shared deep Restricted Boltzmann Machines (RBMs) to model joint distribution over texts and images. Due to the essential feature learned by frontier layers, it can be easier for the latter layers to learn higher-order correlations across the modalities. Similar idea also can be found in work [93]. Ngiam et al. adopted RBMs to learn higher-order correlations of audio-visual representations but they pointed that this model is not robust to missing modality. To avoid this problem, they further proposed a multi-modal auto-encoder, as illustrated in Fig. 4a. Due to the initialization using sparse RBMs, the units could have low expected activation. In this situation, when one of the modalities is set to zero, the model can deal with the incomplete modality. Guo et al. [94] also proposed a two-step method for this model when dealing with incomplete multi-modal social data. Their first step utilizes the data with all modalities to learn a pre-trained model. Further, the incomplete modalities are trained with ignoring corresponding reconstruction error. It can be assumed that the reconstruction for incomplete data is accurate because cross modality correlation has been learned in the first step.

Although the layer-shared structure is simple to interpret, it might cause the unfeasible units. Namely, some units are tuned only for one specific modality while others are tuned for other modality [93]. Therefore, some works soften this strategy and associate sub-networks with correlation regularizations. As shown in Fig. 5, sub-networks are combined by different correlation regularizations so that it is more feasible for application. Operation mentioned in Section 2.2.1 can be adopted for further fusing. Typical example of Fig. 5a includes works [95–97]. Feng et al. [95,96] adopted a simple correlation regularization, i.e. Euclidean distance. The intuition is that semantically similar multi-modal data will be close in the embedding space, and the distance similarity will be kept. Inspired by CCA, Wang et al. [98] presented deep canonically correlated auto-encoders (DCCAE). The main idea is to utilize CCA to associate multi-modal data. Some works also adopted multi-modal deep brief networks and kept the likelihood of multi-modal data [99]. Its architecture is illustrated in Fig. 5b.

B. Supervised manner

Except for the above non-convolutional works, CNN based methods have also been proposed [100,101]. Wei et al. [101] adopted CNN for visual data and full-connected network for textual data. Since the outputs of their neural network are the intrinsic probability distribution over the class labels for image or text, the heterogeneous data are naturally mapped into a commonly shared subspace. Note that, social events usually happen in natural user-posted content in real applications, it is crucial to train the network from zero, especially for social image. A feasible solution could be pre-training. Both works [100,101] pre-trained their CNN by a large-scale image dataset, ImageNet.[5] Since the pre-trained convolutional layers encode valuable and general lower-level visual patterns acquired from the large-scale datasets, their model could achieve faster convergence and better accuracy.

However, above works take handcrafted features as inputs so that the process of selecting features plays an important role in these methods. Some researchers also focus on designing a deep architecture, which can take raw documents as inputs. He et al. [102] individually adopted two types of CNNs to extracting the inner features for images and texts and correlated images and texts with cosine similarity. Specially, the CNN for texts involves word embedding learning, which has been proved effective on text classification [103]. In Hong's work [104], CNN based feature mapping task and multitask learning task are associated to obtain the multi-modal features. Particularly, they improved convolutional layers with manifold regularization. This strategy can keep locality properties of neurons and learn better features. Further, a sparse and low-rank regression model integrates the learned features of each single-modality. Different from their work where the learning step and fusion step are separated, Gao et al. [105] designed a unified two-pathway neural network for social images and texts. Specially, the maximum output neuron activations strategy is chosen to combine each pathway. Namely, for each neuron of the integration layer, its value is determined by the maximum of corresponding image neuron and text neuron. In this situation, this integration layer can infer the most confident activation with respect to the target social event. This network can be viewed as an end-to-end architecture, i.e. taking raw data as inputs and taking event detection results as output.

2.2.6. Hashing based methods

Features extracted by most works are real-valued. However, both the immense storage space and long computational time requirements of real-valued features become the major problems within the social event detection. In the view of this fact, hashing based methods are a practical ways to tackle these issues because they enforce messages to be represented by binary hashing codes. Since some works view event detection as retrieving tasks, hashing is particularly applicable for fast search services of social media [106–108]. Hashing based methods obtain binary features of multi-modal data by projecting different modalities into a commonly shared Hamming space [109–112]. Their basic idea can be formulated as follow:

$$\mathbf{B} = \frac{1}{2}[1 + sgn(\omega_m^T \mathbf{X}^m)] \tag{7}$$

Where $\mathbf{X^m}$ is the feature matrix of modality m and $\boldsymbol{\omega}_\mathbf{m}^\mathbf{T}$ is the projection matrix to map the original data into the Hamming space. $\mathbf{B}$ encodes the shared information with binary elements. For example, by adding additional regularizations, Song et al. [109] explored both inter-modal consistency and intra-modal consistency to ensure that distance between multi-modal data points can be efficiently measured. Specifically, the inter-modal consistency is achieved by leveraging labeled data under the assumption that inter-data points should have similar representation with their associated labels. On the other hand, the intra-modal consistency is ensured by introducing affinity matrices where similar intra-data points share larger similarity. Zhu et al. [110] added graph regularization to preserve the geometrical of data and bit-uncorrelated constraint to guarantee the orthogonality of the learned Hamming space.

2.2.7. Others methods

There are still some approaches, which cannot be easily classified into above categories. For example, considering the temporal attribute of social media streams, there are also some works regarding the data as signal and analyzing them in time and frequency domain. Weng et al. [113] proposed the Event Detection with Clustering of Wavelet-based Signals (EDCoW) method that adopted the Discrete Wavelet Transform (DWT) to extract word features from Twitter stream. In their work, textual stream can be seen as signal and a sliding window is adopted to detect frequency of a word and filter away trivial words. The remaining words are further used to construct a similarity graph as discussed before. Litvak et al. [114] further extended the EDCoW in order to avoid the performance degradation when social events sharing the similar wavelet shape and temporal information. Although these works mainly focus on event happened with textual content, their idea can be extended into multi-modal social event detection.

2.3. Meta-data aggregation for event detection

Either single-modality feature learning methods or multi-modal feature learning methods mainly focus on the media content. However, most of the definitions show that events are closely related to time, space, object and social components. For instance, Hakeem et al. [115] referred to event as a collection of actions performed between one or more agents. Jiang et al. [116] defined event as an activity-centered occurrence that involves objects/people engaged in several activity-driven actions at a specific place and time. Another high-level definition of event is that of a story related to some news topic comprising of patterns that occurred at some specific time and space [117]. Therefore, the basic event description could include following aspects: content, spatial information, temporal information and social information. Therefore, any input message of social media can be described by the following structured format [118]:

$$Message_i = \{C_i, L_i, T_i, S_i\} \tag{8}$$

Where C_i represents data content (single-modality or multi-modality). L_i describes the location of the message sending place or the event place. T_i represents the posting time of the message. Usually, the posting time can be considered as the approximation of occurrence time of events. S_i represents the relationships between users and their followers, forming the connection structure of their message dissemination. Following this idea, several works generated similar descriptions in different scenarios. Abebe et al. [119] proposed a generic Metadata Representation Space Model (MRSM) to represent social media message as a triplet (ID number, time, location, semantic). Shan et al. [120] defined a field (URL, title, body, score, ID, timestamp) to model webpage. Xu et al. [121] constructed a set of attributes, such as authenticity, timestamp, textual feature, visual feature, to describe multi-modal news articles. However, not all attributes of events in Eq. (8) are covered in most above works because of their specific applications.

[5] http://www.image-net.org/

Table 2
A summary of some reviewed works on specific social event detection.

Ref.	Time	Modality	*C*	*L*	*T*	*S*	Platform	Data	Tools
[86]	2015	Multi	✓	–	–	–	Weibo	C1: 31K microblogs C2: 13K microbolgs	Cross-Media-LDA
[88]	2014	Multi	✓	–	–	–	Flickr	60K image/text documents	Multi-modal supervised LDA
[89]	2015	Multi	✓	–	–	–	Flickr	100K image/text documents	Multi-modal supervised LDA
[90]	2014	Multi	✓	–	–	–	Flickr	36K image/text documents	Multi-modal supervised LDA
[91]	2016	Multi	✓	–	–	–	Wikipedia and Google News[a]	130K images and 50K texts	(Incremental) Multi-modal LDA
[105]	2017	Multi	✓	–	–	–	Weibo	3 million microblogs	Multi-modal deep neural networks
[122]	2012	Multi	✓	✓	✓	–	Last.fm [b]	1 million social pictures	SVM classifier
[123]	2012	Multi	✓	✓	✓	✓	Flickr	C1: 270K social photos C2: 73K social photos	SVM classifier and incremental cluster
[124]	1961	Single	✓	–	–	–	IRE Transactions [c]	405 thesis documents	Bayesian classifier
[125]	2010	Single	✓	–	–	–	Wikipedia [d]	104,713 textual items	Gradient Boosted Decision Trees
[126]	2005	Multi	✓	–	–	–	Various sources	100 h sports videos	SVM classifier
[127]	2019	Multi	✓	–	–	–	Twitter	1851 tweets	SVM classifier
[128]	2018	Multi	✓	–	–	–	Twitter	C1: 900,000 tweets C2: 760,000 tweets	Multi-kernel learning and SVM classifier
[129]	2009	Single	✓	✓	✓	✓	Twitter	*did not mention*	Naive Bayes classifier and online cluster
[130]	2017	Multi	✓	–	–	–	Weibo and Xinhua news[e] Twitter	C1: 40k microblogs and news C2: 13,000 tweets	Multi-modal Recurrent Neural Networks
[131]	2016	Single	✓	–	✓	–	Weibo	4K microblogs	Convolutional neural networks
[132]	2011	Multi	✓	✓	✓	–	Flickr	73K social photos	External knowledge and clustering
[133]	2012	Multi	✓	✓	–	–	Flickr	73K social photos	External knowledge and probabilistic model
[134]	2019	Multi	✓	–	–	–	Flickr	74K image/text documents	External knowledge and multi-modal LDA

[a] news.google.com.
[b] http://www.last.fm.
[c] IRE Transactions on Electronic Computers, Vol. EC-8, No. 1. Published by the Professional Group on Electronic Computers.
[d] https://www.wikipedia.org.
[e] https://news.sina.com.cn.

In practice, not all works should follow such structured format to obtain the event features because directly concatenating all components may ignore the statistical property of each attribute. There are also various strategies effectively learning the event features. For instance, some works [70] sequentially consider the content and meta-data while some works [68,71] fuse the attributes into the content and measure the similarity between social messages. Some details will be provided in the next section.

3. Event inference

Event inference is the event discovery stage of deciding whether an input message belongs to a social event or not. Its effectiveness directly affects the accuracy of multi-modal social event detection systems. Motivated by single-modal social event detection works, several works directly convert non-textual media into textual tags and further utilize traditional methods for multi-modal social event detection [135,122,136,123,137]. However, it is hard to comprehensively represent non-textual data only by specific words so that this approach is not able to provide a wide range of application in multi-modal event detection system. Researchers have also developed amount of event inference technologies for multi-modal social event detection in recent years. In this section, the representative event inference methods are reviewed according to the social events attribute and social data attribute: (1) specific/unspecific social events, (2) new/retrospective social events.

3.1. Specific versus unspecific event

A fundamental categorization criterion for event detection is the type of social event, i.e. specific and unspecific events. Specific events are events whose characteristics are usually available, such as disasters, sports games. Therefore, specific events inference technologies usually attempt to fully exploit the information at hand for better detection performance. On the contrary, since no prior information is available for unspecific events (such as hot topics, breaking news), unsupervised manner is suitable for their detection. Tables 2 and 3 summarized the reviewed detection works for specific and unspecific social events respectively. Their details will be provided in the following subsections.

3.1.1. Specific event detection

Typically, specific social events are partially or fully specified with related event attributes, such as occurring time and space, involved participants and event description. Since some prior information of specific event is usually available, supervised techniques are more suitable for specified event than for unspecific event.

Popular supervised ways include classification based methods where classifiers are adopted to categorize the input message into pre-defined event classes. One of the pioneer works is [124]. Maron et al. classified textual documents by Bayesian classifier in terms of their clue words. Researchers also extended substantial classification algorithms for multi-modal social event detection including Decision Tree (DT) [125,150], K-Nearest Neighbor

Table 3
A summary of some reviewed works on unspecific social event detection.

Ref.	Time	Modality	*C*	*L*	*T*	*S*	Platform	Data	Tools
[66]	2015	Multi	✓	✓	✓	✓	Flickr	167K social photos	K-means
[70]	2014	Multi	✓	–	✓	–	Youku[a] and Sina	9K news and 7K videos	Semi-supervised multi-modal NMF
[77]	2016	Multi	✓	–	✓	–	C1: Youtube[b] C2: Youku and Sina	C1: 3660 web videos C2: 7K news and 2K videos	Multi-modal graph learning
[80]	2014	Multi	✓	✓	✓	✓	Flickr	167K social photos	Structural Clustering Algorithm for Networks (SCAN)
[87]	2015	Multi	✓	✓	✓	–	Twitter	10M tweets	Spatio Temporal Multimodal Twitter LDA
[120]	2012	Multi	✓	–	✓	–	Various	Web page, videos, news and microblogs	Heuristic clustering algorithm
[137]	2015	Multi	✓	✓	✓	–	Yahoo[c] and Flickr	99.2M photos and 0.8 M videos	Hierarchical clustering
[138]	2017	Multi	✓	✓	✓	✓	Flickr	167K social photos	Multi-modal NMF and K-means
[139]	2015	Multi	✓	✓	✓	✓	Flickr	167K social photos	Semi-supervised multimodal clustering
[140]	2017	Multi	✓	✓	✓	✓	Flickr	110K social photos	Hybrid clustering and data recovery residual models
[141]	2012	Multi	✓	✓	✓	–	Flickr	70K social photos	Multi-modal spectral clustering
[142]	2012	Multi	✓	–	✓	–	Youtube	80K videos	Star-structured graph based co-clustering
[143]	2019	Multi	✓	–	–	✓	Weibo	2M texts and, 4M images	Multi-modal co-clustering
[144]	2016	Multi	✓	–	–	–	Youtube and Wikipedia	7K videos and 4K articles	Heterogeneous graph learning
[145]	2013	Multi	✓	–	✓	–	Youku and Sina	2K videos and 7K articles	Graph shift
[146]	2016	Multi	✓	–	✓	–	Five US Broadcast	379 news videos	Swendsen–Wang graph cuts
[147]	2015	Multi	✓	–	✓	–	Yahoo and Flickr	99.2M photos and 0.8 M videos	Multi-modal graph clustering
[148]	2017	Multi	✓	–	✓	–	Weibo	3 million microblogs	Graph learning and transfer cut
[149]	2017	Multi	✓	✓	✓	✓	Flickr	110K social photos	SCAN and QCA

[a] https://www.youku.com.
[b] https://www.youtube.com.
[c] https://www.yahoo.com.

(KNN) [151], Back Propagation Neural Networks (BPNN) [152] and Support Vector Machine (SVM) [126,122,153,127]. Blandfort et al. [127] proposed two strategies to classify multi-modal features, i.e. early fusion and late fusion. In early fusion, different features are concatenated into a single feature vector and further this single feature is input into conventional SVM. In late fusion, separate SVMs accomplish classification task for each feature and another final SVM is trained to make the final detecting decision from the probability outputs of the previous SVMs. Similar to the late fusion strategy, Zhao et al. [154] presented a multi-modal microblog classification method in a multi-task learning framework. Specifically, multiple classification tasks are conducted for each modality within utilizing appropriate shared information across different tasks. Then, the outputs are integrated by a SVM with weighting each modality. Generally speaking, in classification based methods, classifiers are typically trained on a set of labeled multi-modal social media data and further adopted to target the input multi-modal message as a specific type of event, such as sports event or non-sport event [126], natural disaster or non-disaster event [153], gang violence or non-violence event [127] and Chile earthquake or other earthquake event [128].

Classifiers also can be utilized to filter out some irrelevant input message when detecting specified social event [155,156, 129,123]. Sankaranarayanan et al. [129] trained a Naive Bayes classifier to mark event or non-event tweet. The junk tweets are discarded so that remaining tweets have a good chance of being related to events. Gao et al. [156] developed an accurate classifier to filter out noise by taking into account the content and social nature of social media data. With embedding textual, low-level visual features, high-level visual semantic features and social relationship, an off-the-shelf classifier, such as SVM, can then be trained and applied to microblog filtering. Above filtering operation would be helpful for discarding a large number of noisy information. Hence, the computational cost would be economically reduced and the performance of further event identifying stage is improved. However, the effectiveness of their approach is sensitive to the empirical threshold in the classification stage because the classifier may also filter out some relevant information when the threshold is not applicable.

Traditional classification algorithm generally deals with social data in a linear manner so that it is inapplicable to tackle the complex and nonlinear practical case. In recent years, some works try to conduct event inference tasks via deep learning [105,130] due to their powerful ability on nonlinear learning. For instance, Gao et al. [105] proposed Multi-modal Multi-instance Deep Network (M2DN) to classify multi-modal microblogs and detect brand related social event. Their model includes (a) two pathways, each of which connects to one single input modality, i.e., image or text; (b) an aggregation layer, where the maximum output neuron activations strategy is adopted to combine each pathway; (c) classification layer, where Softmax layer is adopted to predict the event class of the microblog. Similarly, Jin et al. [130] individually adopted VGGnet for visual data and Long Short Term Memory network for textual data but the features of each modality are fused by directly concatenating. A binary classifier finally decides whether a microblog belongs to rumor or not. Deep learning based methods typically tackle detection tasks with end-to-end model. Namely, they can take the original message as input and further directly output the event detection results. In this situation, there is no need to choose middle features for each modality. However, recent deep learning based works only consider contents while fail to consider the attributes of social event, resulting in degradation of practical detection performance. Although there are single-modal works simultaneously analyzing semantic and temporal information [131], how to fully take into account the multi-modal event attribute is still challenging.

External knowledge also can facilitate the performance of specific event detection approaches. This is particularly true in scheduled social event, such as sports game [132], music concert [133]. For example, transferring existing geographic information in GeoNames[6] to handle the venue location of a topic-specific event [133]. Exploiting music information from last.fm website for prior knowledge acquisition of concert [132]. WordNet [157] is another popular external event directory base for social event detection, where English vocabularies are grouped together based on their meanings. It would help researchers identify the social events which expressed by different yet synonymous words. Xue et al. [134] proposed the knowledge based topic model for multi-modal social event classification where WordNet are adopted as lexical external knowledge. Specifically, their model incorporates external knowledge into multi-modal Latent Dirichlet Allocation [88] by taking them as another observable variable. When sampling a word, the model automatically searches for the related item from the knowledge base. Hence, the external knowledge is able to guide the traditional topic model to discover cognitive-related events.

3.1.2. Unspecific event detection

A characteristic of unspecific event detection methods pertains to their applicability on general topics or breaking news. As for general topics, they are often generated in social media platforms and appeal amount of users discussing together while breaking news often burst in a very short time.

A. General topics detection

Since seldom prior knowledge are available for general topics and it is impossible to manually label all type of social data, unsupervised methods are more feasible for unspecific social event detection, such as clustering. Clustering based method aims at grouping all input messages into a number of clusters, and ultimately the clusters can be viewed as events or non-events. For example, Choi et al. [137] developed a step-by-step clustering method and chose the best cluster subset that meet both relevance and diversity criteria as the social event. Initially, unsupervised clustering using temporal information is carried out for collected media data. Then, similar clustering operation is conducted using spatial information. If the locations of two clusters are similar, these clusters are merged. Finally, Clusters with the similar textual descriptions are merged. Similarly, Shan et al. [120] developed a heuristic clustering algorithm which can run in parallel for different parts of data, reducing total time cost of event detection. However, above systematic methods depends heavily on the parameters since each step needs empirically similarity threshold setting.

K-means algorithm is a popular tool to accomplish clustering tasks. Ma et al. [66] decomposed the multi-modal graph by SVD and further conducted K-means on these low-dimension feature vectors. However, there are two main demerits on K-means. Firstly, the number of clusters should be determined in prior while it is difficult to explicitly know the number. In addition, K-means suffers from the uncertainty of the initial centers. A feasible solution could be the application of Density-Based Spatial Clustering of Applications with Noise (DBSCAN), which is able to automatically specify the cluster numbers and can discover clusters with arbitrary shape [139,158,159]. Capdevila et al. [159] proposed an event discovery technique in twitter (Tweet-SCAN), which can be viewed as multi-modal extensions of DBSCAN. They implemented independent neighbor identification in each modality to group close neighbors into a dense cluster, which is finally associated to an event. However, DBSCAN is time-consuming especially for vast, high-dimensional multi-modal social data. As a result, Yang et al. [140] designed a hybrid clustering model where the number and centers of clusters are initialized by DBSCAN and K-means is adopted to accomplish event inference task. This hybrid strategy can avoid the disadvantages of DBSCAN and K-means to some extent.

Amount of graph based clustering methods are also proposed for multi-modal social event detection. A popular method is similar to the spectral clustering. Namely, after obtaining the multi-modal graph, conventional spectral clustering algorithm can be further conducted to group data and discover the social event [141]. Some works also designed the co-clustering algorithms [142,136,143]. For example, Qian et al. [143] considered three dimensions in their work, i.e. texts, images and users. They firstly determined three joint probability matrices for text and image, user and text, image and user respectively. Further, their co-clustering approach tried to find the optimal cluster sets by minimizing the linear combination of the Bregman information of three joint probability matrices. Generally speaking, the basic idea of co-clustering is that, graphs of each modality is alternatively and iteratively updated, and event clusters can be ultimately determined once a criterion is satisfied such as minimum information entropy loss, minimum square error. There are also some works regarding event inference as dense subgraph detection problem [144,80,145,77,146,147]. For example, Chu et al. [77] adopted the graph shift based subgraph detection method to measure the connection strength of graph nodes and find all the local maximums. Each local maximum indicates a dense subgraph, which is finally defined as a hot topic or social event. Although many works have achieved satisfied results on multi-modal social event detection, graph based clustering is not suitable for large scale of social data analysis because they require the construction of similarity matrix. Moreover, graph based clustering may not correctly capture the unbalanced event distribution of social data due to its nature in dealing with balanced data.

B. Breaking news detection

The underlying assumption for general topics detection methods is that topics are in relation to static social media data collections. However, breaking news driven unspecific social events often keep evolving and emerge in a very short time and new data are constantly being produced. In this situation, aforementioned methods may not feasible for breaking news detection. A simple yet efficient approach is burst words detecting, which aims at estimating the occurrence number of specific words. If the frequency of an actual word occurred more than expected, then this word can be viewed as burst and the social event described by burst word can be discovered [160]. However, specific textual words cannot comprehensively represent non-textual data so that this approach is not suitable for multi-modal breaking news detection system.

Incremental clustering is a popular approach to deal with the continuously generated multi-modal social media data. Its basic idea is to analyze the message sequentially. Once the similarity between a new message and an existing cluster is greater than

6 http://www.geonames.org

a predefined threshold, the input message can be merged into an existing cluster; otherwise, it is regarded as a new event and a new cluster will be formed [135]. However, the predefined threshold is usually empirically selected.

Aforementioned graph based methods are usually applicable when all the social media data are available. To make graph based methods be applicable for breaking news detection, there are also some works developed incremental graph based clustering methods. For example, Zhao et al. [148] incrementally update the constructed graph for new input social data. Considering that, it is difficult to build all the hyper-edges between original graph and new microblogs due to the high computational cost, they only selected top-n most representative new microblogs based on the number of reposts and comments. Further, transfer cut method [161] was adopted to partition the updated graph. Petkos et al. [149] directly adopted Quick Community Adaptation (QCA) for incremental sub-graph detection. QCA can maintain detecting sub-graphs up-to-date by appropriately processing operations when new nodes/edges are added or nodes/edges are deleted. This work is threshold-free. However, it may not applicable when dealing with large-scale multi-modal social data due to the high computational cost of graph construction.

3.2. New versus retrospective event

Similar to TDT, event inference technologies also can be divided into new event detection and retrospective event detection. Different from the unspecific/specific events detection, new/retrospective events detection pay more attention to the data characteristic. New event detection, also known as first story detection, tries to detect new social events from living social data stream in real time. On the other hand, retrospective event detection focuses on detecting occurred events from a collection of historical social media data.

3.2.1. New event detection

New event detection is sometimes quite similar to aforementioned unspecific breaking news detection because it aims at discovering emerging social events in evolving social data. Accordingly, burst words detecting, incremental clustering based methods are feasible for new event detection. To boost the cost efficiency and detection performance, several techniques can be adopted including parallel calculation, choosing representative message, reducing event representation dimension.

Time series mining are also popular to discover social event from constantly generated data. For example, assuming that same social event only emerged in short time, temporal-windows based methods compare the new input message with recent message sequentially to discover the content and time similarity over occurred events [162]. Sequential model, such as Hidden Markov Model (HMM), Kalman Filter (KF), Particle Filter (PF) also can provide a reasonable solution for temporal information modeling of social data stream [163–165,160]. In practice, sequential model is usually adopted to predict social event within available historical social data stream. If they predict social events in short future, they also can be viewed as real-time events detection methods. In addition, generative model can be proposed to model user behavior in social media stream based on media content, location [166], posting behavior [167] etc. Yilmaz et al. [166] introduced an unsupervised probabilistic generative model for social event detection in Twitter. Hashtags as well as geolocations were simultaneously utilized to model the exponential distribution of social events. The expectation–maximization (EM) algorithm was derived to find the maximum likelihood of the model parameters. Alternatively, the generative model also can be used to detecting abnormal social behavior by modeling normal user behavior. For example, Costa et al. [167] observed that the distribution of postings inter-arrival times is characterized by specific patterns such as heavy tails, periodic spikes. Therefore, they proposed a generative model that is able to match all discovered patterns. Their model can be used to mark outliers and detect users with non-normal behaviors. This is valuable to eliminate the effect of abnormal elements when detecting social events.

Note that, new event detection may not be restricted to unspecific breaking news when detection tasks focus on specific event (such as sports games, disaster) or some attributes (such as location) of events are available. Wang et al. [165] input traffic related tweets and GPS probe readings into the Coupled HMM to accurately detect and forecast traffic conditions. To reduce the computation cost, a parallel EM-algorithm was proposed to efficiently estimate the variables of Coupled HMM. Sakaki et al. [163, 164] firstly trained a SVM classifier to mark the input tweets as events-related (earthquakes related) or non-events-related (others). Further, event occurring time is modeled as an exponential distribution while the location and trajectory is estimated by KF and PF. According to the prediction results of KF and PF, their method can detect the earthquakes related events in real time. The reason that above work can focus on interested events is that only related social media constitute the analyzed data.

3.2.2. Retrospective event detection

In retrospective event detection, many approaches are usually involved to group a collection of multi-modal social data into event clusters. Since all data are available, to some extent, retrospective event detection is similar to aforementioned general topic detection. Therefore, aforementioned clustering algorithm [139], topic modeling [88] based methods are naturally feasible to detect retrospective events in historical social data collections. Note that, if the timestamp of historical data are available, new event detection approaches are also applicable for retrospective event detection.

From another perspective, retrospective event detection also can be regarded as retrieving the relevant social media data by performing the given queries (single-modal or multi-modal) over data collections. Therefore, many query-based methods can be designed to accomplish event related message retrieving tasks [106,107,168]. Their main idea is capturing the message relevance (also can be viewed as query) to a specific event by integrating multi-modal feature. For example, a deep common semantic space for cross-modal event retrieval is proposed in [168]. This is achieved by exploiting deep learning models to extract semantic features from images and textual articles jointly. Metzler et al. [106] also introduced the event retrieval method in microblogs. Specially, they retrieved summarized microblogs in response to an event query, rather than retrieving individual microblog messages. These techniques are applicable for search service of social media platform, such as provided by Twitter and Sina Weibo. However, querying specific events on social platform is still challenging due to the sparseness of message and a large number of querying mismatching. Moreover, the relevant message also may not contain query-related terms.

Traditional retrieving tasks only focus on query-related content while ignore the attributes of social events. Recent research efforts have started to exploit retrospective events with considering temporal information and thus generate event evolution map. This is also called as event summarization. Although multi-modal event evolution maps are relatively new and unsophisticated, there are still some works yielding notable achievements. Timeline map is the most popular linear axis to achieve both content coherence and temporal continuity [169–171]. Accordingly, the evolution of events can be easily followed by users since an event overview can be provided. Sahuguet et al. [169] automatically generated a timeline of queried events by mining video information. They first roughly generated a timeline to have a global view for an interested event. Further, videos details were added into each time segmentation for visualization. Yan et al. [171] proposed a graph-based framework to generate a timeline summary for news event with pictures. In their framework, individual sub-summaries were generated by taking into account the mutual dependencies between sentences and images, and further were iteratively refined by considering how they contribute to the global timeline and their coherence. However, the one-dimensional axis only works for simple social events while real-world social events often involve branches, intertwining narratives and side news. Therefore, storyline, also known as metro-map, are generated for yielding richer and clear events structure [172–174]. Xu et al. [174] generated a high-quality map by making different events sharing the same timestamp stay close but their method still fails to reflect the correlation of different social events from semantic level. Wang et al. [173] proposed a graph-based method to generate pictorial temporal storyline. Since nodes of graph can be viewed as a sub-topic, their storyline is naturally represented in 2-dimensional space. By characterizing the interactions, the non-linear 2-dimensional structures can not only illustrate the temporal information of events but also reveal the latent relationships among various aspects of events.

4. Datasets and evaluation

Despite of the amounts of research methods for multi-modal social event detection, the available public datasets for their evaluation are limited. Due to this fact, most of aforementioned works are tested on the self-collected and self-annotated social media datasets. Since both the data and ground truth are missing, it is unrealistic to provide an explicit numerical comparison between these approaches. In this section, we only present some selected works on few public datasets.

• Social Event Detection (SED) dataset. One of the publicly available competition datasets is the SED datasets provided by the MediaEval Initiative.[7] In this dataset, four subsets are released in four successive years (2011–2014). The main task is to discover the interested social event from the vast user-generated Flickr multimedia content together with their surrounding metadata. For each subset, there are several different challenging tasks requiring participants to accomplish. For example, the first challenge of SED 2011 requires the participants to discover all soccer events taking place in Barcelona and Rome. The second challenge aims at finding all events that took place in May 2009 in the Paradiso venue (Amsterdam) and in the Parcdel Forum (Barcelona). The details of subsets can be found in works [175–178].

As shown in Table 4, the comparison is presented under different tasks, Two metrics, F-measures and normalized mutual information (NMI), are adopted to evaluate the performance. From the results of first two datasets, we can see that more satisfied detection performance can be achieved when the social events have a clear definition. For example, results achieved in Challenge 2 and 3 of SED 2012 are worse than that in Challenge 1. This is because "technical events" and "indignados movements" in Challenge 2 and 3 are fuzzy while "soccer events" are much clear. Therefore, there is still space for improvements in unspecific social events detection tasks. On the other hand, generally speaking, methods achieve better results on SED 2013 and 2014. This is mainly because non-events content of SED 2013/2014 datasets are filtered in advance and all the documents can be assumed to belong to a social event. Therefore, filtering irrelevant data may boost the event detection performance in social media sites.

• Brand-Social-Net dataset [155].[8] Brand-Social-Net dataset is a large-scale public dataset on brand-related social media data. There are totally 20 brand-related events in over 3 million microblogs. Microblogs are crawled from one of the largest Chinese social platform, i.e. Sina Weibo. Most of them are post in Chinese, with a large proportion containing images. Several tasks can be performed on this dataset. For example, detecting and tracking brand-related events, analyzing social sentiment and social network.

Table 5 shows the publicly reported results on dataset BSN dataset. Four popular metrics are adopted including precision, recall, F score and accuracy. Note that, the accurate results number is missing in some original works (star-marked) so that we estimate the results from the figures they provided. It is obvious that works [155,105,148] achieved similar performance under the static scenario. The satisfied results in work [156] indicates the effectiveness of embedding the social relationship as well as the high-level feature. However, the performance of real-time mode is not as satisfied as that of static mode. Therefore, there is still space for improvements in real-time social event detection tasks.

• Multi-Modality Social Event dataset (MMSE). Recently, many works conduct experiments on this dataset collected by Qian et al. [88]. Several works also adopted a wider collection of this dataset. For example, to the best of our knowledge, the early version [90] contains 35K multi-modal documents while work [134] expand it into a 75K dataset (also called HFUT-mmdata[9]). However, this dataset fails to simulate the real-world social media scenario due to the following reasons. On the one hand, all the documents correspond to a specific social event in this dataset while social media data usually have a large ratio of event to non-event documents. In addition, real-world event data are usually imbalanced while the number of documents in each event is similar in this dataset.

Table 6 shows the publicly reported performance on MMSE dataset measured by accuracy. It is obviously that the performance of work [83] (text only) is much better than that of work [83] (visual only). This shows that the textual information is more valuable than visual information in the social event detection tasks. The main reason might be the diversity of images. Comparing the single-modality based methods , multi-modality based methods can achieve better results, which shows the effectiveness of multi-modal property. In addition, the satisfied performance of work [134] proved the effectiveness of embedding external knowledge.

5. Conclusion and discussion

Multi-modal social event detection tries to uncover a collection of real-world actions in unprecedentedly vast social media data. The discovered information can be transformed into actionable knowledge. Recent efforts on multi-modal social event detection are mainly distributed into two parts. The first part is event feature learning. Single-modality feature learning is the foundation of social media understanding. Subspace learning is an effective way to tackle the multi-modal data and solve the problem of heterogeneity. In most of works, their effectiveness depends heavily on the selected features because they only take selected middle-level features as input. On the contrary, end-to-end learning structures are more flexible because they are designed to directly obtain the correlation of heterogeneous data. The second part is event inference. Various techniques can be selected by researchers according to the social event characteristics. Generally speaking, supervised methods, such as classification and deep learning, are feasible to detect specified events

7 http://www.multimediaeval.org/

8 http://www.nextcenter.org/Brand-Social-Net/

9 http://scholarhub.cn/ScholarHubProject/MMTM/HFUT-mmdata.zip

Table 4
The comparison results on Social Event Detection dataset.

Sub-dataset	Ref.	Challenge	F-score	NMI
SED 2011	[132]	Challenge 1	0.59	0.25
		Challenge 2	0.69	0.61
	[133]	Challenge 1	–	0.54
	[179]	Challenge 1	–	0.63
		Challenge 2	–	0.67
SED 2012	[66]	Challenge 1	–	0.77
		Challenge 2	–	0.68
		Challenge 3	–	0.65
	[139]	Challenge 1	–	0.76
SED 2013	[180]	Challenge 1	0.93	0.98
		Challenge 2	0.45	–
	[181]	Challenge 1	0.94	0.98
SED 2014	[138]	Challenge 1	–	0.98
	[140]	Challenge 1	0.97	0.99
	[182]	Challenge 1	0.94	0.98

Table 5
The comparison results on Brand-Social-Net dataset.

	Precision	Recall	F score	Accuracy	Description
[155]*	0.64	0.58	0.60	–	Static scenario
[105]*	0.70	0.65	0.67	–	Static scenario
[148]*	0.68	0.65	0.66	–	Static scenario
	0.58	0.64	0.61	–	Real-time scenario
[156]	0.83	0.69	0.78	–	Static scenario
[154]	–	–	–	0.85	Static scenario

Table 6
The comparison results on Multi-Modal Social Event dataset.

Ref.	Dataset [89]	Dataset [134]	Dataset [90]	Dataset [88]	Description
[26]	–	0.68	–	–	Text only
[83]	0.76	–	0.70	0.72	Text only
[83]	0.36	–	0.31	0.40	Visual only
[85]	0.75	0.72	0.72	0.72	Multi-modal
[90]	0.76	0.80	0.72	0.77	Multi-modal
[88]	–	–	0.83	0.88	Multi-modal
[134]	–	0.85	–	–	Multi-modal

because prior knowledge is usually available. In contrast, unsupervised methods are more suitable for unknown event detection. Moreover, although new/retrospective social event detection focuses on the temporal attributes of social media data, they also share some common techniques with (un)specific social event detection. In this situation, researchers can select aforementioned inference methods depending on the practical social event detection systems. Although substantial works have been done in this domain and some significant success has been achieved in recent years, detecting events in multi-modal social data is still a very difficult task. Future works can focus on some open issues.

A. Information enrichment. Textual information often co-occurs in the visual and audio social media. If co-occurring textual information can be jointly considered, it may lead to the improvement of social event detection. For example, compared analyzing text or speech signals only, simultaneously analyzing both modalities can comprehensively exploit speech contents, speaking tones and pause length. Another example is extracting texts in images or videos to accurately understand the visual content. In this situation, Optical Character Recognition (OCR) [183,184] and Automatic Speech Recognition (ASR) [185,186] could be the effective techniques to obtain textual information from visual and audio information. In addition, another feasible enrichment could be the user behavior analyzing, such as like/dislike, commenting, repost. This is useful for hot social event detection. Recently, many research works [187,188] found that user click feature is valuable in justifying the relevance between a query and clicked objects. This is also true in the retrospective event detection, especially the search service of social platform.

B. Accuracy of multi-modal feature learning. A good synthetization of multi-modal social media data is able to capture comprehensive and distinguishing characteristics of data and further improve the event detection performance. Unfortunately, the simple idea of directly concatenating different feature vectors of each modality may result the "curse of dimensionality" problem. Although some considerable progresses have been made to learn the meaningful information from the multi-modal social data, the current approaches still need to be improved for achieving better learning accuracy on social platform. On the other hand, new effective methods are required. For instance, the ideas of other deep architectures such as extension Recurrent Neural Network (RNN), Generative Adversarial Nets (GAN), are valuable to improve the accuracy of multi-modal feature learning.

C. Dealing with evolving voluminous data. Nowadays, millions of statistics are generated per minutes on the social media platform. To process these vast data against social events, immense storage space and efficient computing algorithm are required. In some particular applications, especially when detecting bursty events, systems should also incorporate dynamism and scalable running algorithm so that it is able to handle sudden increased social data and accurately detect social event within stipulated period.

Moreover, in real-world scenario, social events are usually evolving so that the concept of events would shift. In this situation, the basic assumption of traditional detection methods that event distribution of social media is static, is more likely to be violated. These problems pose a great challenge on social event detection.

D. Improvement of data quality. Since the effectiveness of multi-modal social event detection system depends heavily on the social data, the detection techniques should also consider the quality of the raw data. On the one hand, the case that a modality of a message or a post may disappear usually happens in practice. For example, some users may update a post with only texts or images. The spatial information in terms of longitude and latitude may be inaccurate or missing. In this situation, detection approaches are required to consider the underlying incompleteness of raw data. On the other hand, raw social data are usually sparse and unbalanced. Work [9] found that only 2% of collected documents from a random stream were related to events. For example, a multi-modal social event detection system tries to discover events from 10000 documents by clustering method, but it is difficult for only about 200 event-related documents to form distinct clusters. So far, most of works did not consider this problem and test their performance on the collected datasets with a large ratio of event documents and non-event document. Namely, they are not applicable to tackle this challenge when detecting social events in real-world scenario. Possible solutions for this problem could be identifying of event-related message before event inference, because once irrelevant documents have been identified, they can be filtered out to improve the accuracy.

E. Combining information from cross-platform. Current works mainly focus on event detection from single social platform. Since missing information of one platform is very likely to be available on other platforms, a good synthetization of information from multiple social platforms can provide comprehensive as well as complementary understanding of real-world social data. A challenging way to achieve this goal is learning common features shared by multiple sources. Therefore, detection methods can be applied to simultaneously analyze social data of multiple platforms. Alternatively, platform-specific social events can be detected by applying detection methods on each platform and further be emerged to obtain the final results. This is also called late-fusion strategy. Moreover, in recent years, transfer learning has been proved useful on extracting information from related external knowledge. This idea is also true in transferring knowledge of a social platform into another social platform.

Note that, to transform the valuable relevant events information into actionable knowledge, there are still many further works need to be done after event detection, such as event evolutionary analysis [189], topic opinion analysis [190] and public decision making [191,192]. These tasks are vibrant research area that draws on techniques from different fields such as linguistics, sociology, data mining and natural language processing. Combining knowledge of various fields can effectively and efficiently develop a practical system for real-life events analysis.

CRediT authorship contribution statement

Han Zhou: Conceptualization, Writing - original draft. **Hongpeng Yin:** Supervision, Project administration, Funding acquisition, Writing review & editing. **Hengyi Zheng:** Writing - review & editing, Investigation. **Yanxia Li:** Writing - review & editing, Visualization.

Acknowledgments

This work is jointly supported by Chongqing Nature Science Foundation of Fundamental Science and Frontier Technologies, China (cstc2015jcyjB0569), China Central Universities foundation (2019CDYGZD001) Scientific Reserve Talent Programs of Chongqing University (cqu2018CDHB1B04), General Program of National Natural Science Foundation of China (61773080), Graduate Research and Innovation Foundation of Chongqing, China (CYS19072) and Kehui Graduate innovation competition of Chongqing, China (09301119).

References

[1] Dave Evans, Social Media Marketing: The Next Generation of Business Engagement, John Wiley & Sons, 2010.
[2] Clayton J. Hutto, Eric Gilbert, Vader: A parsimonious rule-based model for sentiment analysis of social media text, in: Eighth International AAAI Conference on Weblogs and Social Media, 2014.
[3] Rajiv Ratn Shah, Yi Yu, Akshay Verma, Suhua Tang, Anwar Dilawar Shaikh, Roger Zimmermann, Leveraging multimodal information for event summarization and concept-level sentiment analysis, Knowl.-Based Syst. 108 (2016) 102–109.
[4] Mahmood Hajli, A research framework for social commerce adoption, Inf. Manage. Comput. Secur. 21 (3) (2013) 144–154.
[5] Maryam Khodabakhsh, Mohsen Kahani, Ebrahim Bagheri, Zeinab Noorian, Detecting life events from twitter based on temporal semantic features, Knowl.-Based Syst. 148 (2018) 1–16.
[6] Maia Zaharieva, Matthias Zeppelzauer, Christian Breiteneder, Automated social event detection in large photo collections, in: Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, ACM, 2013, pp. 167–174.
[7] James Allan, Introduction to topic detection and tracking, in: Topic Detection and Tracking, Springer, 2002, pp. 1–16.
[8] Farzindar Atefeh, Wael Khreich, A survey of techniques for event detection in Twitter, Comput. Intell. 31 (1) (2015) 132–164.
[9] Georgios Petkos, Symeon Papadopoulos, Vasileios Mezaris, Raphael Troncy, Philipp Cimiano, Timo Reuter, Yiannis Kompatsiaris, Social event detection at MediaEval: a three-year retrospect of tasks and results, in: Proc. ACM ICMR 2014 Workshop on Social Events in Web Multimedia, SEWM, 2014.
[10] Anuradha Goswami, Ajey Kumar, A survey of event detection techniques in online social networks, Soc. Netw. Anal. Min. 6 (1) (2016) 107.
[11] Muskan Garg, Mukesh Kumar, Review on event detection techniques in social multimedia, Online Inf. Rev. 40 (3) (2016) 347–361.
[12] Mário Cordeiro, João Gama, Online social networks event detection: a survey, in: Solving Large Scale Learning Tasks. Challenges and Algorithms, Springer, 2016, pp. 1–41.
[13] Nikolaos Panagiotou, Ioannis Katakis, Dimitrios Gunopulos, Detecting events in online social networks: Definitions, trends and challenges, in: Solving Large Scale Learning Tasks. Challenges and Algorithms, Springer, 2016, pp. 42–84.
[14] Christos Tzelepis, Zhigang Ma, Vasileios Mezaris, Bogdan Ionescu, Ioannis Kompatsiaris, Giulia Boato, Nicu Sebe, Shuicheng Yan, Event-based media processing and analysis: A survey of the literature, Image Vis. Comput. 53 (2016) 3–19.
[15] Matthias Zeppelzauer, Daniel Schopfhauser, Multimodal classification of events in social media, Image Vis. Comput. 53 (2016) 45–56.
[16] Houkui Zhou, Huimin Yu, Roland Hu, Junguo Hu, A survey on trends of cross-media topic evolution map, Knowl.-Based Syst. 124 (2017) 164–175.
[17] Tianpeng Liu, Feng Xue, Jian Sun, Xiao Sun, A survey of event analysis and mining from social multimedia, Multimedia Tools Appl. (2019) 1–18.
[18] Shiliang Sun, Chen Luo, Junyu Chen, A review of natural language processing techniques for opinion mining systems, Inf. Fusion 36 (2017) 10–25.
[19] Ewoud Pons, Loes M.M. Braun, M.G. Myriam Hunink, Jan A. Kors, Natural language processing in radiology: a systematic review, Radiology 279 (2) (2016) 329–343.
[20] Waseem Rawat, Zenghui Wang, Deep convolutional neural networks for image classification: A comprehensive review, Neural Comput. 29 (9) (2017) 2352–2449.
[21] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, Ramesh Jain, Content-based image retrieval at the end of the early years, IEEE Trans. Pattern Anal. Mach. Intell. (12) (2000) 1349–1380.
[22] Deepti Deshwal, Pardeep Sangwan, Divya Kumar, Feature extraction methods in language identification: A survey, Wirel. Pers. Commun. 107 (4) (2019) 2071–2103.

[23] Karen Spärck Jones, A statistical interpretation of term specificity and its application in retrieval, J. Doc. (2004).
[24] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (6) (1990) 391.
[25] Thomas Hofmann, Probabilistic latent semantic indexing, in: ACM SIGIR Forum, vol. 51, ACM, 2017, pp. 211–218.
[26] David M. Blei, Andrew Y. Ng, Michael I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (Jan) (2003) 993–1022.
[27] Kang Xu, Guilin Qi, Junheng Huang, Tianxing Wu, Xuefeng Fu, Detecting bursts in sentiment-aware topics from social media, Knowl.-Based Syst. 141 (2018) 44–54.
[28] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, Christian Jauvin, A neural probabilistic language model, J. Mach. Learn. Res. 3 (Feb) (2003) 1137–1155.
[29] Tomas Mikolov, Kai Chen, Greg Corrado, Jeffrey Dean, Efficient estimation of word representations in vector space, 2013, arXiv preprint arXiv:1301.3781.
[30] Max Jaderberg, Andrea Vedaldi, Andrew Zisserman, Deep features for text spotting, in: European Conference on Computer Vision, Springer, 2014, pp. 512–528.
[31] Xiao Yang, Craig Macdonald, Iadh Ounis, Using word embeddings in Twitter election classification, Inf. Retr. J. 21 (2–3) (2018) 183–207.
[32] Mingyang Jiang, Yanchun Liang, Xiaoyue Feng, Xiaojing Fan, Zhili Pei, Yu Xue, Renchu Guan, Text classification based on deep belief network and softmax regression, Neural Comput. Appl. 29 (1) (2018) 61–70.
[33] Weiling Chen, Chai Kiat Yeo, Chiew Tong Lau, Bu Sung Lee, Leveraging social media news to predict stock index movement using RNN-boost, Data Knowl. Eng. 118 (2018) 14–24.
[34] Nut Limsopatham, Nigel Henry Collier, Bidirectional LSTM for named entity recognition in Twitter messages, COLING 2016, 2016.
[35] John R. Smith, Shih-Fu Chang, Tools and techniques for color image retrieval, in: Storage and Retrieval for Still Image and Video Databases IV, vol. 2670, International Society for Optics and Photonics, 1996, pp. 426–437.
[36] Peizhong Liu, Jing-Ming Guo, Kosin Chamnongthai, Heri Prasetyo, Fusion of color histogram and LBP-based features for texture image retrieval and classification, Inform. Sci. 390 (2017) 95–111.
[37] Dennis Dunn, William E. Higgins, Optimal Gabor filters for texture segmentation, IEEE Trans. Image Process. 4 (7) (1995) 947–964.
[38] Tim F. Cootes, Christopher J. Taylor, Statistical models of appearance for medical image analysis and computer vision, in: Medical Imaging 2001: Image Processing, vol. 4322, International Society for Optics and Photonics, 2001, pp. 236–248.
[39] Xinyue Zhao, Zaixing He, Shuyou Zhang, Dong Liang, Robust pedestrian detection in thermal infrared imagery using a shape distribution histogram feature and modified sparse representation classification, Pattern Recognit. 48 (6) (2015) 1947–1960.
[40] Hongpeng Yin, Xuguo Jiao, Yi Chai, Bin Fang, Scene classification based on single-layer SAE and SVM, Expert Syst. Appl. 42 (7) (2015) 3368–3380.
[41] Jingkuan Song, Hanwang Zhang, Xiangpeng Li, Lianli Gao, Meng Wang, Richang Hong, Self-supervised video hashing with hierarchical binary auto-encoder, IEEE Trans. Image Process. 27 (7) (2018) 3210–3221.
[42] Alex Krizhevsky, Ilya Sutskever, Geoffrey E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[43] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, Yoshua Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[44] Karen Simonyan, Andrew Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[45] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1–9.
[46] Alec Radford, Luke Metz, Soumith Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, 2015, arXiv preprint arXiv:1511.06434.
[47] Lawrence Rabiner, Biing-Hwang Juang, Fundamentals of Speech Recognition, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1993.
[48] Daniel Jurafsky, James Martin, Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, volume 2, 2008.
[49] Selver Ezgi Küçükbay, Mustafa Sert, Audio-based event detection in office live environments using optimized MFCC-SVM approach, in: Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing, IEEE ICSC 2015, IEEE, 2015, pp. 475–480.
[50] Min Xu, Namunu C. Maddage, Changsheng Xu, Mohan Kankanhalli, Qi Tian, Creating audio keywords for event detection in soccer video, in: 2003 International Conference on Multimedia and Expo. ICME'03. Proceedings (Cat. No. 03TH8698), vol. 2, IEEE, 2003, pp. II–281.
[51] Elliot Singer, Pedro A Torres-Carrasquillo, Terry P Gleason, William M Campbell, Douglas A Reynolds, Acoustic, phonetic, and discriminative approaches to automatic language identification, in: Eighth European Conference on Speech Communication and Technology, 2003.
[52] Moira Yip, Tone, Cambridge University Press, 2002.
[53] Shervin Malmasi, Aoife Cahill, Measuring Feature Diversity in Native Language Identification, 2015, http://dx.doi.org/10.3115/v1/W15-0606.
[54] Shervin Malmasi, Eshrag Refaee, Mark Dras, Arabic dialect identification using a parallel multidialectal corpus, in: Conference of the Pacific Association for Computational Linguistics, Springer, 2015, pp. 35–53.
[55] Laura Fiorini, Gianmaria Mancioppi, Francesco Semeraro, Hamido Fujita, Filippo Cavallo, Unsupervised emotional state classification through physiological parameters for social robotics applications, Knowl.-Based Syst. (2019) 105217.
[56] Kay L O'Halloran, Sabine Tan, Duc-Son Pham, John Bateman, Andrew Vande Moere, A digital mixed methods research design: Integrating multimodal analysis with data mining and information visualization for big data analytics, J. Mix. Methods Res. 12 (1) (2018) 11–30.
[57] Harold Hotelling, Relations between two sets of variates, Biometrika 28 (3/4) (1936) 321–377.
[58] Fei Wu, Hong Zhang, Yueting Zhuang, Learning semantic correlations for cross-media retrieval, in: IEEE International Conference on Image Processing, 2007, pp. 1465–1468.
[59] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert RG Lanckriet, Roger Levy, Nuno Vasconcelos, A new approach to cross-modal multimedia retrieval, in: Proceedings of the 18th ACM International Conference on Multimedia, ACM, 2010, pp. 251–260.
[60] Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Nikhil Rasiwasia, Gert RG Lanckriet, Roger Levy, Nuno Vasconcelos, On the role of correlation and abstraction in cross-modal multimedia retrieval, IEEE Trans. Pattern Anal. Mach. Intell. 36 (3) (2014) 521–535.
[61] Shotaro Akaho, A kernel method for canonical correlation analysis, 2006, arXiv preprint cs/0609071.
[62] Galen Andrew, Raman Arora, Jeff Bilmes, Karen Livescu, Deep canonical correlation analysis, in: International Conference on Machine Learning, 2013, pp. 1247–1255.
[63] Jie Shao, Leiquan Wang, Zhicheng Zhao, Anni Cai, et al., Deep canonical correlation analysis with progressive and hypergraph learning for cross-modal retrieval, Neurocomputing 214 (2016) 618–628.
[64] Po-Yao Huang, Junwei Liang, Jean-Baptiste Lamare, Alexander G. Hauptmann, Multimodal filtering of social media for temporal monitoring and event analysis, in: Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval, ACM, 2018, pp. 450–457.
[65] Kah Seng, Li Minn Ang, Chien Ooi, A combined rule-based and machine learning audio-visual emotion recognition approach, IEEE Trans. Affect. Comput. PP (99) (2018) 1.
[66] Yun Ma, Qing Li, Zhenguo Yang, Zheng Lu, Haiwei Pan, Antoni B. Chan, An SVD-based multimodal clustering method for social event detection, in: IEEE International Conference on Data Engineering Workshops, 2015, pp. 202–209.
[67] Sunil Kumar Gupta, Dinh Phung, Brett Adams, Svetha Venkatesh, Regularized nonnegative shared subspace learning, Data Min. Knowl. Discov. 26 (1) (2013) 57–97.
[68] Xiaohui Huang, Yunming Ye, Liyan Xiong, Shaokai Wang, Xiaofei Yang, Clustering time-stamped data using multiple nonnegative matrices factorization, Knowl.-Based Syst. 114 (2016) 88–98.
[69] Sunil Kumar Gupta, Dinh Phung, Brett Adams, Truyen Tran, Svetha Venkatesh, Nonnegative shared subspace learning and its application to social media retrieval, in: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2010, pp. 1169–1178.
[70] Zhe Xue, Guorong Li, Weigang Zhang, Junbiao Pang, Qingming Huang, Topic detection in cross-media: a semi-supervised co-clustering approach, Int. J. Multimed. Inf. Retr. 3 (3) (2014) 193–205.
[71] Liu Yang, Liping Jing, Michael K. Ng, Robust and non-negative collective matrix factorization for text-to-image transfer learning, IEEE Trans. Image Process. 24 (12) (2015) 4701–4714.
[72] Ziyu Guan, Lijun Zhang, Jinye Peng, Jianping Fan, Multi-view concept learning for data representation, IEEE Trans. Knowl. Data Eng. 27 (11) (2015) 3016–3028.
[73] Gianluca Monaci, Philippe Jost, Pierre Vandergheynst, Boris Mailhe, Sylvain Lesage, Rémi Gribonval, Learning multi-modal dictionaries: Application to audiovisual data, in: International Workshop on Multimedia Content Representation, Classification and Security, Springer, 2006, pp. 538–545.

[74] Fan Zhu, Ling Shao, Mengyang Yu, Cross-modality submodular dictionary learning for information retrieval, in: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, ACM, 2014, pp. 1479–1488.
[75] Zhou Yu, Fei Wu, Yi Yang, Qi Tian, Jiebo Luo, Yueting Zhuang, Discriminative coupled dictionary hashing for fast cross-media retrieval, in: Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval, ACM, 2014, pp. 395–404.
[76] Xing Xu, Yang Yang, Atsushi Shimada, Rin-ichiro Taniguchi, Li He, Semi-supervised coupled dictionary learning for cross-modal retrieval in internet images and texts, in: Proceedings of the 23rd ACM International Conference on Multimedia, ACM, 2015, pp. 847–850.
[77] Lingyang Chu, Yanyan Zhang, Guorong Li, Shuhui Wang, Weigang Zhang, Qingming Huang, Effective multimodality fusion framework for cross-media topic detection, IEEE Trans. Circuits Syst. Video Technol. 26 (3) (2016) 556–569.
[78] Xiao Cai, Feiping Nie, Heng Huang, Farhad Kamangar, Heterogeneous image feature integration via multi-modal spectral clustering, in: CVPR 2011, IEEE, 2011, pp. 1977–1984.
[79] Abhishek Kumar, Piyush Rai, Hal Daume, Co-regularized multi-view spectral clustering, in: Advances in Neural Information Processing Systems, 2011, pp. 1413–1421.
[80] Georgios Petkos, Symeon Papadopoulos, Emmanouil Schinas, Yiannis Kompatsiaris, Graph-based multimodal clustering for social event detection in large collections of images, in: International Conference on Multimedia Modeling, Springer, 2014, pp. 146–158.
[81] Hanghang Tong, Jingrui He, Mingjing Li, Changshui Zhang, Wei-Ying Ma, Graph based multi-modality learning, in: Proceedings of the 13th Annual ACM International Conference on Multimedia, ACM, 2005, pp. 862–871.
[82] Petros Daras, Stavroula Manolopoulou, Apostolos Axenopoulos, Search and retrieval of rich media objects supporting multiple multimodal queries, IEEE Trans. Multimed. 14 (3) (2011) 734–746.
[83] Yang Bao, Nigel Collier, Anindya Datta, A partially supervised cross-collection topic model for cross-domain text classification, in: Proceedings of the 22nd ACM International Conference on Information & Knowledge Management, ACM, 2013, pp. 239–248.
[84] Daniel Ramage, Paul Heymann, Christopher D Manning, Hector Garcia-Molina, Clustering the tagged web, in: Proceedings of the Second ACM International Conference on Web Search and Data Mining, ACM, 2009, pp. 54–63.
[85] Jitao Sang, Changsheng Xu, Right buddy makes the difference: An early exploration of social relation analysis in multimedia applications, in: Proceedings of the 20th ACM International Conference on Multimedia, ACM, 2012, pp. 19–28.
[86] Jingwen Bian, Yang Yang, Hanwang Zhang, Tat-Seng Chua, Multimedia summarization for social events in microblog stream, IEEE Trans. Multimed. 17 (2) (2015) 216–228.
[87] Hongyun Cai, Yang Yang, Xuefei Li, Zi Huang, What are popular: exploring Twitter features for event detection, tracking and visualization, in: Proceedings of the 23rd ACM International Conference on Multimedia, ACM, 2015, pp. 89–98.
[88] Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, Boosted multi-modal supervised latent Dirichlet allocation for social event classification, in: 2014 22nd International Conference on Pattern Recognition, ICPR, IEEE, 2014, pp. 1999–2004.
[89] Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, M Shamim Hossain, Social event classification via boosted multimodal supervised latent Dirichlet allocation, ACM Trans. Multimed. Comput. Commun. Appl. 11 (2) (2015) 27.
[90] Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, Multi-modal supervised latent dirichlet allocation for event classification in social media, in: Proceedings of International Conference on Internet Multimedia Computing and Service, ACM, 2014, p. 152.
[91] Shengsheng Qian, Tianzhu Zhang, Changsheng Xu, Jie Shao, Multi-modal event topic model for social event analysis, IEEE Trans. Multimed. 18 (2) (2016) 233–246.
[92] Nitish Srivastava, Ruslan R. Salakhutdinov, Multimodal learning with deep boltzmann machines, in: Advances in Neural Information Processing Systems, 2012, pp. 2222–2230.
[93] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y. Ng, Multimodal deep learning, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 689–696.
[94] Quan Guo, Jia Jia, Guangyao Shen, Lei Zhang, Lianhong Cai, Zhang Yi, Learning robust uniform features for cross-media social data by using cross autoencoders, Knowl.-Based Syst. 102 (2016) 64–75.
[95] Fangxiang Feng, Xiaojie Wang, Ruifan Li, Ibrar Ahmad, Correspondence autoencoders for cross-modal retrieval, ACM Trans. Multimed. Comput. Commun. Appl. 12 (1s) (2015) 26.
[96] Fangxiang Feng, Xiaojie Wang, Ruifan Li, Cross-modal retrieval with correspondence autoencoder, in: Proceedings of the 22nd ACM International Conference on Multimedia, ACM, 2014, pp. 7–16.
[97] Chaoqun Hong, Jun Yu, Jian Wan, Dacheng Tao, Meng Wang, Multimodal deep autoencoder for human pose recovery, IEEE Trans. Image Process. 24 (12) (2015) 5659–5670.
[98] Weiran Wang, Raman Arora, Karen Livescu, Jeff Bilmes, On deep multi-view representation learning, in: International Conference on Machine Learning, 2015, pp. 1083–1092.
[99] Fangxiang Feng, Ruifan Li, Xiaojie Wang, Deep correspondence restricted Boltzmann machine for cross-modal retrieval, Neurocomputing 154 (2015) 50–60.
[100] Sungeun Hong, Woobin Im, Hyun S. Yang, Content-based video-music retrieval using soft intra-modal structure constraint, CoRR abs/1704.06761 (2017).
[101] Y. Wei, Y. Zhao, C. Lu, S. Wei, L. Liu, Z. Zhu, S. Yan, Cross-modal retrieval with CNN visual features: A new baseline, IEEE Trans. Cybern. 47 (2) (2016) 449–460.
[102] Yonghao He, Shiming Xiang, Cuicui Kang, Jian Wang, Chunhong Pan, Cross-modal retrieval via deep and bidirectional representation learning, IEEE Trans. Multimed. 18 (7) (2016) 1363–1377.
[103] Yoon Kim, Convolutional neural networks for sentence classification, 2014, arXiv preprint arXiv:1408.5882.
[104] Chaoqun Hong, Jun Yu, Jian Zhang, Xiongnan Jin, Kyong-Ho Lee, Multi-modal face pose estimation with multi-task manifold deep learning, IEEE Trans. Ind. Inf. (2018).
[105] Yue Gao, Hanwang Zhang, Xibin Zhao, Shuicheng Yan, Event classification in microblogs via social tracking, ACM Trans. Intell. Syst. Technol. (TIST) 8 (3) (2017) 35.
[106] Donald Metzler, Congxing Cai, Eduard Hovy, Structured event retrieval over microblog archives, in: Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, 2012, pp. 646–655.
[107] Dong Zhou, Séamus Lawless, Vincent Wade, Improving search via personalized query expansion using social media, Inf. Retr. 15 (3–4) (2012) 218–242.
[108] Theresa Marie Bernardo, Andrijana Rajic, Ian Young, Katie Robiadek, Mai T Pham, Julie A Funk, Scoping review on search queries and social media for disease surveillance: a chronology of innovation, J. Med. Internet Res. 15 (7) (2013) e147.
[109] Jingkuan Song, Yang Yang, Yi Yang, Zi Huang, Heng Tao Shen, Inter-media hashing for large-scale retrieval from heterogeneous data sources, in: Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data, ACM, 2013, pp. 785–796.
[110] Lei Zhu, Zi Huang, Xiaobai Liu, Xiangnan He, Jiande Sun, Xiaofang Zhou, Discrete multimodal hashing with canonical views for robust mobile landmark search, IEEE Trans. Multimed. 19 (9) (2017) 2066–2079.
[111] Jinhui Tang, Zechao Li, Weakly supervised multimodal hashing for scalable social image retrieval, IEEE Trans. Circuits Syst. Video Technol. 28 (10) (2017) 2730–2741.
[112] Lei Gao, Ling Guan, Information fusion via multimodal hashing with discriminant correlation maximization, in: 2019 IEEE International Conference on Image Processing, ICIP, IEEE, 2019, pp. 2224–2228.
[113] Jianshu Weng, Bu-Sung Lee, Event detection in Twitter, in: Fifth International AAAI Conference on Weblogs and Social Media, 2011.
[114] Marina Litvak, Natalia Vanetik, Efi Levi, Michael Roistacher, What's up on Twitter? Catch up with TWIST!, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations, 2016, pp. 213–217.
[115] Asaad Hakeem, Yaser Sheikh, Mubarak Shah, CASEˆ e: A Hierarchical Event Representation for the Analysis of Videos, in: AAAI, 2004, pp. 263–268.
[116] Lu Jiang, Alexander G. Hauptmann, Guang Xiang, Leveraging high-level and low-level features for multimedia event detection, in: Proceedings of the 20th ACM International Conference on Multimedia, ACM, 2012, pp. 449–458.
[117] Nisha Pahal, Santanu Chaudhury, Vishwanath Gaur, Brejesh Lall, Anupama Mallik, Detecting and correlating video-based event patterns: An ontology driven approach, in: Proceedings of the 2014 IEEE/WIC/ACM International Joint Conferences on Web Intelligence, WI and Intelligent Agent Technologies, IAT-Volume 01, IEEE Computer Society, 2014, pp. 438–445.
[118] Xiangmin Zhou, Lei Chen, Event detection over Twitter social media streams, VLDB J. 23 (3) (2014) 381–400.
[119] Minale A. Abebe, Joe Tekli, Fekade Getahun, Richard Chbeir, Gilbert Tekli, Generic metadata representation framework for social-based event detection, description, and linkage, Knowl.-Based Syst. (2019).

[120] Dongdong Shan, Wayne Xin Zhao, Rishan Chen, Baihan Shu, Ziqi Wang, Junjie Yao, Hongfei Yan, Xiaoming Li, Eventsearch: a system for event discovery and retrieval on multi-type historical data, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2012, pp. 1564–1567.

[121] Shize Xu, Liang Kong, Yan Zhang, A cross-media evolutionary timeline generation framework based on iterative recommendation, in: Proceedings of the 3rd ACM Conference on International Conference on Multimedia Retrieval, ACM, 2013, pp. 73–80.

[122] Timo Reuter, Philipp Cimiano, Event-based classification of social media streams, in: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ACM, 2012, p. 22.

[123] Yanxiang Wang, Hari Sundaram, Lexing Xie, Social event detection with interaction graph modeling, in: Proceedings of the 20th ACM International Conference on Multimedia, ACM, 2012, pp. 865–868.

[124] Melvin Earl Maron, Automatic indexing: an experimental inquiry, J. ACM 8 (3) (1961) 404–417.

[125] Ana-Maria Popescu, Marco Pennacchiotti, Detecting controversial events from twitter, in: Proceedings of the 19th ACM International Conference on Information and Knowledge Management, ACM, 2010, pp. 1873–1876.

[126] David A. Sadlier, Noel E. O'Connor, Event detection in field sports video using audio-visual features and a support vector machine, IEEE Trans. Circuits Syst. Video Technol. 15 (10) (2005) 1225–1233.

[127] Philipp Blandfort, Desmond U. Patton, William R. Frey, Svebor Karaman, Surabhi Bhargava, Fei-Tzin Lee, Siddharth Varia, Chris Kedzie, Michael B. Gaskell, Rossano Schifanella, et al., Multimodal social media analysis for gang violence prevention, in: Proceedings of the International AAAI Conference on Web and Social Media, vol. 13, 2019, pp. 114–124.

[128] Samar M. Alqhtani, Suhuai Luo, Brian Regan, A multiple kernel learning based fusion for earthquake detection from multimedia Twitter data, Multimedia Tools Appl. 77 (10) (2018) 12519–12532.

[129] Jagan Sankaranarayanan, Hanan Samet, Benjamin E Teitler, Michael D Lieberman, Jon Sperling, Twitterstand: news in tweets, in: Proceedings of the 17th Acm Sigspatial International Conference on Advances in Geographic Information Systems, ACM, 2009, pp. 42–51.

[130] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, Jiebo Luo, Multimodal fusion with recurrent neural networks for rumor detection on microblogs, in: Proceedings of the 25th ACM International Conference on Multimedia, ACM, 2017, pp. 795–816.

[131] Zhihong Lin, Huidong Jin, Bella Robinson, Xunguo Lin, Towards an accurate social media disaster event detection system based on deep learning and semantic representation, in: Proceedings of the 14th Australasian Data Mining Conference, Canberra, Australia, 2016, pp. 6–8.

[132] Xueliang Liu, Benoit Huet, Raphaël Troncy, EURECOM@ MediaEval 2011 social event detection task, in: MediaEval, 2011.

[133] Markus Brenner, Ebroul Izquierdo, Social event detection and retrieval in collaborative photo collections, in: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ACM, 2012, p. 21.

[134] Feng Xue, Richang Hong, Xiangnan He, Jianwei Wang, Shengsheng Qian, Changsheng Xu, Knowledge embedding based topic model for multi-modal social event analysis, IEEE Trans. Multimed. (2019).

[135] Hila Becker, Mor Naaman, Luis Gravano, Learning similarity metrics for event identification in social media, in: Proceedings of the Third ACM International Conference on Web Search and Data Mining, ACM, 2010, pp. 291–300.

[136] Lu Liu, Lifeng Sun, Yong Rui, Yao Shi, Shiqiang Yang, Web video topic discovery and tracking via bipartite graph reinforcement model, in: Proceedings of the 17th International Conference on World Wide Web, ACM, 2008, pp. 1009–1018.

[137] Jaeyoung Choi, Eungchan Kim, Martha Larson, Gerald Friedland, Alan Hanjalic, Evento 360: Social event discovery from web-scale multimedia collection, in: Proceedings of the 23rd ACM International Conference on Multimedia, ACM, 2015, pp. 193–196.

[138] Zhenguo Yang, Qing Li, Wenyin Liu, Yun Ma, Min Cheng, Dual graph regularized NMF model for social event detection from Flickr data, World Wide Web 20 (5) (2017) 995–1015.

[139] Zhenguo Yang, Qing Li, Zheng Lu, Yun Ma, Zhiguo Gong, Haiwei Pan, Semi-supervised multimodal clustering algorithm integrating label signals for social event detection, in: 2015 IEEE International Conference on Multimedia Big Data, IEEE, 2015, pp. 32–39.

[140] Zhenguo Yang, Qing Li, Zheng Lu, Yun Ma, Zhiguo Gong, Wenyin Liu, Dual structure constrained multimodal feature coding for social event detection from Flickr data, ACM Trans. Internet Technol. 17 (2) (2017) 19.

[141] Georgios Petkos, Symeon Papadopoulos, Yiannis Kompatsiaris, Social event detection using multimodal clustering and integrating supervisory signals, in: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ACM, 2012, p. 23.

[142] Jian Shao, Shuai Ma, Weiming Lu, Yueting Zhuang, A unified framework for web video topic discovery and visualization, Pattern Recognit. Lett. 33 (4) (2012) 410–419.

[143] Xueming Qian, Mingdi Li, Yayun Ren, Shuhui Jiang, Social media based event summarization by user–text–image co-clustering, Knowl.-Based Syst. 164 (2019) 107–121.

[144] Daichi Takehara, Ryosuke Harakawa, Takahiro Ogawa, Miki Haseyama, Hierarchical content group detection from different social media platforms using web link structure, in: 2016 IEEE International Conference on Image Processing, ICIP, IEEE, 2016, pp. 479–483.

[145] Yanyan Zhang, Guorong Li, Lingyang Chu, Shuhui Wang, Weigang Zhang, Qingming Huang, Cross-media topic detection: a multi-modality fusion framework, in: 2013 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2013, pp. 1–6.

[146] Weixin Li, Jungseock Joo, Hang Qi, Song-Chun Zhu, Joint image-text news topic detection and tracking by multimodal topic and-or graph, IEEE Trans. Multimed. 19 (2) (2016) 367–381.

[147] Manos Schinas, Symeon Papadopoulos, Georgios Petkos, Yiannis Kompatsiaris, Pericles A Mitkas, Multimodal graph-based event detection and summarization in social media streams, in: Proceedings of the 23rd ACM International Conference on Multimedia, ACM, 2015, pp. 189–192.

[148] Sicheng Zhao, Yue Gao, Guiguang Ding, Tat-Seng Chua, Real-time multimedia social event detection in microblog, IEEE Trans. Cybern. 48 (11) (2017) 3218–3231.

[149] Georgios Petkos, Manos Schinas, Symeon Papadopoulos, Yiannis Kompatsiaris, Graph-based multimodal clustering for social multimedia, Multimedia Tools Appl. 76 (6) (2017) 7897–7919.

[150] Bibek Luitel, Y.V. Srinivasa Murthy, Shashidhar G. Koolagudi, Sound event detection in urban soundscape using two-level classification, in: Distributed Computing, Vlsi, Electrical Circuits and Robotics, 2017.

[151] Seema Wazarkar, Bettahally N. Keshavamurthy, Ahsan Hussain, Region-based segmentation of social images using soft KNN algorithm, Procedia Comput. Sci. 125 (2018) 93–98.

[152] Maofu Liu, Weili Guan, Jie Yan, Huijun Hu, Correlation identification in multimodal weibo via back propagation neural network with genetic algorithm, J. Vis. Commun. Image Represent. 60 (2019) 312–318.

[153] Benjamin Bischke, Damian Borth, Christian Schulze, Andreas Dengel, Contextual enrichment of remote-sensed events with social media streams, in: Proceedings of the 24th ACM International Conference on Multimedia, ACM, 2016, pp. 1077–1081.

[154] Sicheng Zhao, Hongxun Yao, Sendong Zhao, Xuesong Jiang, Xiaolei Jiang, Multi-modal microblog classification via multi-task learning, Multimedia Tools Appl. 75 (15) (2016) 8921–8938.

[155] Yue Gao, Fanglin Wang, Huanbo Luan, Tat-Seng Chua, Brand data gathering from live social media streams, in: Proceedings of International Conference on Multimedia Retrieval, ACM, 2014, p. 169.

[156] Yue Gao, Yi Zhen, Haojie Li, Tat-Seng Chua, Filtering of brand-related microblogs using social-smooth multiview embedding, IEEE Trans. Multimed. 18 (10) (2016) 2115–2126.

[157] C. Fellbaum, G. Miller, WordNet:An Electronic Lexical Database, MIT Press, 1998.

[158] Imran Memon, Ling Chen, Abdul Majid, Mingqi Lv, Ibrar Hussain, Gencai Chen, Travel recommendation using geo-tagged photos in social media for tourist, in: International Conference on Computer & Emerging Technologies March, 2014, pp. 1347–1362.

[159] Joan Capdevila, Jess Cerquides, Jordi Nin, Jordi Torres, Tweet-SCAN: An event discovery technique for geo-located Tweets, Pattern Recognit. Lett. (2016) 110–119.

[160] Xiaoming Zhang, Xiaoming Chen, Yan Chen, Senzhang Wang, Zhoujun Li, Jiali Xia, Event detection and popularity prediction in microblogging, Neurocomputing 149 (2015) 1469–1480.

[161] Zhenguo Li, Xiao-Ming Wu, Shih-Fu Chang, Segmentation using superpixels: A bipartite graph partitioning approach, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 789–796.

[162] Giovanni Stilo, Paola Velardi, Efficient temporal mining of micro-blog texts and its application to event discovery, Data Min. Knowl. Discov. 30 (2) (2016) 372–402.

[163] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo, Earthquake shakes Twitter users: real-time event detection by social sensors, in: Proceedings of the 19th International Conference on World Wide Web, ACM, 2010, pp. 851–860.

[164] Takeshi Sakaki, Makoto Okazaki, Yutaka Matsuo, Tweet analysis for real-time event detection and earthquake reporting system development, IEEE Trans. Knowl. Data Eng. 25 (4) (2012) 919–931.

[165] Senzhang Wang, Xiaoming Zhang, Fengxiang Li, S Yu Philip, Zhiqiu Huang, Efficient traffic estimation with multi-sourced data by parallel coupled hidden Markov model, IEEE Trans. Intell. Transp. Syst. (2018).

[166] Yasin Yılmaz, Alfred O. Hero, Multimodal event detection in Twitter hashtag networks, J. Signal Process. Syst. 90 (2) (2018) 185–200.

[167] Alceu Ferraz Costa, Yuto Yamaguchi, Agma Juci Machado Traina, Caetano Traina Jr, Christos Faloutsos, Rsc: Mining and modeling temporal activity in social media, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, 2015, pp. 269–278.

[168] Runwei Situ, Zhenguo Yang, Jianming Lv, Qing Li, Wenyin Liu, Cross-modal event retrieval: A dataset and a baseline using deep semantic learning, in: Pacific Rim Conference on Multimedia, Springer, 2018, pp. 147–157.
[169] Mathilde Sahuguet, Benoit Huet, Socially motivated multimedia topic timeline summarization, in: Proceedings of the 2nd International Workshop on Socially-Aware Multimedia, ACM, 2013, pp. 19–24.
[170] Song Tan, Chong-Wah Ngo, Hung-Khoon Tan, Lei Pang, Cross media hyperlinking for search topic browsing, in: Proceedings of the 19th ACM International Conference on Multimedia, ACM, 2011, pp. 243–252.
[171] Rui Yan, Xiaojun Wan, Mirella Lapata, Wayne Xin Zhao, Pu-Jen Cheng, Xiaoming Li, Visualizing timelines: Evolutionary summarization via iterative reinforcement between text and image streams, in: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, ACM, 2012, pp. 275–284.
[172] Xiao Wu, Yi-Jie Lu, Qiang Peng, Chong-Wah Ngo, Mining event structures from web videos, IEEE Multimedia 18 (1) (2011) 38–51.
[173] Dingding Wang, Tao Li, Mitsunori Ogihara, Generating pictorial storylines via minimum-weight connected dominating set approximation in multi-view graphs, in: Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.
[174] Shize Xu, Shanshan Wang, Yan Zhang, Summarizing complex events: a cross-modal solution of storylines extraction and reconstruction, in: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013, pp. 1281–1291.
[175] Symeon Papadopoulos, Raphael Troncy, Vasileios Mezaris, Benoit Huet, Ioannis Kompatsiaris, Social event detection at MediaEval 2011: Challenges, dataset and evaluation, in: MediaEval, 2011.
[176] Symeon Papadopoulos, Emmanouil Schinas, Vasileios Mezaris, Raphaël Troncy, Ioannis Kompatsiaris, The 2012 social event detection dataset, in: Proceedings of the 4th ACM Multimedia Systems Conference, ACM, 2013, pp. 102–107.
[177] Timo Reuter, Symeon Papadopoulos, Giorgos Petkos, Vasileios Mezaris, Yiannis Kompatsiaris, Philipp Cimiano, Christopher de Vries, Shlomo Geva, Social event detection at mediaeval 2013: Challenges, datasets, and evaluation, in: Proceedings of the MediaEval 2013 Multimedia Benchmark Workshop Barcelona, Spain, October 18-19, 2013, 2013.
[178] Georgios Petkos, Symeon Papadopoulos, Vasileios Mezaris, Yiannis Kompatsiaris, Social event detection at MediaEval 2014: Challenges, datasets, and evaluation, in: MediaEval, Citeseer, 2014.
[179] Markus Brenner, Ebroul Izquierdo, MediaEval benchmark: Social event detection in collaborative photo collections, in: MediaEval, 2011.
[180] Truc-Vien Nguyen, Minh-Son Dao, Riccardo Mattivi, Emanuele Sansone, Francesco GB De Natale, Giulia Boato, Event clustering and classification from social media: Watershed-based and kernel methods, in: MediaEval, 2013.
[181] Sina Samangooei, Jonathon Hare, David Dupplaw, Mahesan Niranjan, Nicholas Gibbins, Paul H Lewis, Jamie Davies, Neha Jain, John Preston, Social Event Detection Via Sparse Multi-Modal Feature Selection and Incremental Density Based Clustering, 2013.
[182] Maia Zaharieva, Daniel Schopfhauser, Manfred Del Fabro, Matthias Zeppelzauer, Clustering and retrieval of social events in Flickr, in: MediaEval, 2014.
[183] Nafiz Arica, Fatos T. Yarman-Vural, Optical character recognition for cursive handwriting, IEEE Trans. Pattern Anal. Mach. Intell. 24 (6) (2002) 801–813.
[184] Chirag Patel, Atul Patel, Dharmendra Patel, Optical character recognition by open source OCR tool tesseract: A case study, Int. J. Comput. Appl. 55 (10) (2012) 50–56.
[185] Martin Cooke, Phil Green, Ljubomir Josifovski, Ascension Vizinho, Robust automatic speech recognition with missing and unreliable acoustic data, Speech Commun. 34 (3) (2001) 267–285.
[186] Dong Yu, Li Deng, Automatic Sppech Recognition, Springer, 2016.
[187] Jun Yu, Min Tan, Hongyuan Zhang, Dacheng Tao, Yong Rui, Hierarchical deep click feature prediction for fine-grained image recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2019).
[188] Jun Yu, Dacheng Tao, Meng Wang, Yong Rui, Learning to rank using user clicks and visual features for image retrieval, IEEE Trans. Cybern. 45 (4) (2014) 767–779.
[189] Xiaoyang Liu, Daobing He, Chao Liu, Information diffusion nonlinear dynamics modeling and evolution analysis in online social network based on emergency events, IEEE Trans. Comput. Soc. Syst. 6 (1) (2019) 8–19.
[190] Ramesh S. Wadawadagi, Veerappa B. Pagi, Sentiment analysis on social media: Recent trends in machine learning, in: Handbook of Research on Emerging Trends and Applications of Machine Learning, IGI Global, 2020, pp. 508–527.
[191] Nicola Capuano, Francisco Chiclana, Hamido Fujita, Enrique Herrera-Viedma, Vincenzo Loia, Fuzzy group decision making with incomplete information guided by social influence, IEEE Trans. Fuzzy Syst. 26 (3) (2017) 1704–1718.
[192] Hamido Fujita, Angelo Gaeta, Vincenzo Loia, Francesco Orciuoli, Hypotheses analysis and assessment in counter-terrorism activities: a method based on OWA and fuzzy probabilistic rough sets, IEEE Trans. Fuzzy Syst. (2019).