Mining Multi-Aspect Reflection of News Events in Twitter: Discovery, Linking and Presentation

Jingjing Wang*, Wenzhu Tong*, Hongkun Yu*, Min Li*, Xiuli Ma†, Haoyan Cai*, Tim Hanratty‡ and Jiawei Han*
*Department of Computer Science, University of Illinois at Urbana-Champaign, IL, USA
†School of Electronics Engineering and Computer Science, Peking University, Beijing, China
‡Information Sciences Directorate, Army Research Laboratory, MD, USA
*{jwang112, wtong8, hyu50, minli3, hcai6, hanj}@illinois.edu, †xlma@pku.edu.cn, ‡timothy.p.hanratty@mail.mil

***Abstract*— A major event often has repercussions on both news media and microblogging sites such as Twitter. Reports from mainstream news agencies and discussions from Twitter complement each other to form a complete picture. An event can have multiple aspects (sub-events) describing it from multiple angles, each of which attracts opinions/comments posted on Twitter. Mining such reflections is interesting to both policy makers and ordinary people seeking information. In this paper, we propose a unified framework to mine multi-aspect reflections of news events in Twitter. We propose a novel and efficient dynamic hierarchical entity-aware event discovery model to learn news events and their multiple aspects. The aspects of an event are linked to their reflections in Twitter by a bootstrapped dataless classification scheme, which elegantly handles the challenges of selecting informative tweets under overwhelming noise and bridging the vocabularies of news and tweets. In addition, we demonstrate that our framework naturally generates an informative presentation of each event with entity graphs, time spans, news summaries and tweet highlights to facilitate user digestion.**

I. Introduction

Once an influential event takes place, mainstream news media immediately react to it. News reports deliver real-time status of the event, covering every aspect with fairly standard languages. Informed by these reports, people post their opinions/comments and raise discussions on the event via microblogging sites such as Twitter. The different natures of these two sources provide a complementary view of an event: A reasonably objective and comprehensive presentation of an event, and a view full of opinions and sentiments from the public. Linking them together to provide a complete picture of an event can be of great interest to both policy makers and ordinary people seeking information.

Preliminary research towards this direction include [1], which finds the most relevant news articles to enrich a given tweet; and [2], which retrieves related social media utterances to a given news article. However, either a single tweet or a single news article has limited expressing power, even if the original piece of information is enriched by the retrieved counterpart.

In this paper, we take a global perspective and offer event level summaries of both sources simultaneously. Consider a newly inaugurated mayor who would like to know what the public opinions are about major events in the past two weeks. The following capabilities are desirable: 1) What are the major events; 2) who are the key players in each event; 3) how people talk about each event; and 4) when is the event and how long does the event last?

In addition, we notice that a major event can have multiple aspects. For example, the `Sony Pictures Entertainment Hack`[1] event around December 2014 a) raises doubts on if North Korea is responsible for the hack; b) unexpectedly promotes the film "the Interview" and leads to a big success for its online release; and c) attracts attention from the White House. Each aspect has different focuses both in the sense of key players involved and the diverse public opinions. Therefore, the mining process should be able to distinguish different aspects for each event to present a holistic view.

To this end, we propose a unified framework for mining multi-aspect reflections of news events in Twitter. We aim to detect major events as well as the multiple aspects of each event. An aspect of an event is characterized by both a set of news articles which emphasize objective facts and a set of relevant tweets which contain rich opinions and sentiments. Using the previous example, aspect (b) of the `Sony Hack` event can be described by news articles with headlines like
"Sony plans a limited release for film on Christmas Day"
and tweets like
"I told y'all Sony would release The Interview. This has been the most shameless promotion for a film I've ever seen."

Challenges. We aim to address two major challenges while designing our model. First, we need to discover the "anchors" to link news and tweets. With a collection of news articles and huge numbers of random or informative tweets, the challenge is how to discover and formally define the interesting events and aspects, based on which to link the two sources. Second, the language models of news and tweets are drastically different. The linking process should be able to bridge the vocabularies between news and tweets as well as to accommodate different modeling of long text (news) and short text (tweets). While news-related tweets do share a critical proportion of keywords with the news articles, there exists non-negligible vocabulary gap between them [3].

Anchor Discovery. In our proposal, anchor discovery is achieved by a comprehensive study of news solely instead

[1]http://en.wikipedia.org/wiki/Sony_Pictures_Entertainment_hack
We use `Sony Hack` in the rest of this paper for brevity.

1550-4786/15 $31.00 © 2015 IEEE
DOI 10.1109/ICDM.2015.112

of mixing these two sources at the early stage, in light of the high quality, less noise and broad coverage of news articles. To learn the optimal representation of the news events and their multiple aspects, we propose a novel and efficient *generative model* with an elegant recursive *decomposition strategy* for dynamic hierarchical entity-aware event/aspect discovery. The hierarchical structure is illustrated in Figure 1. The root node denotes the entire news collection, from which events are learned. Each event has a number of child nodes which denote aspects of this event[2]. *The event/aspect discovery is essentially a top-down hierarchical clustering procedure which recursively applies the generative model.* Our proposed decomposition strategy (Section IV-C) complies with the fact that aspect nodes originate from the same theme of their parent event node, while each aspect has its distinct emphasis. The generative model integrates the most critical dimensions for clustering, including text, entities (person/location/organization) and time in a unified manner. These dimensions mutually reinforce each other to boost coherence. A node is characterized by a word distribution, a set of entity distributions (with respect to person, location and organization), and a time distribution. These distributions form an accurate multidimensional *descriptor* for an event/aspect comprehensively.

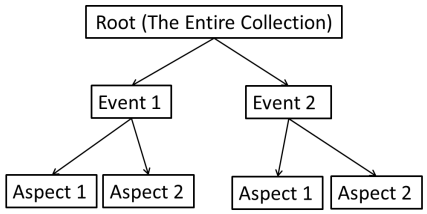

Fig. 1. Event-Aspect Hierarchy.

Linking. The event/aspect descriptors are then utilized to guide the reflection mining. The goal is to investigate how various aspects of an event are discussed in Twitter. This is formulated as a bootstrapped dataless[3] multi-class classification problem [4]. Specifically, for each event, we first form a pool of candidate tweets out of the high-volume tweet stream by information retrieval with the multidimensional event descriptor. A retrieval model is proposed to retrieve tweets which achieve simultaneously textual, entity and temporal relevance to the event. Within the candidate pool, we use the aspect descriptors to select their corresponding initial confident sets of tweets (seeds). Then by bootstrapping we select and classify the candidates into different aspects until the number of tweets for each aspect meets a threshold. We can see that the entire process is unsupervised and no labeled data is required. Furthermore, the classifier is able to accommodate various local or global features. More significantly, the bootstrapping scheme not only benefits the classification accuracy itself, but also naturally handles the vocabulary gap between news and tweets.

Presentation. Aside from discovery and linking, how to present the well-sorted information to the end-users is non-trivial. For each aspect of an event, our framework naturally supports a user friendly presentation with an entity graph, a time span, a news summary and a tweet highlight for user digestion.

The last contribution of the paper is the capability to create an aspect-specific and time-aware event dataset for an arbitrary time period, which prepares fine input for various applications such as opinion mining/comparison, multi-corpus text summarization and information diffusion.

The rest of the paper is organized as follows. We state the problem in Section II, followed by our proposed solution in Section III. We present the key components of our solution, event/aspect discovery and tweets linking in Sections IV and V. In Section VI, we evaluate the proposed solution. Then we review the related work in Section VII, and conclude in Section VIII.

II. Problem Formulation

We formulate our problem in this section. The notations used in this paper are summarized in Table I.

TABLE I
Summary of Notations

Symbol	Description
$\mathbf{X}^{\mathtt{w}}$	word matrix
$\mathbf{X}^{\mathtt{e}}$	entity matrix, $\mathtt{e} = \mathtt{p}, \mathtt{l}, \mathtt{o}$
$\mathbf{t}$	time vector
$\mathcal{I}$	input data associated with a node in the hierarchy. $\mathcal{I}_0$ denotes the root node, $\mathcal{I}_z$ denotes an event/aspect node
$\phi^{\mathtt{w}}$	word distribution
$\phi^{\mathtt{e}}$	entity distribution, $\mathtt{e} = \mathtt{p}, \mathtt{l}, \mathtt{o}$
μ, σ	parameters of time distribution
z	event/aspect ID; event/aspect descriptor: $z = \{\phi^{\mathtt{w}}, \{\phi^{\mathtt{e}}\}, \mu, \sigma\}$
θ	per-document topic distribution. The topic can be event or aspect

DEFINITION 1 (News Article) *A news article is defined by a bag-of-words/entities model with a timestamp. The entities can be persons, locations or organizations*[4].

A collection of news articles are thus compactly represented by 1) a $N^{\mathtt{w}} \times D$ word matrix $\mathbf{X}^{\mathtt{w}}$ where an entry $x^{\mathtt{w}}_{wd}$ denotes how many times the w-th word appears in the d-th news article; 2) three $N^{\mathtt{e}} \times D$ entity matrices $\{\mathbf{X}^{\mathtt{e}}\}$, where $\mathtt{e}$ can be the type person, location, or organization, *i.e.*, $\mathtt{e} = \mathtt{p}, \mathtt{l}, \mathtt{o}$. An entry $x^{\mathtt{e}}_{ed}$ denotes how many times the e-th entity appears in the d-th news article; 3) a time vector $\mathbf{t}$ where t_d denotes the timestamp of the d-th news article.

DEFINITION 2 (Tweet) *A tweet is also defined by a bag-of-words/entities model with a timestamp.*

[2]Our algorithm allows each aspect to have sub-aspects as well.

[3]The name *labelless* classification may be more accurate and intuitive but we follow the terminology *dataless* due to historical reasons.

[4]Entities are extracted by NLP tools from the news content in a preprocessing step. See details in the Experiment section.

DEFINITION 3 (Event/Aspect) *Events and aspects are nodes in a topically coherent hierarchy. Both an event node and an aspect node is defined by textual, entity and temporal dimensions. Formally, it is defined by 1) a multinomial word distribution* $\phi^{\mathtt{w}}$*; 2) a set of entity distributions* $\{\phi^{\mathtt{e}}\}$*,* $\mathtt{e} = \mathtt{p}, \mathtt{l}, \mathtt{o}$*, where* $\phi^{\mathtt{p}}, \phi^{\mathtt{l}}, \phi^{\mathtt{o}}$ *are all multinomial distributions; and 3) a Gaussian time distribution* $\mathcal{N}(\mu, \sigma)$*.*

DEFINITION 4 (Event/Aspect Descriptor) *We denote an event/aspect descriptor by* $z = \{\phi^{\mathtt{w}}, \{\phi^{\mathtt{e}}\}, \mu, \sigma\}$*.*

DEFINITION 5 (Reflection) *The reflection of an aspect of a news event is the set of relevant tweets to the aspect, which will be identified by the event and aspect descriptors.*

With the definitions above, we are now able to formulate our problem as follows.

PROBLEM 1 ***Event-Based Multi-Aspect Reflection Mining*** *Given a collection of news articles and a collection of tweets within a query time period, learn the events during the period and the multiple aspects of each event; find the reflections in twitter; and present the multi-aspect events and their reflections to end users.*

III. Overview of the Event-Based Multi-Aspect Reflection Mining Framework

To mine the reflections of multiple aspects of a news event, we propose a framework that can be divided into two main parts: event and aspect discovery in news, and linking with relevant tweets. Our process for event and aspect discovery in news involves a dynamic hierarchical entity-aware generative model with an elegant recursive decomposition strategy. After learning the accurate event and aspect descriptors via the generative model, we perform a bootstrapped dataless multi-class classification using the descriptors for identifying relevant tweets.

The goal of our generative model is to provide accurate descriptors for each event and aspect. The model leverages text, entities and time jointly in the generative process to enforce coherence through all these dimensions. The estimated distributions of words, entities and time form comprehensive event/aspect descriptors, which are the input for the following tweets linking part. For the construction of the event-aspect hierarchy, we propose a recursive decomposition strategy which naturally a) encodes the intuition that aspect nodes originate from the same theme of their parent event node, while each aspect has its distinct emphasis, b) supports a lazy learning protocol for efficient query processing: the aspects of an event are not learned until a user queries to expand the event.

Tweets are by nature noisy, informally written and filled up with all kinds of information. Identifying the relevant tweets discussing a particular aspect of an event is useful yet challenging. We address this by proposing a *retrieval + bootstapped dataless classification* procedure. For each event, with the event descriptor, we first devise a multidimensional retrieval model to retrieve an initial pool of tweets. Then with the aspect descriptors, we select informative tweets for each aspect iteratively by bootstrapping, which elegantly bridges the vocabulary gap between news and tweets. We expound upon our event/aspect discovery algorithm and tweets linking procedure in Section IV and Section V, respectively.

IV. Event and Aspect Discovery in News

As discussed in Section II, events and aspects are viewed as nodes in a topically coherent hierarchy. We propose a unified generative model for recursive construction of the hierarchy in a top-down manner. Essentially, it is a top-down hierarchical clustering process.

Step 1. Construct $\mathcal{I}_0 = \{\mathbf{X}^{\mathtt{w}}, \{\mathbf{X}^{\mathtt{e}}\}, \mathbf{t}\}$ using the entire collection of news. $\mathcal{I}_0$ is the input associated with the root node for inducing the event nodes.

Step 2. Induce the child nodes (events) of the root node taking $\mathcal{I}_0$ as input using the proposed generative model. The model estimates the descriptor $z = \{\phi^{\mathtt{w}}, \{\phi^{\mathtt{e}}\}, \mu, \sigma\}$ for each child node. We associate node z[5] with $\mathcal{I}_z$, which is generated by decomposing $\mathcal{I}_0$ to node z. Specifically, $\mathcal{I}_z = \{\mathbf{X}_z^{\mathtt{w}}, \{\mathbf{X}_z^{\mathtt{e}}\}, \mathbf{t}\}$, where $\sum_z \mathbf{X}_z^{\mathtt{w}} = \mathbf{X}^{\mathtt{w}}, \sum_z \mathbf{X}_z^{\mathtt{e}} = \mathbf{X}^{\mathtt{e}}, \mathtt{e} = \mathtt{p}, \mathtt{l}, \mathtt{o}$.

Step 3. Apply Step 2 to each event node z to induce the child nodes (aspects).

Recursively applying step 2 will further give sub-aspects, sub-sub-aspects and so on. Whether to split a node and how many child nodes to use depend on how "big" the current node is. We make this decision based on the logarithm of the amplitude of the matrices in $\mathcal{I}_z$. In this paper, a two-level hierarchy is constructed, *i.e.*, the event level and the aspect level. However, our experiment system is implemented in a user-interactive fashion where users can decide the layout of the hierarchy with varying depth and branching numbers. We describe the key Step 2 (the generative model and the decomposition strategy for hierarchy construction) in the following sections.

A. The Generative Model

Our model assumes that each news article is generated by a mixture of topics (At the event level the topic denotes an event and at the aspect level it denotes an aspect.) governed by a multinomial topic distribution θ. The topic distribution is shared by each dimension, *i.e.*, text, entities and time. This is motivated by the intuition that all the above dimensions should be coherent for each topic: a news article is more likely to belong to a particular topic when its text, entities and timestamp all have high probabilities of belonging to the topic. For instance, a news article which contains words like *"film"*, *"release"*, entities like *"sony entertainment"*, *"Seth Rogen"*(the director of the film), and was published around *December 25, 2014*, would have high probability of belonging to the "film release" aspect of the `Sony Hack` event. Any single dimension is not sufficient for the conclusion.

[5]In this paper, we slightly abuse the notation of z which is used both as a descriptor of a node and the ID of the node.

Another important design of our model is to introduce a background topic B[6], which not only serves the traditional purpose of attracting the collection's aggregate characteristics for making other discovered topics more discriminative, but also enables an elegant decomposition strategy to construct the hierarchy. Under our decomposition strategy, we will see in what follows that the descriptor of the background topic for a set of nodes turns out to be exactly the descriptor of their parent node. In other words, the background topic of an aspect has the same representation with that of the corresponding parent event. This matches the intuition that a news article is a mixture of an event background topic and a set of aspect topics.

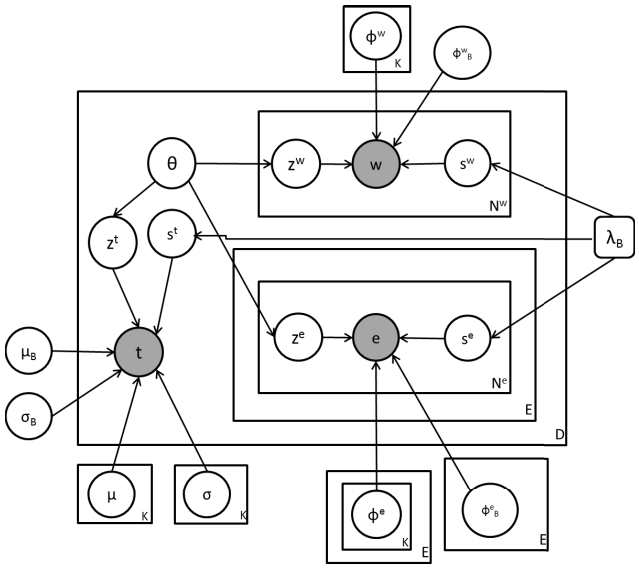

Fig. 2. Plate Notation: News Learning Module

The plate notation for the generative model is shown in Figure 2. We observe words, entities and the timestamp for each news article and estimate the parameters $\mathbf{\Theta} = \{\{\theta\}, \{\phi^{\mathtt{w}}\}, \{\phi^{\mathtt{p}}\}, \{\phi^{\mathtt{l}}\}, \{\phi^{\mathtt{o}}\}, \{\mu\}, \{\sigma\}\}$. The generative process is as follows:

To generate each word in news article d,

1. Draw a switch variable $s^{\mathtt{w}} \sim Bernoulli(\lambda_B)$. λ_B is the topic proportion of the background topic B.
2. If $s^{\mathtt{w}} = 1$,

   draw a word w from the background topic B: $w \sim \phi_B^{\mathtt{w}}$;

   Else,

   draw a topic $z^{\mathtt{w}}$ from the topic distribution θ_d,
   draw a word w from the topic $z^{\mathtt{w}}$: $w \sim \phi_{z^{\mathtt{w}}}^{\mathtt{w}}$.

To generate a timestamp t_d for news article d,

1. Draw a switch variable $s^{\mathtt{t}} \sim Bernoulli(\lambda_B)$.
3. If $s^{\mathtt{t}} = 1$,

   draw a timestamp t_d from the background time distribution B: $t_d \sim \mathcal{N}(\mu_B, \sigma_B)$;

   Else,

   draw a topic $z^{\mathtt{t}}$ from the topic distribution θ_d,
   draw a timestamp t_d from the topic $z^{\mathtt{t}}$: $t_d \sim \mathcal{N}(\mu_{z^{\mathtt{t}}}, \sigma_{z^{\mathtt{t}}})$.

For $\mathtt{e}$ in $\{\mathtt{p}, \mathtt{l}, \mathtt{o}\}$,
To generate each entity e in news article d,

1. Draw a switch variable $s^{\mathtt{e}} \sim Bernoulli(\lambda_B)$.
2. If $s^{\mathtt{e}} = 1$,

   draw an entity e from the background topic B: $e \sim \phi_B^{\mathtt{e}}$;

   Else,

   draw a topic $z^{\mathtt{e}}$ from the topic distribution θ_d,
   draw an entity e from the topic $z^{\mathtt{e}}$: $e \sim \phi_{z^{\mathtt{e}}}^{\mathtt{e}}$.

As shown in the above process, the posterior distribution of topics depends on the information from five dimensions – text, person, location, organization and time. Despite the fact that entities and time are by themselves interesting dimensions to describe each event/aspect, another important motivation to model them jointly with text is that they impose a regularization effect to the posterior distribution of topics which introduces mutual reinforcement among different dimensions.

B. Inference

We learn the parameters by Maximum Likelihood Estimation (MLE), searching the parameters that maximize the likelihood of the observations

$$\mathcal{L} = P(\mathbf{X}^{\mathtt{w}}, \{\mathbf{X}^{\mathtt{e}}\}, \mathbf{t}|\mathbf{\Theta}) \tag{1}$$

The objective function is thus

$$\begin{aligned}\mathbf{\Theta} =& \arg\max_{\mathbf{\Theta}} \mathcal{L} \\ =& \arg\max_{\mathbf{\Theta}} \alpha^{\mathtt{w}} \sum_{w,d} x_{wd}^{\mathtt{w}} \log \sum_z \phi_{zw}^{\mathtt{w}} \theta_{dz} + \\ & \sum_{\mathtt{e}} \alpha^{\mathtt{e}} \sum_{e,d} x_{ed}^{\mathtt{e}} \log \sum_z \phi_{ze}^{\mathtt{e}} \theta_{dz} + \\ & \sum_d \log \sum_z P(t_d|\mu_z, \sigma_z) \theta_{dz}\end{aligned} \tag{2}$$

To balance the influence from different dimensions, a tunable weight vector $[\alpha^{\mathtt{w}}, \alpha^{\mathtt{p}}, \alpha^{\mathtt{l}}, \alpha^{\mathtt{o}}, 1]$ is used to rescale the likelihoods [5], as is also common in speech recognition when the acoustic and language models are combined. The relative weight of text dimension to others determines the strength of the regularization effects. A natural setting is to allow α's to normalize the likelihoods from all the dimensions to the same scale.

We use the standard Expectation-Maximization (EM) algorithm that iteratively infers the model parameters $\mathbf{\Theta}$. The estimation of the topic distribution θ is given by

$$P(z|d) \propto \alpha^{\mathtt{w}} \sum_w x_{wd}^{\mathtt{w}} P(z|w,d) + \sum_{\mathtt{e}} \alpha^{\mathtt{e}} \sum_e x_{ed}^{\mathtt{e}} P(z|e,d) + P(z|t_d) \tag{3}$$

The first term resembles the estimation of the topic distribution in standard topic modeling, the second term integrates the entity dimensions, and the third term integrates the temporal dimension.

[6]The background topic B is also defined by multiple dimensions with the collection's word distribution $\phi_B^{\mathtt{w}}$, the collection's entity distributions $\{\phi_B^{\mathtt{e}}\}$ and the collection's temporal distribution $\mathcal{N}(\mu_B, \sigma_B)$ where μ_B and σ_B are the mean and standard deviation of the collection's timestamps.

C. Hierarchy Construction

To construct the event-aspect hierarchy, we first apply our generative model to the entire collection $\mathcal{I}_0$ for event discovery. Then we decompose $\mathcal{I}_0$ based on the event descriptors to prepare input $\mathcal{I}_z$ for each event node z. The recursion begins at this point where we apply our generative model to each $\mathcal{I}_z$ for aspect discovery.

The key lies in an effective decomposition from $\mathcal{I}_0$ to $\mathcal{I}_z$. We outline the desired properties of the decomposition as follows.

- The word matrix and the entity matrices in $\mathcal{I}_z$ extract the portion of words/entities belonging to event z.
- The distributions in an event descriptor form the background topic descriptor of its child aspects.

The first property is intuitive. The second property is to ensure that aspects of an event originate from the same theme, while each aspect has its distinct emphasis.

We propose the following decomposition strategy based on the topic (event) membership of each word/entity in a document, which naturally embeds the above requirements.

$$\mathbf{X}_z^{\mathtt{w}}(w,d) = \mathbf{X}^{\mathtt{w}}(w,d) \times P(z|w,d) \tag{4}$$

$$\mathbf{X}_z^{\mathtt{e}}(e,d) = \mathbf{X}^{\mathtt{e}}(e,d) \times P(z|e,d), \mathtt{e} = \mathtt{p},\mathtt{l},\mathtt{o} \tag{5}$$

$P(z|w,d)$ denotes the posterior probability that the w-th word in the d-th document belongs to event z. Each entry $\mathbf{X}^{\mathtt{w}}(w,d)$ of the original word matrix $\mathbf{X}^{\mathtt{w}}$ is thus split to different events based on the posterior probability. The decomposition of entity matrices is done in the same way.

To see why the second property holds, let $(\phi_B^{\mathtt{w}})_z, (\phi_B^{\mathtt{e}})_z$ denote the background word and entity distributions computed with input $\mathcal{I}_z$, and let $\phi_z^{\mathtt{w}}, \phi_z^{\mathtt{e}}$ denote the word and entity distributions of event z estimated from the event discovery step. We have

$$(\phi_B^{\mathtt{w}})_z(w) = \frac{\sum_d \mathbf{X}_z^{\mathtt{w}}}{\sum_{w,d} \mathbf{X}_z^{\mathtt{w}}} = \frac{\sum_d \mathbf{X}^{\mathtt{w}}(w,d)P(z|w,d)}{\sum_{w,d} \mathbf{X}^{\mathtt{w}}(w,d)P(z|w,d)} = \phi_z^{\mathtt{w}}(w) \tag{6}$$

$$(\phi_B^{\mathtt{e}})_z(e) = \frac{\sum_d \mathbf{X}_z^{\mathtt{e}}}{\sum_{e,d} \mathbf{X}_z^{\mathtt{e}}} = \frac{\sum_d \mathbf{X}^{\mathtt{e}}(e,d)P(z|e,d)}{\sum_{w,d} \mathbf{X}^{\mathtt{e}}(e,d)P(z|e,d)} = \phi_z^{\mathtt{e}}(e) \tag{7}$$

The first equal sign in both equations follows by definition of the background topic. The second equal sign demonstrates our decomposition strategy. And the third equal sign follows from the updating formula in M-step of the inference procedure.

V. Tweets Linking

In the previous section, we have learned a word distribution, three entity distributions and a time distribution for each event and each aspect in the hierarchy. These distributions form comprehensive descriptors, which are used to find in Twitter the "reflection" of each aspect of a news event[7].

In this section, we first describe the candidate pool retrieval for each event with the event descriptor, and then elaborate the bootstrapping procedure which selects tweets for each aspect with the aspect descriptors.

[7]A substantial number of tweets contain a URL to a news article and the contents are just the news titles. Identifying these tweets are trivial in the linking task and do not add much value for users. In this paper, we skip these cases and consider the tweets without URLs only.

A. Candidate Pool Retrieval with Event Descriptor

A candidate pool of tweets are retrieved for each event by information retrieval (IR). Specifically, we propose a language model which simultaneously investigate text, entities and time to determine the relevance of a tweet to an event.

The event descriptor is fed in as a query. Documents (tweets) are ranked by the probability of being generated by the query. This IR step is motivated by the fact that high volumes of tweets make it impossible to investigate every single tweet. The event descriptor provides a feasible way to retrieve a highly relevant candidate pool for identifying the reflections. The score for ranking is derived as follows:

$$\begin{aligned}&\log P(d|z) \quad (d \text{ is a tweet and } z \text{ is an event descriptor})\\ =&\alpha^{\mathtt{w}} \log P(d^{\mathtt{w}}|z) + \sum_{\mathtt{e}=\mathtt{p},\mathtt{l},\mathtt{o}} \alpha^{\mathtt{e}} \log P(d^{\mathtt{e}}|z) + \log P(d^{\mathtt{t}}|z)\end{aligned} \tag{8}$$

where $d^{\mathtt{w}}$ denotes all the words in d, $d^{\mathtt{e}}$ denotes all the type $\mathtt{e}$ entities in d, and $d^{\mathtt{t}}$ is the timestamp of d. The likelihoods from different dimensions are rescaled with α's by the same philosophy as in Section IV-B. Apply Bayes's rule to the first two terms as in standard query likelihood model, we obtain

$$\begin{aligned}&\log P(d|z)\\ \propto&\alpha^{\mathtt{w}} \log P(z|d^{\mathtt{w}}) + \sum_{\mathtt{e}=\mathtt{p},\mathtt{l},\mathtt{o}} \alpha^{\mathtt{e}} \log P(z|d^{\mathtt{e}}) + \log P(d^{\mathtt{t}}|z)\\ =&\alpha^{\mathtt{w}} \sum_w \phi_{zw}^{\mathtt{w}} \log P(w|d) + \sum_{\mathtt{e}=\mathtt{p},\mathtt{l},\mathtt{o}} \alpha^{\mathtt{e}} \sum_e \phi_{ze}^{\mathtt{e}} \log P(e|d)\\ &+ \log P(d^{\mathtt{t}}|z)\end{aligned} \tag{9}$$

This is the final score used for ranking, where $P(d^{\mathtt{t}}|z) \sim t|N(\mu_z, \sigma_z)$, $P(w|d)$ and $P(e|d)$ are obtained by a Dirichlet smoothing to the language model of a tweet d:

$$P(w|d) = \frac{\#(w,d) + \mu P(w)}{\#w + \mu} \tag{10}$$

$$P(e|d) = \frac{\#(e,d) + \mu P(e)}{\#e + \mu} \quad e = p,l,o \tag{11}$$

B. Dataless Bootstrapping

We select and rank tweets for each aspect by a bootstrapped multi-class dataless classification scheme. We classify the tweets in the candidate pool into different aspects and select the top ones for each aspect. In addition to the multidimensional relevance requirement, this step is motivated by a) the existence of vocabulary gap between news and tweets; and b) the existence of noisy tweets which are irrelevant to any aspect.

Bootstrapping provides a way to weigh the semantic representation extracted from news that best fits the specific tweet collection. It starts with a confident seed set for each aspect obtained using the aspect descriptor. These are viewed as the first batch of labeled data. In each iteration, a multi-class classifier is trained using the current labeled data. And then the classifier labels more data by selecting the most confident tweets from the unlabeled ones. After each iteration, the accuracy of the classifier is improved and more labeled data are incorporated. The procedure is summarized as follows:

Step 1: Initialize M most confident seed tweets for each aspect using the aspect descriptors. The confidence is measured by the score from the language model as in Eq. (9).

Step 2: For each iteration, train a classifier based on the current set of labeled data and label N more tweets for each aspect.

Step 3: Repeat Step 2 until a desired number of tweets for each aspect are retrieved or the confidence score is lower than a threshold.

The classifier can be any multi-class classifier taking arbitrary features. In this study, we use logistic regression with L2 regularization. The features we use are listed as follows.

- Tf-idf word features. The values are scaled to range $[0, 1]$.
- Tf-idf entity features. The values are scaled to range $[0, 1]$.
- Time vector. For a tweet with a timestamp t, the i-th element in the time vector is the probability density at t computed using the time distribution of the i-th aspect. The vector is normalized to sum to 1.

VI. Experiments

We perform empirical study to answer the following questions: 1) how effective is the event-aspect hierarchy learning? and 2) how well is the tweets linking quality? At the end, we demonstrate that our framework naturally supports a user friendly presentation with entity graphs, time spans, news summaries and tweet highlights.

A. Dataset Description

We consider two datasets in our experiments.

TopStory We crawled the top stories (full text) from Google News[8] every 30 minutes from Dec 20, 2014 to Jan 4, 2015. For each news, we query the Twitter Search API[9] with the extracted noun phrases from the title and snippet. Tweets containing at least one of the noun phrases are returned. We collected tweets that are posted within one day after the published time of the news. The dataset consists of $3,747$ news and $36,149,019$ tweets in total.

Link This dataset is provided by Guo *et al.* [1], which contains explicit URL links from each tweet to a related news article. They crawled $12,704$ CNN and NYTIMES news (title + snippets) from RSS feeds from Jan 11 to Jan 27, 2013. $34,888$ tweets that contain a single link to a CNN or NYTIMES news were collected during the same period. This dataset is a gold standard dataset to test the performance of the tweets linking module.

For both datasets, entities including persons, locations, and organizations are extracted using DBpedia Spotlight[10].

B. Implementation Details

For all the methods in our experiments, we set the number of iterations to be 20. The topic modeling parameters are initialized by the results from Kmeans clustering with 50 random initializations. Specifically, Kmeans is run on the tf-idf vectors of news articles. Topic distributions are initialized by the cluster assignments[11]. The word/entity/time distributions are initialized by the aggregate statistics of the documents in each cluster. The weights α's are tuned for each dataset on a develop set containing $1/10$ of the dataset. Specifically, we first let α_0's to scale the likelihoods from different dimensions after the first iteration to the same value. Then we search in a grid centered at α_0's and select the configuration which leads to the highest pointwise mutual information (PMI) [6]. In the tweets linking procedure, we set $M = 50$ and $N = 10$.

C. Effectiveness and efficiency of the Event-Aspect Hierarchy Learning

We investigate the benefit from integrating multiple dimensions (entities and time) and compare with the state-of-art topical hierarchy construction method CATHYHIN [7]. Pointwise mutual information (PMI) [6] is used to measure the clustering quality, which is generally preferred over other quantitative metrics such as perplexity or the likelihood of held-out data [8]. We compare the average PMI over all events[12]. An efficiency analysis is presented at the end. Methods for comparison are summarized as follows.

- Our model with text dimension only;
- Our model with text + entity dimensions;
- Our full model with text + entity + time;
- CATHYHIN [7]. CATHYHIN takes a collection of documents and entities from a network perspective. They take the same input as our model and build the hierarchy recursively as well. But they work on networks formed by multiple combination of the matrix multiplications and conduct network clustering for topical hierarchy construction. For example, $\mathbf{X}^{\mathrm{w}} \times \mathbf{X}^{\mathrm{p}^T}$ forms a word-by-person network. CATHYHIN requires human to specify several types of networks and models the edge weight generation using a Poisson distribution. By default, all the entity type combinations are considered in the clustering process.

We list the results with two different number of events settings, *i.e.*, 30 events and 150 events. Similar results were observed for other numbers of events. As shown in Table II, integrating entities and time increases the topical coherence. CATHYHIN has comparable performance with our Text+Entities model on the Link dataset but is significantly worse on the TopStory dataset. In fact, the Link dataset only contains titles and snippets which are of high quality. This makes the clustering task relatively easy. As CATHYHIN primarily relies on the co-occurrence matrices of all possible entity type combinations, it performs better on a smaller and cleaner dataset. Another significant observation is that

[8]https://news.google.com/

[9]https://dev.twitter.com/rest/public/search

[10]https://github.com/dbpedia-spotlight/dbpedia-spotlight

[11]For *e.g.*, if there are K clusters and document d is assigned to cluster 2, the topic distribution becomes $(s, 1-(K-1)s, ..., s)$. $s = 1/K^2$ is a smoothing parameter.

[12]The comparison is done for the event level because all the methods start with the same root node but the event clusters can be different which makes aspect level PMI incomparable.

TABLE II
AVERAGED PMI OVER EVENTS USING TOP 20 WORDS FROM THE WORD DISTRIBUTIONS

	TopStory	Link
30 events		
CATHYHIN	0.5239	0.2769
Text	0.702	0.2503
Text+Entities	0.7423	0.2803
Full	**0.773**	**0.2866**
150 events		
CATHYHIN	0.316	0.3123
Text	0.4065	0.2883
Text+Entities	0.4281	0.3151
Full	**0.4485**	**0.3222**

our method is far more efficient than CATHYHIN since we work on the sparse document by words/entities matrices while CATHYHIN works on the co-occurrence matrices which is usually much denser especially for long text. Although we take the same amount of input knowledge, the running time of our method is in the order of several minutes but CATHYHIN takes several hours[13]. The running time of our method with varying event number is plotted in Figure 3. The results show that our model scales linearly with the event number. In fact, the complexity for each iteration of the inference process is dominated by the text dimension in the M-step, which is $O(K|\mathbf{X}^{\mathtt{w}}|)$, where K is the number of events and $|\mathbf{X}^{\mathtt{w}}|$ is the number of non-zero entries in the matrix. Thus our model scales linearly with the number of events and the size of the collection.

D. Tweets Linking

To quantitatively evaluate the linking procedure, we use the Link dataset which has explicit links between news and tweets. We compare with the WTMF-G method proposed in [1], which learns a latent representation (vector) for each news and tweet also considering multi-dimensional information such as text, entities and time. They use cosine similarity of the latent vectors to measure the relevance of a news and a tweet. The number of events is set to 150 because WTMF-G was reported to work best at this setting. We design the following experiment to study the precision and recall. Each news article d is assigned to the event z^* by $z^* = \arg\max_z \theta_{dz}$. We take the top 20 events measured by the total number of news articles contained. For each of these events, our method select the top $k \times \#$(articles in the event) tweets. To compare with WTMF-G, we take the news assignments as given and consider two baselines derived from WTMF-G: 1) retrieve the top k tweets for each news article to form a same length of ranking list; 2) use the centroid of the latent vectors of the news in an event to retrieve $k \times \#$(articles in the event) tweets. We compute the average precision and recall for the top 20 events and randomly select one of them to evaluate the average precision and recall of its aspects.

[13]Both test are on a 16GB memory Matlab platform. For CATHYHIN, we used the implementation from the authors. CATHYHIN finishes in 3-4 hours for TopStory dataset and 10-20 minutes for Link dataset.

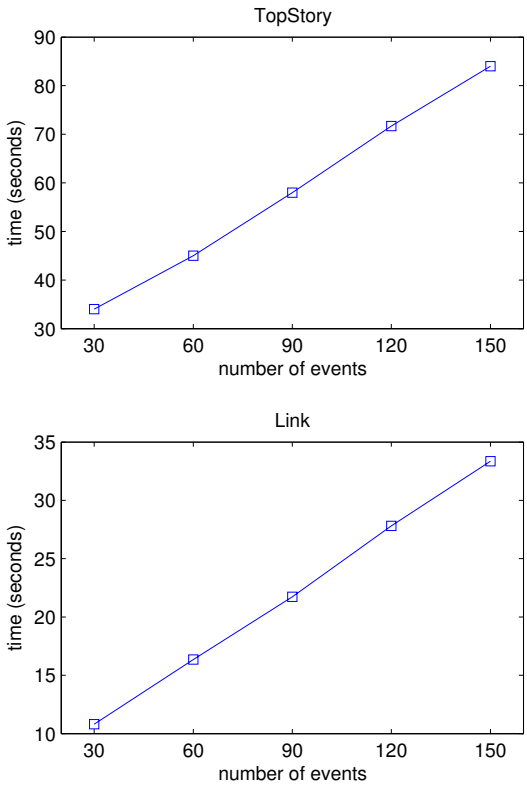

Fig. 3. Running Time with Varying Event Number

The precisions/recalls are computed at the positions $1, 5, 10, 20$ and are plotted in Figure 4. Our method clearly outperforms both baselines. This demonstrates the effectiveness of our event/aspect descriptors and the bootstrapped dataless classification procedure.

E. Presenting Results for User Digestion

It is always important yet challenging to present learned results to users in an informative way. Our framework naturally supports a user friendly presentation with entity graphs, time spans, news summaries and tweet highlights. We use the `Sony Hack` event in the TopStory dataset to illustrate each component. The overall visualization of this event is given at Event_description_overall.html[14] where each aspect is given at Event_description_aspect1.html[15] (change 1 from 2 to 6 to see other aspects).

For each aspect of an event, we offer a view with an entity graph, a time span, a ranked list of news articles(the headlines are displayed) and a ranked list of tweets. We also offer an event view which integrates all the information of its aspects.

[14]The text should be clickable. If not, go to https://dl.dropboxusercontent.com/u/155956218/Event_description_overall.html Make sure there is no space after ".com" and the underscores are correctly typed.

[15]https://dl.dropboxusercontent.com/u/155956218/Event_description_aspect1.html

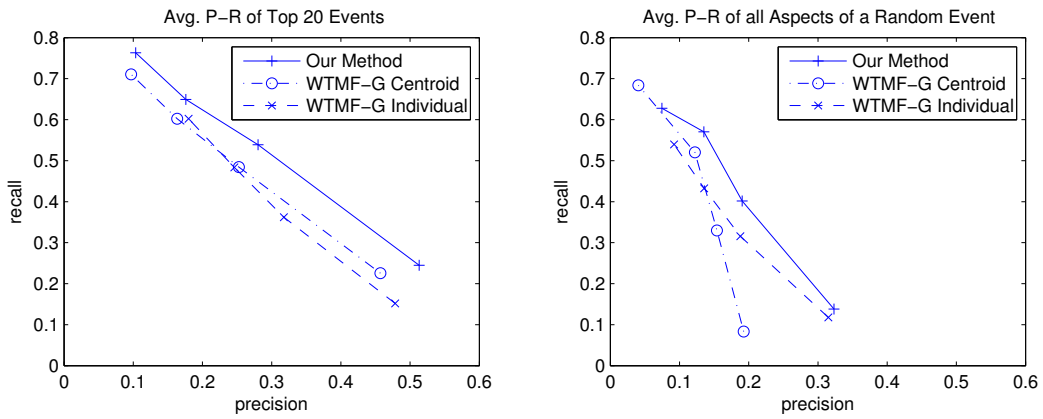

Fig. 4. Precision-Recall Curves. The four points on each curve correspond to the precision/recall @ $1, 5, 10, 20$.

In the following paragraphs, we explain how each component is generated. We use the `Sony Hack` event with a sample aspect about the "North Korea Internet Outage" as a running example.

An event $z = \{\phi^{\mathtt{w}}_z, \{\phi^{\mathtt{e}}_z\}, \mu_z, \sigma_z\}$ is associated with $\mathcal{I}_z$, which is used as input to discover aspects. Let $za = \{\phi^{\mathtt{w}}_{za}, \{\phi^{\mathtt{e}}_{za}\}, \mu_{za}, \sigma_{za}\}$ be the descriptor of the a-th aspect in event z, and let $\mathcal{I}_{za} = \{\mathbf{X}^{\mathtt{w}}_{za}, \{\mathbf{X}^{\mathtt{e}}_{za}\}, \mathbf{t}\}$ associate with node za.

1) Entity Graphs: The recursive hierarchy construction leads to a natural visualization of the entity graph. For an aspect a in event z. The edge weight matrix $\mathbf{W}_{za}$ is given by

$$\mathbf{X}all_{za} = \text{vertical stack of}(\mathbf{X}^{\mathtt{p}}_{za}, \mathbf{X}^{\mathtt{l}}_{za}, \mathbf{X}^{\mathtt{o}}_{za}) \quad (12)$$

$$\mathbf{W}_{za} = \mathbf{X}all_{za}\mathbf{X}all^T_{za} \quad (13)$$

and the node weight is given by $\{\phi^{\mathtt{e}}_{za}\}$. For an event, an entity graph is constructed by combining all of its aspect entity graphs to form a multigraph, *i.e.*, two entities can be connected by multiple edges denoting their interaction in multiple aspects. The edge weights are the same as in individual aspect graphs while the node weights are given by $\{\phi^{\mathtt{e}}_z\}$. We give each aspect a unique color and let the node size (edge width) be proportional to the corresponding weight of a node (an edge).

The entity graph of the `Sony Hack` event is shown in Figure 5. Each node denotes an entity where the entities of the same type are in the same color. Each edge denotes the correlation between two entities where different colors represent the correlations in different aspects. We can see that *"Sony", "North Korea", "Kim Jong-un","Barack Obama", "Seth Rogen" and "James Franco"* are most influential in this event. If we zoom into the view of the red aspect, as shown in Figure 6, we can examine the entities in this particular aspect.

2) Time Spans: We use the Gaussian parameters μ_{za}, σ_{za} to generate the time distribution of each aspect. The time spans of different aspects in this event are shown in Figure 7, where the colors are consistent with the edges in the entity graph.

3) News Summaries and Tweet Highlights: While sophisticated news summarization can be performed to extract news summaries and tweet highlights, in this visualization we adopt a simple strategy. For the aspect a in z, we rank news articles by their posterior weight on a $P(a|d) = \theta_{da}$. We list the top five news articles in Figure 8. Tweets are ranked by the output score of the classifier and we list the top five tweets together with the news summaries. The top five keywords from the word distribution are also listed. Obviously, the summaries, highlights together with the entity graph and the time span are of great help in understanding this aspect.

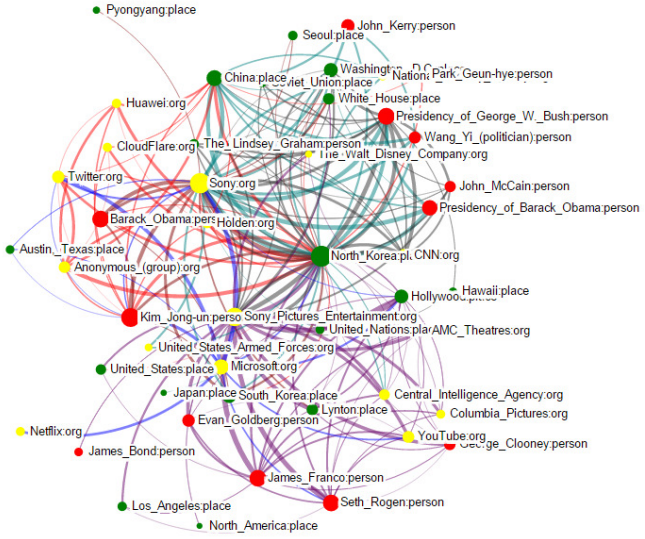

Fig. 5. Entity Graph in the Event View. Red node: person; Green node: place; Yellow node: organization. The size of the node denotes the influence of the entity in this event. The width of the edge denotes the strength of the correlation between two entities. Different colors of edges represent the correlations in different aspects. We can see the influential entities in this event are: *"Sony", "North Korea", "Kim Jong-un","Barack Obama", "Seth Rogen (director and actor of the film)" and "James Franco (actor of the film)"* .

Aspect 1: News Summaries	Aspect 1: Keywords	Aspect 1: Tweet Hightlights
North Korean Web goes dark days after Obama pledges response to Sony hack	internet	I unplugged North Koreas internet #HappyHolidays
North Korean Internet Goes Dark; A US Government Attack 'Would Be Way Worse'	korea	Internet down for everyone in North Korea. Aka, 3 households arw without internet in North Korea #NorthKorea #Sony #SonyHack
North Korean Internet Goes Dark in Wake of Sony Hack	north	HAHAHA SCREW YOU NORTH KOREA NO INTERNET 4 U
N. Korea Internet Service Restored	outage	The most surprising news about North Korea's internet problems is that North Korea had access to the internet in the first place.
North Korea's Internet Service Appears Erratic After Outage	dyn	Internet in North Korea taken down by the millions of people trying to see if Internet in North Korea works. #DoSbyTesting

Fig. 8. The News Summary, Keywords and the Tweet Highlight of the aspect "North Korea Internet Outage" in the `Sony Hack` Event

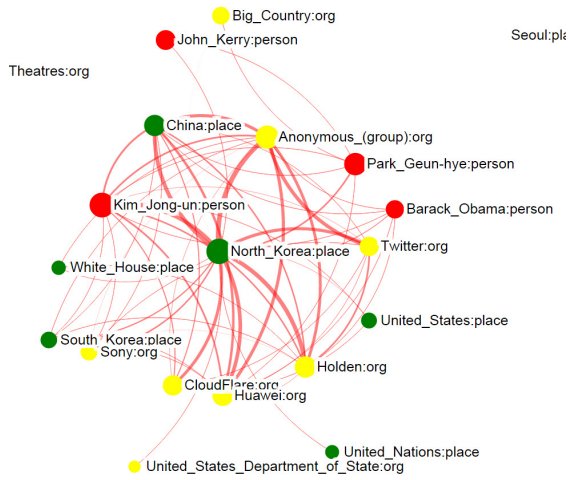

Fig. 6. Zoom in to the Entity Graph of the Red Aspect about "North Korea Internet Outage". The size of a node denotes the influence of the entity in the aspect. The width of an edge denotes the strength of the correlation between two entities.

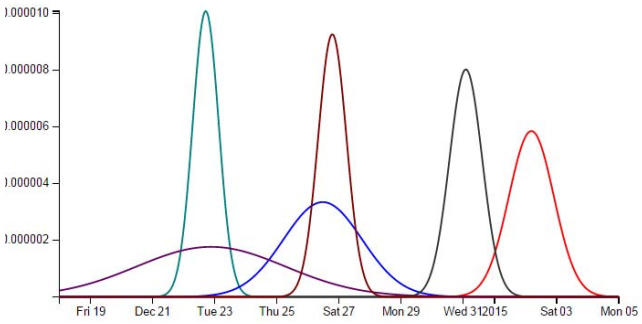

Fig. 7. Time Spans in the Event View. The colors are consistent with the edges in the entity graph.

VII. Related Work

To the best of our knowledge, this is the first work addressing the task of event based multi-aspect linking between news and tweets. Yet our work is related to topic modeling, event detection and several joint studies of news media and social media. In this section, aside from the related work we mentioned previously, we review the recent literature and make connections with them.

A. Topic modeling

There has been a substantial amount of research on topic modeling. Inspired by entity topic models [9], [10], dynamic topic models [5], [11] and hierarchical topic models [12], we tailor our model to integrate multi-dimensional information for event/aspect learning. We follow the universal document-topic-word philosophy. In the mean time, we integrate entities and temporal information to jointly describe an event/aspect as well as to regularize the topic distributions. The proposed decomposition strategy provides a natural way for efficient hierarchy construction. Our model also provides an effective presentation for both user digestion and the tweets linking task afterwards. Provided the evidence by Masada *et al.* [13] that no meaningful difference between LDA and pLSI are observed for dimensionality reduction in document clustering, we intentionally leave out the prior for document-topic distributions as in LDA but take a pLSI style for an efficient EM optimization procedure, which is critical in hierarchical inference once the document collection becomes large. It is worth noting that our topic modeling algorithm scales linearly with the number of events and the length of the corpus.

B. Event Detection in Tweets

In the literature, there have been numerous research efforts aimed at event discovery in tweets [14]–[18], where various clustering methods taking well-calibrated features have been proposed. These studies focused on the single collection of tweets where huge number of random posts irrelevant to any news events interfere as noise. Our task distinguishes itself from this line of work by taking an opposite perspective. We discover events by investigating news articles, carefully learning different aspects and identifying their reflections in tweets, which is a more targeted and fine-grained task.

C. Joint Study of News Media and Microblogging Sites

Joint studies of news media and microblogging sites have attracted much attention recently due to a broad spectrum of potential applications. Zhao *et al.* [19] conducted a comparative study on the high level categories (politics,

sports, etc.) and types (event-oriented, long-standing, etc.) of topics discovered from News and Tweets by running separate topic models in the two sources. Subavsic and Berendt [20] performed a case study to investigate text/headline/sentiment/entity divergence between news and tweets in an aggregate sense, concluding that a major role of Twitter authors consists of neither creating nor peddling, but extending them by commenting on news, which justifies the significance of our work. Gao *et al.* [21] studied the sentence level complementary news sentences and tweets and Wei and Gao [22] studied news highlights extraction utilizing tweets for a given event which can benefit from our event detection and representation. Within an event, Gao *et al.* [21] modeled dimensions such as location and time as latent aspects which were also characterized by word distributions, while they disregarded topical aspects. In our work we explicitly extract the entities from these dimensions, model them directly and go beyond events to find fine-grained topical aspects. Kothari *et al.* [23] and Masada *et al.* [24] utilized various features to classify tweets into comments or non-comments. These features can be well integrated to our classifier for tweets linking as well.

VIII. Conclusions and Future Work

In this paper, we proposed a unified framework to mine multi-aspect reflections of news events in Twitter. We proposed an effective time and entity-aware event/aspect discovery model to learn accurate descriptors of news events and their multiple aspects; the aspects of an event are linked to their reflections in Twitter by a bootstrapped dataless classification scheme, which elegantly handles the challenges of selecting informative tweets under overwhelming noise and bridging the vocabulary gap between news and tweets. Experimental results demonstrated that our framework can effectively retrieve the relevant tweets for fine-grained aspects of news events. While the scope of this paper is to accurately identify the "reflections" of news events in twitter, discovering new aspects in Twitter which are not emphasized in news is an interesting future direction. We also demonstrated that our framework naturally generates an informative presentation of each event with entity graphs, time spans, news summaries and tweet highlights to facilitate user digestion. The capability of creating a high-quality aspect-specific and time-aware event dataset is of considerable practical benefits for various interesting applications such as comparative opinion mining and multi-corpus text summarization, which can lead to a broad spectrum of future research.

Acknowledgment

Research was sponsored in part by the U.S. Army Research Lab. under Cooperative Agreement No. W911NF-09-2-0053 (NSCTA), National Science Foundation IIS-1017362, IIS-1320617, and IIS-1354329, HDTRA1-10-1-0120, and grant 1U54GM114838 awarded by NIGMS through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative (www.bd2k.nih.gov), and MIAS, a DHS-IDS Center for Multimodal Information Access and Synthesis at UIUC.

References

[1] W. Guo, H. Li, H. Ji, and M. T. Diab, "Linking tweets to news: A framework to enrich short text data in social media," in *ACL*, 2013, pp. 239–249.

[2] M. Tsagkias, M. de Rijke, and W. Weerkamp, "Linking online news and social media," in *WSDM*. New York, NY, USA: ACM, 2011, pp. 565–574.

[3] T.-A. Hoang-Vu, A. Bessa, L. Barbosa, and J. Freire, "Bridging vocabularies to link tweets and news," *International Workshop on the Web and Databases, WebDB*, 2014.

[4] Y. Song and D. Roth, "On dataless hierarchical text classification," in *AAAI*, 7 2014.

[5] X. Wang and A. McCallum, "Topics over time: A non-markov continuous-time model of topical trends," in *KDD*. New York, NY, USA: ACM, 2006, pp. 424–433.

[6] D. Newman, E. V. Bonilla, and W. L. Buntine, "Improving topic coherence with regularized topic models," in *NIPS*, 2011, pp. 496–504.

[7] C. Wang, M. Danilevsky, J. Liu, N. Desai, H. Ji, and J. Han, "Constructing topical hierarchies in heterogeneous information networks," in *ICDM*, 2013, pp. 767–776.

[8] J. Tang, M. Zhang, and Q. Mei, "One theme in all views: Modeling consensus topics in multiple contexts," in *KDD*. New York, NY, USA: ACM, 2013, pp. 5–13.

[9] D. Newman, C. Chemudugunta, and P. Smyth, "Statistical entity-topic models," in *KDD*. New York, NY, USA: ACM, 2006, pp. 680–686.

[10] H. Kim, Y. Sun, J. Hockenmaier, and J. Han, "Etm: Entity topic models for mining documents associated with entities," in *ICDM*. Washington, DC, USA: IEEE Computer Society, 2012, pp. 349–358.

[11] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *ICML*. New York, NY, USA: ACM, 2006, pp. 113–120.

[12] D. M. Blei, T. L. Griffiths, and M. I. Jordan, "The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies," *J. ACM*, vol. 57, no. 2, pp. 7:1–7:30, Feb. 2010.

[13] T. Masada, S. Kiyasu, and S. Miyahara, "Comparing lda with plsi as a dimensionality reduction method in document clustering," in *Large-Scale Knowledge Resources. Construction and Application*. Springer, 2008, pp. 13–26.

[14] L. Shou, Z. Wang, K. Chen, and G. Chen, "Sumblr: Continuous summarization of evolving tweet streams," in *SIGIR*. New York, NY, USA: ACM, 2013, pp. 533–542.

[15] J. Vosecky, D. Jiang, K. W.-T. Leung, and W. Ng, "Dynamic multi-faceted topic discovery in twitter," in *CIKM*. New York, NY, USA: ACM, 2013, pp. 879–884.

[16] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling, "Twitterstand: News in tweets," in *GIS*. New York, NY, USA: ACM, 2009, pp. 42–51.

[17] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes twitter users: Real-time event detection by social sensors," in *WWW*. New York, NY, USA: ACM, 2010, pp. 851–860.

[18] A. Angel, N. Sarkas, N. Koudas, and D. Srivastava, "Dense subgraph maintenance under streaming edge weight updates for real-time story identification," *VLDB*, vol. 5, no. 6, pp. 574–585, Feb. 2012.

[19] W. X. Zhao, J. Jiang, J. Weng, J. He, E.-P. Lim, H. Yan, and X. Li, "Comparing twitter and traditional media using topic models," in *ECIR*. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 338–349.

[20] I. Subavsic and B. Berendt, "Peddling or creating? investigating the role of twitter in news reporting," in *ECIR*. Berlin, Heidelberg: Springer-Verlag, 2011, pp. 207–213.

[21] W. Gao, P. Li, and K. Darwish, "Joint topic modeling for event summarization across news and social media streams," in *CIKM*. New York, NY, USA: ACM, 2012, pp. 1173–1182.

[22] Z. Wei and W. Gao, "Utilizing microblogs for automatic news highlights extraction," in *COLING*, 2014, pp. 872–883.

[23] A. Kothari, W. Magdy, K. Darwish, A. Mourad, and A. Taei, "Detecting comments on news articles in microblogs," in *ICWSM*, 2013.

[24] T. Štajner, B. Thomee, A.-M. Popescu, M. Pennacchiotti, and A. Jaimes, "Automatic selection of social media responses to news," in *KDD*. New York, NY, USA: ACM, 2013, pp. 50–58.