Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence

ET-LDA: Joint Topic Modeling for Aligning Events and Their Twitter Feedback

Yuheng Hu[†] **Ajita John**[‡] **Fei Wang**[§] **Subbarao Kambhampati**[†]
[†] Department of Computer Science, Arizona State University, Tempe, AZ 85281
[‡] Collaborative Applications Research, Avaya Labs, Basking Ridge, NJ 07920
[§] IBM T. J. Watson Research Lab, Hawthorne, NY 10532
[†]{yuhenghu, rao}@asu.edu [‡]ajita@avaya.com [§]feiwang03@gmail.com

Abstract

During broadcast events such as the Superbowl, the U.S. Presidential and Primary debates, etc., Twitter has become the *de facto* platform for crowds to share perspectives and commentaries about them. Given an event and an associated large-scale collection of tweets, there are two fundamental research problems that have been receiving increasing attention in recent years. One is to extract the topics covered by the event and the tweets; the other is to segment the event. So far these problems have been viewed separately and studied in isolation. In this work, we argue that these problems are in fact inter-dependent and should be addressed together. We develop a joint Bayesian model that performs topic modeling and event segmentation in one unified framework. We evaluate the proposed model both quantitatively and qualitatively on two large-scale tweet datasets associated with two events from different domains to show that it improves significantly over baseline models.

1 Introduction

During public broadcast events such as the Superbowl, the U.S. Presidential and Primary debates, the last episode of a TV drama series, etc., Twitter has become the *de facto* platform for crowds to share perspectives and commentaries about these events. Given an event and an associated large-scale collection of tweets, we face two fundamental problems in analyzing and understanding them, namely, extracting the topics covered in the event and tweets, and segmenting the event into topically coherent segments. Tackling the two problems is critical to applications like computational advertising, community detection, journalistic investigation, storytelling, playback of events, etc. While both topical modeling and event segmentation have received considerable attention in recent years, they have been mainly viewed as separate problems and studied in isolation. For example, there have been significant efforts on developing Bayesian models to discover the patterns that reflect the underlying topics from the document (Blei, Ng, and Jordan 2003; Griffiths et al. 2004; Wang and McCallum 2006; Titov and McDonald 2008). Similarly, there is also a rich body of work devoted to segmentation of events/discourses/meetings via heuristics, machine learning, etc. (Hearst 1993; Boykin and Merlino 2000; Galley et al. 2003; Dielmann and Renals 2004).

Directly applying these current solutions to analyze the event and its associated tweets however has a major drawback: they treat event and tweets independently, thus ignoring the topical influences of the event on its associated tweets. In reality they are obviously inter-dependent. For example, in practice, when tweets are generated by the crowds to express their interests in the event, their content is essentially influenced by the topics covered in the event in some way. Based on such dependencies, i.e., topical influences, a person can respond to the event in a variety of ways. For example, she may choose to comment directly on a specific topic in the event which is of concern and/or interest to her. So, her tweets would be deeply influenced by that specific topic. In another situation, she could also comment broadly about the event. Therefore, the tweets would be less influenced by the specific topics but more by the general topics of the event.

In this paper, we are interested in jointly modeling the topics of the event and its associated tweets, as well as segmenting the event in one unified model. Our work is motivated by the observation that the topical influences from the event on its associated tweets are not only used for indicating the topics mentioned in the event but also indicating the content/topics in tweets and the tweeting behaviors of the crowd. Besides, by accounting for such influences on tweets, we can obtain a richer context about the evolution of topics and the topical boundaries in the event which is critical to the event segmentation, as mentioned in (Shamma, Kennedy, and Churchill 2009).

We build our joint model based on Latent Dirchlet Allocation (LDA), a Bayesian model proven to be effective for topic modeling. In our model, an event may consist of many paragraphs, each of which discusses a particular set of topics. These topics evolve over the timeline of the event. We assume that whether the topic mixture of a paragraph changes from the one in its preceding paragraph follows a binomial distribution parameterized by the similarity between their topic distributions. With some probability, the two paragraphs are merged to form a segment; otherwise, a new segment is created. Additionally, we assume the event (in fact the segments) can impose topical influences on the associated tweets. Under such influences, the words in the

tweets can belong to two distinct types of topics: general topics, which are high-level and constant across the entire event, and specific topics, which are detailed and relate to specific segments of the event. We define a tweet in which most words belong to general topics as a "general tweet", indicating a weak topical influence from the event, whereas a tweet with more words about the specific topics is defined as a "specific tweet", indicating a strong topical influence from one segment of the event. Similar to the event segmentation, whether the event has strong or weak influence on tweets depends on a binomial distribution. To learn our model, we derive inference and estimate parameters using Gibbs sampling. In the update equations, we can observe how the tweets help regularize the topic modeling process via topical influences and *vice versa*. To test our model, we apply it to two large-scale tweet datasets associated with two events from different domains (a) President Obama's Middle East speech on May 19, 2011 and (b) the Republican Primary debate on September 9, 2011. We examine the results both quantitatively and qualitatively to demonstrate that our model improves significantly over baseline models.

2 Related Work

Topic modeling methods, such as Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) have achieved great success in discovering underlying topics from text documents. Recently, there has been increasing interest in developing better and sophisticated topic modeling schemes. One line of such research is to extend topic models on networked documents, e.g., research publications, blogs etc. PHITS (Hofmann 2001) models the documents and their interconnectivity based on topic-specific distributions. Further extensions include (Dietz, Bickel, and Scheffer 2007), Link-PLSA-LDA (Nallapati et al. 2008) and RTM (Chang and Blei 2009). In addition, some works consider the dynamics of topics which include dynamic topic model (Blei and Lafferty 2006). Also, recent efforts apply topic modeling on social media such as (Ramage, Dumais, and Liebling 2010; Hu and Liu 2012).

In parallel, there is a rich body of work on automatic topic segmentation of events/texts/meetings. Many approaches have been developed. For example, (Hearst 1993) uses a measure of lexical cohesion between adjoining paragraphs for segmenting texts. LCSeg (Galley et al. 2003) uses a similar approach on both text and meeting transcripts and gains better performance than that achieved by applying text/monologue-based techniques. In addition to lexical approaches, machine learning methods have also been considered. (Beeferman, Berger, and Lafferty 1999) combines a variety of features such as statistical language modeling, cue phrases, discourse information to segment broadcast news. Recent advances have used generative models such as (Purver et al. 2006).

The focus of most of the above work is either to model topics in documents (where documents are assumed to be the *homogenous*, e.g, research papers) or segment the events alone. However, they do not provide insights into how to characterize one source of text (tweets) in response to another (event). A distinct difference in our work is that, the event and the associated tweets are *heterogenous*: the topics in a tweet may be sampled from different types of topic mixtures (general or specific). Additionally, the topic mixtures in an event evolve over its timeline. While the current paper focuses on the technical development and evaluation of the ET-LDA framework, our companion papers (Hu, John, and Seligmann 2011; Hu et al. 2012) elaborate on the motivations for joint analysis and alignment of events and tweets.

Table 1: Mathematical Notation

Notation	Description
S	a set of paragraphs in the event's transcript
N_s	the number of words in paragraph s
T	a set of tweets associated with the event
M_t	the number of words in tweet t
$\theta^{(s)}$	topic mixture of the specific topics from a paragraph s of the event
$\psi^{(t)}$	topic mixture of the general topics from tweets corpus
$\delta^{(s)}$	parameter for choosing to draw topics in paragraph s from $\theta^{(s)}$ or $\theta^{(s-1)}$
$c^{(s)}$	indicates whether the topic of a paragraph is drawn from current or previous segment's topics.
$\lambda^{(t)}$	parameter for choosing to draw topics in t from θ or ψ
$c^{(t)}$	indicates whether the topic of a tweet is drawn from specific or general topics
$s^{(t)}$	a referred segment, to which a specific topic in a tweet is associated
w_s, w_t	words in event's transcript, tweets, respectively
z_s, z_t	topic assignments of words in event, tweets, respectively.
α, β	Dirichlet/beta parameters of the Multinomial/Bernoulli distributions

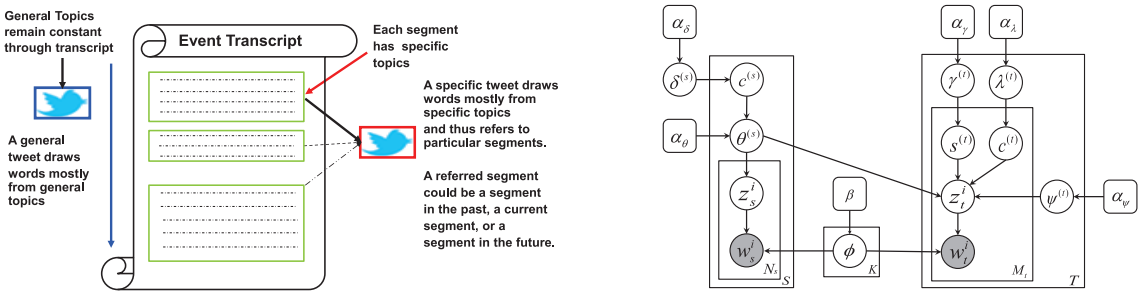
3 Joint modeling of Event and Twitter Feeds

In this section, we show how to represent the event and its Twitter feeds by a hierarchical Bayesian model based on Latent Dirichlet Allocation (LDA), so that the topic modeling of the event/tweets and event segmentation can be achieved. Table 1 lists the notation used in this paper.

3.1 Model

Our proposed model called the joint Event and Tweets LDA (ET-LDA), aims to model (1) the event's topics and their evolution (event segmentation), as well as (2) the associated tweets' topics and the crowd's tweeting behaviors. Therefore, the model has two major components with each capturing one perspective of our target. The conceptual model of ET-LDA is shown in Fig. 1a and its graphical model representation is in Fig. 1b. Both parts have the LDA-like model, and are connected by the link which captures the topical influences from the event on its Twitter feeds.

More specifically, in the event part, we assume that an event is formed by discrete sequentially-ordered segments, each of which discusses a particular set of topics. A segment consists of one or many coherent paragraphs available



(a) Conceptual model of ET-LDA

(b) Plate model of ET-LDA

Figure 1: The graphical model representation of ET-LDA

from the transcript of the event.[1] Each paragraph s is associated with a particular distribution of topics $\theta^{(s)}$. To model the topic evolutions in the event, we apply the Markov assumption on $\theta^{(s)}$: with some probability, $\theta^{(s)}$ is the same as the distribution of topics of previous paragraph $s-1$, measured by the delta function $\delta(\theta^{(s-1)}, \theta^{(s)})$; otherwise, a new distribution of topics $\theta^{(s)}$ is sampled for s, chosen from a $Dirichlet(\alpha_\theta)$. This pattern of dependency is produced by associating a binary variable $c^{(s)}$ with each paragraph, indicating whether its topic is the same as that of the previous paragraph or different. If the topic remains the same, these paragraphs are merged to form one segment. This variable is associated with a binomial distribution $\delta^{(s)}$ parameterized by a symmetric beta prior α_δ.

In the tweets part, we assume that a tweet consists of words which can belong to two distinct types of topics: *general* topics, which are high-level and constant across the entire event, and *specific* topics, which are detailed and relate to the segments of the event. As a result, the distribution of general topics is fixed for a tweet. However, the distribution of specific topics keeps varying with respect to the development of the event. We define a tweet in which most words belong to general topics as a general tweet, indicating a weak topical influence from the event. In contrast, a tweet with more words about the specific topics is defined as a specific tweet, indicating a strong topical influence from one segment of the event. In other words, a specific tweet refers to a segment of the event. Similar to the event segmentation, each word in a tweet is associated with a distribution of topics. It can be either sampled from a mixture of specific topics $\theta^{(s)}$ or a mixture of general topics $\psi^{(t)}$ over K topics depending on a binary variable $c^{(t)}$ sampled from a binomial distribution $\lambda^{(t)}$. In the first case, $\theta^{(s)}$ is from a referring segment s of the event, where s is chosen according to a categorical distribution $s^{(t)}$. Unlike $\delta^{(s)}$, $\lambda^{(t)}$ is controlled by an asymmetrical beta prior parameterized by the preference parameter α_{λ_γ} (for specific topics) and α_{λ_ψ} (for general topics).

[1]For many publicly televised events, transcripts are readily published by news services like NY Times etc. Paragraph outlines in the transcripts are usually determined through human interpretation and may not necessarily correspond to topic changes in the event.

An important property of the categorical distribution $s^{(t)}$ is to allow choosing any segment in the event. This reflects the fact that a person may compose a tweet on topics discussed in a segment that (1) was in the past (2) is currently occurring, or (3) will occur after the tweet is posted (usually when she expects certain topics to be discussed in the event)

To summarize, we have the following generative process:

Procedure Generation Process in ET-LDA

foreach *paragraph* $s \in S$ **do**
  draw a segment choice indicator $c^{(s)} \sim Bernoulli(\delta^{(s)})$
  if $c^{(s)} = 1$ **then**
    draw a topic mixture $\theta^{(s)} \sim Dirichlet(\alpha_\theta)$
  else
    draw a topic mixture $\theta^{(s)} \sim \delta(\theta^{(s-1)}, \theta^{(s)})$
  foreach *word* $w_s^i \in s$ **do**
    draw a topic $z_s^i \sim Multinomial(\theta^{(s)})$
    draw a word $w_s^i \sim \phi_{z_s^i}$
foreach *tweet* $t \in T$ **do**
  foreach *word* $w_t^i \in t$ **do**
    draw a topic changing indicator $c^{(t)} \sim Bernoulli(\lambda^{(t)})$
    if $c^{(t)} = 1$ **then**
      draw a topic mixture $\psi^{(t)} \sim Dirichlet(\alpha_\psi)$
      draw a general topic $z_t^i \sim Multinomial(\psi^{(t)})$
    else
      draw a paragraph $s \sim Categorical(\gamma^{(t)})$
      draw a specific topic $z_t^i \sim Multinomial(\theta^{(s)})$
    draw a word from its associated topic $w_t^i \sim \phi_{z_t^i}$

With the model hyperparameters α, β, the joint distribution of observed and hidden variables $\mathbf{w}_s$, $\mathbf{w}_t$, $\mathbf{z}_s$, $\mathbf{z}_t$, $\mathbf{c}_s$, $\mathbf{c}_t$, and $\mathbf{s}_t$ can be written as blow.

$$
\begin{aligned}
&P(\mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t | \alpha_\delta, \alpha_\theta, \alpha_\gamma, \alpha_\lambda, \alpha_\psi, \beta) = \\
&\int \cdots \int P(\mathbf{w}_s|\mathbf{z}_s, \phi) P(\mathbf{w}_t|\mathbf{z}_t, \phi) P(\phi|\beta) P(\mathbf{s}_t|\gamma^{(t)}) P(\gamma^{(t)}|\alpha_\gamma) \\
&P(\mathbf{z}_s|\theta^{(s)}) P(\mathbf{z}_t|\theta^{(s)}, \mathbf{s}_t, \mathbf{c}_t = 0) P(\theta^{(s)}|\alpha_\theta, \mathbf{c}_s) P(\mathbf{c}_s|\delta^{(s)}) P(\delta^{(s)}|\alpha_\delta) \\
&P(\mathbf{z}_t|\psi^{(t)}, \mathbf{c}_t = 1) P(\psi^{(t)}|\alpha_\psi) P(\mathbf{c}_t|\lambda^{(t)}) P(\lambda^{(t)}|\alpha_{\lambda_\gamma}, \alpha_{\lambda_\psi}) \\
&d\gamma^{(t)} d\theta^{(s)} d\delta^{(s)} d\lambda^{(t)} d\psi^{(t)} d\phi
\end{aligned} \tag{1}
$$

3.2 Inference in the Model via Gibbs Sampling

The computation of the posterior distribution of the hidden variables $\mathbf{z}_s$, $\mathbf{z}_t$, $\mathbf{c}_s$, $\mathbf{c}_t$ and $\mathbf{s}_t$ is intractable for the ET-LDA model because of the coupling between α, β. Therefore, in this paper, we utilize approximate methods like collapsed Gibbs sampling algorithm (Griffiths et al. 2004) for parameter estimation. Note that Gibbs sampling allows the learning of a model by iteratively updating each latent variable given the remaining variables.

To begin with, we need to compute conditional probability $P(z_t, z_s, w_s, w_t, c_s, c_t, s_t | \mathbf{z}'_t, \mathbf{z}'_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}'_s, \mathbf{c}'_t, \mathbf{s}'_t)$, where $\mathbf{z}'_t, \mathbf{z}'_s, \mathbf{c}'_s, \mathbf{c}'_t, \mathbf{s}'_t$ are vectors of assignments of topics, segment indicators, topic switching indicators and segment choice indicators for all words in the collection except for the one at position i in a tweet or an event's transcript. According to the Bayes rule, we can compute this conditional probability in terms of the joint probability distribution of the latent and observed variables shown in Eq.1. Next, to make the sampling procedure clearer, we factorize this joint probability as:

$$P(\mathbf{w}_s, \mathbf{w}_t, \mathbf{z}_s, \mathbf{z}_t, \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t) = P(\mathbf{w}_s, \mathbf{w}_t | \mathbf{z}_s, \mathbf{z}_t) P(\mathbf{z}_s, \mathbf{z}_t | \mathbf{c}_s, \mathbf{c}_t, \mathbf{s}_t) P(\mathbf{c}_s) P(\mathbf{c}_t) P(\mathbf{s}_t) \quad (2)$$

By integrating out the parameter ϕ we can obtain the first term in Eq.2:

$$P(\mathbf{w}_s, \mathbf{w}_t | \mathbf{z}_s, \mathbf{z}_t) = \left(\frac{\Gamma(W\beta)}{\Gamma(\beta)^W} \right)^K \prod_{k=1}^{K} \frac{\prod_{w=1}^{W} \Gamma(n^k_{sw} + n^k_{tw} + \beta)}{\Gamma(n^k_{s(.)} + n^k_{t(.)} + W\beta)} \quad (3)$$

where W is the size of the vocabulary, n^k_{sw} and n^k_{tw} are the numbers of times topic k assigned to word w in the event and the tweets. $n^k_{s(.)}$ and $n^k_{t(.)}$ are the total number of words in the event and tweets assigned to topic k. $\Gamma(\cdot)$ is the gamma function.

To evaluate the second term in Eq.2, we need to consider the value of a tweet's topic switching indicator $\mathbf{c}_t$ because it determines whether a word's topic $\mathbf{z}_t$ is general, i.e, sampling from $\psi^{(t)}$ (when $\mathbf{c}_t = 1$) or specific, i.e., sampling from $\theta^{(s)}$ (when $\mathbf{c}_t = 0$). Based on the model structure in Fig.1b, we factor the second term as $P(\mathbf{z}_s, \mathbf{z}_t | \mathbf{c}_s, \mathbf{c}_s, \mathbf{s}_t) = P(\mathbf{z}_s) P(\mathbf{z}_t | \mathbf{c}_s, \mathbf{c}_t = 0, \mathbf{s}_t) P(\mathbf{z}_t | \mathbf{c}_t = 1, \mathbf{s}_t)$ and compute each of these factors individually. So when $\mathbf{c}_t = 0$, by integrating out $\theta^{(s)}$ and canceling the factor that does not depend on this value, we obtain:

$$P(\mathbf{z}_s) P(\mathbf{z}_t | \mathbf{c}_s, \mathbf{c}_t = 0, \mathbf{s}_t) = \left(\frac{\Gamma(K\alpha_\theta)}{\Gamma(\alpha_\theta)^K} \right)^{\mathcal{S}} \prod_{i=1}^{\mathcal{S}} \frac{\prod_{k=1}^{K} \Gamma(n^{\mathcal{S}_i}_k + nt^{\mathcal{S}_i}_k + \alpha_\theta)}{\Gamma(n^{\mathcal{S}_i}_{(.)} + nt^{\mathcal{S}_i}_{(.)} + K\alpha_\theta)} \quad (4)$$

where $\mathcal{S}$ is a set of segments of the event. $n^{\mathcal{S}_i}_k$ is the number of times topic k appears in the segment $\mathcal{S}_i$, and $nt^{\mathcal{S}_i}_k$ is the number of times topic k appears in tweets, where these tweets refer to the content in segment $\mathcal{S}_i$. Similarly, when $\mathbf{c}_t = 1$, by integrating out $\psi^{(t)}$ and canceling the factor that does not depend on this value, we have:

$$P(\mathbf{z}_t | \mathbf{c}_t = 1, \mathbf{s}_t) = \left(\frac{\Gamma(K\alpha_\psi)}{\Gamma(\alpha_\psi)^K} \right)^T \prod_{i=1}^{T} \frac{\prod_{k=1}^{K} \Gamma(n^i_k + \alpha_\psi)}{\Gamma(n^i_{(.)} + K\alpha_\psi)} \quad (5)$$

in which T is the total number words in tweet t which are under the general topics, and n^i_k is the number of times topic k assigns to words i.

Next, we evaluate the third term in Eq.2. By integrating out $\delta^{(s)}$ we compute:

$$P(\mathbf{c}_s) = \frac{\Gamma(2\alpha_\delta)}{\Gamma(\alpha_\delta)^2} \frac{\Gamma(S^0_s + \alpha_\delta)\Gamma(S^1_s + \alpha_\delta)}{\Gamma(S + 2\alpha_\delta)} \quad (6)$$

where S is the total number of paragraphs in an event. S^1_s is the number of segments (the number of times the topic of paragraph s differs from its preceding paragraph, i.e., $\mathbf{c}_s = 1$).

Similarly, for the fourth term in Eq.2, we integrate out $\lambda^{(t)}$ and get:

$$P(\mathbf{c}_t) = \prod_{t \in T} \frac{\Gamma(\alpha_{\lambda_\gamma} + \alpha_{\lambda_\psi})}{\Gamma(\alpha_{\lambda_\gamma})\Gamma(\alpha_{\lambda_\psi})} \frac{\Gamma(M^0_t + \alpha_{\lambda_\gamma})\Gamma(M^1_t + \alpha_{\lambda_\psi})}{\Gamma(M_t + \alpha_{\lambda_\gamma} + \alpha_{\lambda_\psi})} \quad (7)$$

where M_t is the total number of words in tweet t, M^0_t is the number of words that are under the specific topics, and M^1_t is the number of words in t that are under the general topics. Last, we need to derive the fifth term. Again, by integrating out $\gamma^{(t)}$ we have:

$$P(\mathbf{s}_t) = \left(\frac{\Gamma(K\alpha_\gamma)}{\Gamma(\alpha_\gamma)^K} \right)^T \prod_{i=1}^{T} \frac{\prod_{s=1}^{S} \Gamma(n^i_s + \alpha_\gamma)}{\Gamma(n^i_{(.)} + S\alpha_\gamma)} \quad (8)$$

where n^i_s is the number of times paragraph s (in fact its associated segment) is referred by tweet t.

Now, the conditional probability can be obtained by multiplying and canceling of terms in Eq.3–8. We show the core case (when $c_s = 0$) here while the other case (when $c_s = 1$) is omitted due to the space limit.

$$P(z_t, z_s, w_s, w_t, c_s = 0, c_t = 0, s_t | \mathbf{z}'_t, \mathbf{z}'_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}'_s, \mathbf{c}'_t, \mathbf{s}'_t) = \frac{n^k_{sw} + n^k_{tw} + \beta - 1}{n^k_{s(.)} + n^k_{t(.)} + W\beta - 1} \times \frac{n^{S_i}_k + nt^{S_i}_k + \alpha_\theta - 1}{n^{S_i}_{(.)} + nt^{S_i}_{(.)} + K\alpha_\theta - 1} \times \frac{n^i_s + \alpha_\gamma - 1}{n^i_{(.)} + S\alpha_\gamma - 1} \frac{M^0_t + \alpha_{\lambda_\gamma} - 1}{M_t + \alpha_{\lambda_\gamma} + \alpha_{\lambda_\psi} - 1} \times \frac{S^0_t + \alpha_\delta - 1}{M_t + 2\alpha_\delta - 1} \quad (9)$$

and when $c_t = 1$ we have the conditional probability:

$$P(z_t, z_s, w_s, w_t, c_s = 0, c_t = 1, s_t | \mathbf{z}'_t, \mathbf{z}'_s, \mathbf{w}_s, \mathbf{w}_t, \mathbf{c}'_s, \mathbf{c}'_t, \mathbf{s}'_t) = \frac{n^k_{sw} + n^k_{tw} + \beta - 1}{n^k_{s(.)} + n^k_{t(.)} + W\beta - 1} \times \frac{n^i_k + \alpha_\psi - 1}{n^i_{(.)} + K\alpha_\psi - 1} \times \frac{n^i_s + \alpha_\gamma - 1}{n^i_{(.)} + S\alpha_\gamma - 1} \frac{M^1_t + \alpha_{\lambda_\gamma} - 1}{M_t + \alpha_{\lambda_\gamma} + \alpha_{\lambda_\psi} - 1} \times \frac{S^1_t + \alpha_\delta - 1}{M_t + 2\alpha_\delta - 1} \quad (10)$$

In both of these expressions, counts are computed without taking into account assignments of the considered word w_s^i and w_t^i. After algebraic manipulation to Eq.9 and 10, we can easily derive Gibbs update equations for variables z_t, z_s, c_s, c_t and s_t which are omitted here.[2] Sampling with these questions is fast and in practice convergence can be achieved in time similar to that needed by LDA implementations.

4 Experiments

In this section, we examine the effectiveness of our proposed joint model against other baselines. Three main tasks are undertaken to evaluate the ET-LDA: (1) the topics extracted from the whole corpus (tweets and transcripts of events) are compared with those separately extracted from the LDA model, (2) the capability of predicting topical influences of the events on unseen tweets in the test set is compared with LDA, and (3) the quality of event segmentation is compared with LCSeg – a popular HMM-based segmenting tool in the literature (Galley et al. 2003).

Data Sets and Experimental Setup We use two large-scale tweet datasets associated with two events from different domains: (1) President Obama's Middle East speech on May 19, 2011 and (2) the Republican Primary debate on Sept 7, 2011. The first tweet dataset consists of 25,921 tweets tagged with "#MESpeech" and the second dataset consists of 121,256 tweets tagged with "#ReaganDebate". Both tweet datasets were crawled via the Twitter API using these two hashtags (which were officially posted by the White House and NBC News, respectively, before the event). In the rest of this paper, we use the hashtags to refer to these events. Furthermore, we split both tweet datasets into a 80-20 training and test sets.

We obtained the transcripts of both events from the New York Times,[3] where **MESpeech** has 73 paragraphs and **ReaganDebate** has 230 paragraphs. We applied preprocessing to both tweets and transcripts by removing non-English tweets, retweets, punctuation and stopwords and stemming all terms. Further, it is known that topic modeling methods (including LDA) behave badly when applied to short documents (Hu et al. 2009). To remedy this, we follow the scheme in (Sahami and Heilman 2006) to augment a tweet's context. First, we treat a tweet as a query and send it to a search engine. After generating a set of top-n query snippets $d_1, ..., d_n$, we compute the TF-IDF term vector v_i for each d_i. Finally, we pick the top-m terms from v_i and concatenate them to t to form an expanded tweet. In the experiments, we used the Google custom search engine for retrieving snippets and set $n = 5$ and $m = 10$. Such augmentation was applied to both tweet datasets.

In the experiments, we used the Gibbs sampling algorithm for training ET-LDA on the tweets dataset with the transcript of associated event. The sampler was run for 1000 iterations for both datasets. Coarse parameter tuning for the prior distributions was performed. We varied the number of topics K in ET-LDA and chose the one which maximizes the log-likehood $P(\mathbf{w}_s, \mathbf{w}_t|K)$, a standard approach in Bayesian statistics (Griffiths and Steyvers 2004). As a result, we set K = 20. In addition, we set model hyperparameters $\alpha_\delta = 0.1, \alpha_\theta = 0.1, \alpha_\gamma = 0.1, \alpha_{\lambda_\gamma} = \alpha_{\lambda_\psi} = 0.5, \alpha_\psi = 0.1$, and $\beta = 0.01$.

Topic Modeling: We first study the performance of ET-LDA on topic modeling for the two events and their tweets against the baseline LDA (which was trained on the event transcripts and tweet datasets separately with $K = 20$). Table 2 and 3 present the top words, i.e., the highest probability words from topics for (**i**) top specific topics discovered from a sample of 3 (out of 7 for **MESpeech**, or, out of 14 for **ReaganDebate**) segments of the events, and (**ii**) for a sample of the general topics from the tweets collection using the ET-LDA model. The results of segmentation are shown next. For comparison, both tables also list the top words for the topics discovered from (**iii**) the event and (**iv**) tweets individually using LDA. Note that all of the topics have been manually labeled for identification (e.g. "Arab Spring", "Immigration") to reflect our interpretation of their meaning from their top words.

It is clear that all specific and general topics from the ET-LDA model are very reasonable from a reading of the transcripts. Furthermore, we observe that the specific topics are sensitive to the event's context and keep evolving as the event progresses. On the other hand, general topics and their top words capture the overall themes of the event well. But unlike specific topics, these are repeatedly used across the entire event by the crowd in their tweets, in particular when expressing their views on the themes of the event (e.g., "Arab spring", "Immigration") or even on some popular issues that are neither related nor explicitly discussed in the event (e.g., "Obama" in **MESpeech**, "Conservative" in **ReaganDebate**).

The results of LDA seem less reasonable by comparison. Although LDA may extract general topics like "Israel Palestine issues" just like ET-LDA since these topics remain constant throughout the document, LDA cannot extract specific topics for the event. In fact, "Israel Palestine issues" shows the advantage of ET-LDA: it is the top general topic for entire tweet collection (which is very relevant to and influenced by the event) whereas LDA fails to identify that (its top topic is 'Obama' which is less relevant). The data showed that people tweeted about this issue a lot. Besides, some top words for LDA topics are not so related to the event. This lack of correspondence is more pronounced for LDA when it is applied to the tweet datasets, e.g., *GOP Job Approval* in topic "Obama" of the tweets corpus by LDA. This is mainly because ET-LDA successfully characterizes the topical influences from the event on its Twitter feeds such that the content/topics of tweets are regularized, whereas the LDA method ignores these influences and thus gives less reasonable results.

[2] For each variable, in order to derive its update equation, one can first pick the factors that depend on it in Eq.2 and then select the corresponding factors in Eq.9 or Eq.10 under the condition that $c_s = 0$. For $c_s = 1$, the derivation is the same.)

[3] *http://www.nytimes.com/2011/05/20/world/middleeast/20prexy-text html* and *http://www.nytimes.com/2011/09/08/us/politics/08republican-debate-text.html*

Table 2: Top words in topics from **MESpeech** for ET-LDA and LDA

	Topics		Top words
ET-LDA (Top specific topic of each sampled segment of the event)	S1	"Foreign policy"	Hillary Clinton State Dept Sec Leaders Rights Transition MiddleEast Democracy
	S2	"Terrorism"	Bin Laden Mass Murderer Dignity Power Blood Dictator Message peaceful
	S4	"Aid Egypt"	Aid Trade Reform Greatest Resource Billion Debt Egypt Support Help Good News
ET-LDA (Top general topics from the tweets collection)	"Arab spring" "Obama" "Israel & Palestine"		Arabia Bahrain Mosques Stepped Mespeech Syrian Leader Government Religion Obama Economics Failed President Job Tough Critique Jews Policies Weakness Israel Palestine Borders State Negotiations Lines Hamas Permanent Occupation
LDA on event (3 out of a total of 20 topics)	"MiddleEast/Arab" "Security/Terrorism" "Israel Palestine issues"		Young People Deny month Country Region Democracy Women violence Cairo Iraq Many America Home Transition State Peace Security Conflict Hate Blood Al Qaeda Country Palestinian security Israel Know Between Leader resolve Issue Boarder
LDA on tweets (3 out of a total of 20 topics)	"Arab Spring" "Security/Peace" "Obama"		Obama Town Assad Month Syria Libya Countries Leave Dictators must Jews Iran Bin Laden Dead Oil Region War Murder Iraq Risk Nuclear Peace Army Wonderful Obama Job Approval GOP Middle East Mespeech Talking Economics

Table 3: Top words in topics from **ReaganDebate** for ET-LDA and LDA

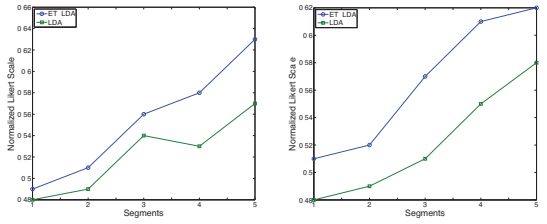
	Topics		Top words
ET-LDA (Top specific topic from each sampled segment)	S2	"Job market"	Job Payroll Economy Crisis Market Commitment Income Tax Wages Pledges Cured
	S3	"Health care"	Obamacare Wrong Health Care Regulations Bush People Law Everyone Stage Treats
	S10	"Social sec."	Perry Social Security Benefits Ponzi Scheme Republican Annual Circumstances
ET-LDA (Top general topics from Tweets collection)	"Conservative" "Obama" "Immigration"		Ron Paul President Real Blessed GOP Tea Party Purpose Government Conservative Support President Job Obamacare Health Care Critique Policies Democrats Low Rate U.S country Legislative Legal Immigration Law Solution Fence Economics Committed Debate Taxpayer
LDA on event (out of 20 topics)	"Social Sec." "Regulations" "Health care"		Constitution Law Government Wrong Federal Question Funding Monstrous Financially Fed Funding Expenditures Devastating Economy Policies Hurt Admin. Democratic President Plan Romney Cheaper Free Debate Mandate Individual Obamacare Question Better
LDA on tweets (out of 20 topics)	"Social Sec." "Obama" "Economics"		Perry Ridiculous Crazy Ponzi Exaggerated Provocative Tcot Blasts Investment Reformed Warfare Job Creation Obamacare Approval Poll Troop Withdraw Working Country Iraq Vote Monster Stupid Bloody Congress Obama Jobs Apple Lost One Sided Blasting Hopeless

(a) Segmentation of MESpeech **(b)** Segmentation of ReaganDebate

Figure 2: Segmentation of the Event

Prediction Performance: Next, we study the prediction performance of ET-LDA. Specifically, we are interested in the prediction of topical influences from the event on the unseen tweets in our test set (20% of total tweets). Thus, we first run the Gibbs sampling algorithm, described in previous section, on the training set for each event/tweet dataset. Then we extend the sampler state with samples from the test set. For comparison, we adopt LDA as our baseline approach. However, since LDA treats the event and tweets individually, we measure the topical influences by the distance of topic mixtures of the unseen tweets to the ones of the segments of the event (as determined by ET-LDA in advance). This distance is measured by the Jensen-Shannon divergence.

To evaluate the "goodness" of prediction results by our proposed model, we asked 30 graduate students from the engineering school of our university (who were selected as they follow the news closely and tweet at least three times per week) to manually label the quality and strength of the predicted topical influences from events on the unseen tweet datasets on a Likert scale of 1 to 5 rating. We then averaged these ratings over the value diversity (i.e., normalization). In



(a) MESpeech **(b)** ReaganDebate

Figure 3: Predictive performance of ET-LDA compared with LDA model on 5 randomly sampled segments.

Fig. 3a and 3b, we present the results of the two methods on 5 randomly sampled segments.

In light of the observed differences in Fig. 3a and 3b, we study the statistical significance of ET-LDA with respect to LDA. We perform paired-t-tests for models with a significance level of $\alpha = 0.05$ ($p = 0.0161$ and $p = 0.0029$ for **MESpeech** and **ReaganDebate**, respectively). This reveals that the improvement in prediction performance of ET-LDA is statistically significant.

Event Segmentation: Finally, we study the quality and effectiveness of ET-LDA on the segmentation of the two events based on their transcripts. The results of the event segmentation (obtained using $K = 20$ in ET-LDA) are shown in Fig. 2a and 2b. To evaluate our model, we compare its results with the ones from a popular HMM-based tool *LC-*

Table 4: Comparisons of segmentation results on two events

	MESpeech		ReaganDebate	
	ET-LDA	LCSeg	ET-LDA	LCSeg
P_k	0.295	0.361	0.31	0.397

Seg (trained on 15-state HMM) on the P_k measure (Beeferman, Berger, and Lafferty 1999). Note that this measure is the probability that a randomly chosen pair of words from the event will be incorrectly separated by a hypothesized segment boundary. Therefore, the lower P_k indicates better agreement with the human-annotated segmentation results, i.e., better performance. In practice, we first ask four graduate students in our department to annotate the segments of the events based on their transcripts (two for each event) and later ask another graduate student to judge, for one event, which human annotation is better. We pick the better one of each event and treat it as the hypothesized segmentation. Then, we compute the P_k value. The results of two methods are shown in Table. 4.

The results show that our model significantly outperforms the *LCSeg* – as the latter cannot merge topic mixtures in paragraphs according to their similarity, and thus places a lot of segmentation boundaries (i.e., over-segmented), resulting in poor performance.

5 Conclusion

In this paper, we have described a joint statistical model ET-LDA that characterizes topical influences between an event and its associated Twitter feeds (tweets). Our model enables the topic modeling of the event/tweets and the segmentation of the event in one unified framework. We evaluated ET-LDA both quantitatively and qualitatively through three tasks. Based on the experimental results, our model shows significant improvements over the baseline methods.

We believe this paper presents the first step towards understanding complex interactions between events and social media feedback. In fact, beyond the transcripts of publicly televised events that we used in this paper, ET-LDA can also handle other forms of text sources that describe an event. For example, one can explore how people respond to an event (and how it is different from journalists' responses in media) by applying our model to the news articles and the social media feedback about this event. We also believe that this paper reveals a perspective that is useful for the extraction of a variety of further dimensions such as sentiment and polarity. For example, one can examine how the crowd's mood is affected by the event based on the topical influences.

Acknowledgements: ET-LDA was initially formulated and prototyped at Avaya Labs during a summer internship there by Yuheng Hu, who thanks Dorée Duncan Seligmann for helpful discussions. Hu and Kambhampati's research at ASU is supported in part by the ONR grant N000140910032.

References

Beeferman, D.; Berger, A.; and Lafferty, J. 1999. Statistical models for text segmentation. *Machine learning*.

Blei, D., and Lafferty, J. 2006. Dynamic topic models. In *ICML*. ACM.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*.

Boykin, S., and Merlino, A. 2000. Machine learning of event segmentation for news on demand. *Communications of the ACM*.

Chang, J., and Blei, D. 2009. Relational topic models for document networks. In *Artificial Intelligence and Statistics*.

Dielmann, A., and Renals, S. 2004. Dynamic bayesian networks for meeting structuring. In *ICASSP*. IEEE.

Dietz, L.; Bickel, S.; and Scheffer, T. 2007. Unsupervised prediction of citation influences. In *ICML*. ACM.

Galley, M.; McKeown, K.; Fosler-Lussier, E.; and Jing, H. 2003. Discourse segmentation of multi-party conversation. In *ACL*. Association for Computational Linguistics.

Griffiths, T., and Steyvers, M. 2004. Finding scientific topics. *PNAS*.

Griffiths, T. L.; Steyvers, M.; Blei, D. M.; and Tenenbaum, J. B. 2004. Integrating topics and syntax. In *NIPS*.

Hearst, M. 1993. Texttiling: A quantitative approach to discourse segmentation. *Sequoia*.

Hofmann, D. 2001. The missing link-a probabilistic model of document content and hypertext connectivity. In *NIPS*. The MIT Press.

Hu, X., and Liu, H. 2012. Text analytics in social media. *Mining Text Data*.

Hu, X.; Sun, N.; Zhang, C.; and Chua, T. 2009. Exploiting internal and external semantics for the clustering of short texts using world knowledge. In *CIKM*. ACM.

Hu, Y.; John, A.; Seligmann, D.; and Wang, F. 2012. What were the tweets about? topical associations between public events and twitter feeds. In *Proceedings of ICWSM*. AAAI.

Hu, Y.; John, A.; and Seligmann, D. 2011. Event analytics via social media. In *Proceedings of the 2011 ACM workshop on Social and behavioural networked media access*. ACM.

Nallapati, R.; Ahmed, A.; Xing, E.; and Cohen, W. 2008. Joint latent topic models for text and citations. In *KDD*. ACM.

Purver, M.; Griffiths, T.; Körding, K.; and Tenenbaum, J. 2006. Unsupervised topic modelling for multi-party spoken discourse. In *ACL*. Association for Computational Linguistics.

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing microblogs with topic models. In *ICWSM*. The AAAI Press.

Sahami, M., and Heilman, T. 2006. A web-based kernel function for measuring the similarity of short text snippets. In *WWW*. ACM.

Shamma, D.; Kennedy, L.; and Churchill, E. 2009. Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*. ACM.

Titov, I., and McDonald, R. 2008. Modeling online reviews with multi-grain topic models. In *WWW*. ACM.

Wang, X., and McCallum, A. 2006. Topics over time: a non-markov continuous-time model of topical trends. In *KDD*. ACM.