What Were the Tweets About? Topical Associations between Public Events and Twitter Feeds

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

Abstract

Social media channels such as Twitter have emerged as platforms for crowds to respond to public and televised events such as speeches and debates. However, the very large volume of responses presents challenges for attempts to extract sense from them. In this work, we present an analytical method based on joint statistical modeling of topical influences from the events and associated Twitter feeds. The model enables the auto-segmentation of the events and the characterization of tweets into two categories: (1) *episodic* tweets that respond specifically to the content in the segments of the events, and (2) *steady* tweets that respond generally about the events. By applying our method to two large sets of tweets in response to President Obama's speech on the Middle East in May 2011 and a Republican Primary debate in September 2011, we present what these tweets were about. We also reveal the nature and magnitude of the influences of the event on the tweets over its timeline. In a user study, we further show that users find the topics and the episodic tweets discovered by our method to be of higher quality and more interesting as compared to the state-of-the-art, with improvements in the range of 18-41%.

1 Introduction

Social media channels such as Twitter have emerged as valuable platforms for information sharing and communication, in particular during public and televised events such as the State of the Union addresses by the President of the United States, the annual Academy Awards ceremony, etc. During such events large amounts of commentary have been contributed by crowds via Twitter. For example, over 22,000 tweets were posted around President Obama's hour long speech on the Middle East in May 2011. Likewise, we retrieved more than 110,000 tweets about the Republican Primary debate in September 2011 within two hours.

This burst of information, on the one hand, enriches the user experience for the live event. On the other hand, it poses tremendous challenges for attempts to extract sense from the tweets, which is critical to applications for journalistic investigation, playback of events, storytelling, etc. How can we identify *what these tweets were about*? And *did these tweets refer to specific parts of the event and if so, what parts?* Furthermore, *what was the nature and magnitude of the influence of the event over the tweeting behavior of the crowd?*

We answer these questions by devising a computational method geared towards extracting sense from the crowd's tweets in the context of the public events that they are in response to. Therefore, the focus in this paper is quite different from the literature of sensemaking of tweets in that the existing techniques tend to focus on either tweets in isolation from the context of the event or their usage patterns, e.g., volumes of tweets, networks of audience members and their tag relations, etc. (see Section 2 for related work).

Our approach is based on the characterization of topical influences of the event on its Twitter feeds. Intuitively, since tweets are generated by the crowd to express their interest in the event, they are essentially influenced by the topics covered in the event in some way. In order to characterize such influences, we first propose that rather than enforcing tweets to be correlated only with the topics of the event that occur within time-windows around the tweets' timestamps (a common approach in the literature, e.g., (Shamma, Kennedy, and Churchill 2009)), they should correlate to any topic in the event. Next, we claim that a person can compose her tweets in a variety of ways to respond to the event. To take an example, she may choose to comment directly on a specific topic in the event which is concerning and/or interesting to her. So, her tweets would be deeply influenced by that topic. In another situation, she could comment broadly about the event. In consequence, the tweets would be less influenced by the specific topics but more by the general themes of the event.

Our approach models exactly these two distinct tweeting behaviors of the crowd, and later experimentally confirm their existence by finding these types of tweets. We deem the tweets *episodic* tweets if they are generated in the first way, since their content refers to the specific topics of the event. To determine what these topics are about and where they appear, our approach splits the entire event into several sequential segments in which a particular set of topics is covered. On the other hand, we deem the tweets *steady* tweets if they are generated in the second way, because their topics stay steady on the general themes across the event rather than being affected by its varying context. The patterns of

[*]Most of the work was performed while the author was an intern at Avaya Labs, NJ. Subsequently at ASU, he was supported by the ONR grant N000140910032.

episodic and steady tweets and their correlations to the event shows how people responded to the event.

Our Contributions. We have developed a joint statistical topic model for making sense of a vast amount of tweets in response to public events. Based on the event's topical influences, it finds segments in the event's transcript and concurrently classifies tweets into two categories (episodic/steady) in an unsupervised manner. Enabled by this model, people would gain much deeper insights about the event (e.g., which topic was most interesting to the crowd) and the tweets around it (e.g., what they were about). In addition, the model also sheds light on the *nature* of the crowd's tweeting behaviors in the following ways: (1) Reveals the topical context of the tweets, and (2) Shows how the tweets evolve over the event's timeline. Such work, to our knowledge, has not been investigated before and is not feasible with other alternative methods. For example, manual coding of the tweets is prohibitively expensive, and pure topic modeling (such as LDA (Blei, Ng, and Jordan 2003)) does not easily enable the segmentation of the event and distinguishing between two types of tweets.

Our results. We perform quantitative studies of the proposed model over two large sets of tweets in response to President Obama's speech on the Middle East in May 2011 and a Republican Primary debate in Sept. 2011. Our results reveal several key insights into how people responded to the event: (1) We find the crowd's tweeting behavior varies with the timeline of the event. More episodic tweets were witnessed during the event and less were found before or after the event (the percentages on average are 55%/35%/38%). (2) We also discover that people showed a greater level of engagement (the total number of tweets and the percentage of episodic tweets) in the Republican debate which centered around national issues as opposed to President Obama's Middle East speech. (3) We find that, as the event evolved, the crowd tended to comment on any topic in the event – that topic could have been discussed before, was being discussed currently, or was expected to be discussed later on.

We also address the issue of evaluating results in the absence of ground truth. This is accomplished with a user study with 31 active Twitter users in a university. We evaluate the goodness of sampled topics and episodic tweets by different methods based on the participants' perception of their quality. From the participant responses in the user study, we observe that our approach yields better quality, with improvements in the range of 18%–41% over the state-of-the-art.

The rest of the paper is organized as follows. In Section 2 we discuss related work. Section 3 presents our observation of the crowd's tweeting patterns to an event. In Section 4 we present our approach. Section 5 and 6 present quantitative studies and subjective evaluations, followed by a discussion of their results. Section 7 concludes the paper.

2 Related Work

While the topic of making sense of a crowd's responses to a media event is relatively new, there have been some recent attempts to characterize events by the tweets around them. These works include inferring structures of events using Twitter usage patterns (Shamma, Kennedy, and Churchill 2009), event detection or summarization via tweets (Weng et al. 2011; Chakrabarti and Punera 2011), exploring events by the classification of audience types and categories on Twitter (Vieweg et al. 2010), and sentimental analysis of tweets to understand the events (Diakopoulos and Shamma 2010).

There is also a rich body of work that investigates tweets outside the context of events. This includes studies of why people tweet (Java et al. 2007; Zhao and Rosson 2009), representations of tweet content using a labeled topic model (Ramage, Dumais, and Liebling 2010), characterizing individuals' activity on Twitter through a content-based categorization of the type of their tweets (Naaman, Boase, and Lai 2010), and also quantifying and predicting social influence on Twitter and other social media (Cui et al. 2011; Bakshy et al. 2011).

The focus of most of the above works is to either better understand events or to analyze tweets on their own. Thus, they do not provide insights on how to extract sense from the tweets around the events. Furthermore, no analytical method has been proposed to study the correlations between tweets and events. We have already accomplished novel work in this direction and published relevant results in (Hu, John, and Seligmann 2011). In this paper, we propose a new approach to capture influences of the event on its associated Twitter feeds and provide a comprehensive and in-depth analysis of the associated textual content.

3 Understanding Tweeting Behavior

In this section, we present a preliminary understanding of a crowd's response to an event they are interested in. As an example, Figure 1 shows how the crowd interacted over the timeline of the Republican Primary debate, namely, *before*, *during* and *after* the event. The total number of tweets we collected for this event was over 110,000.

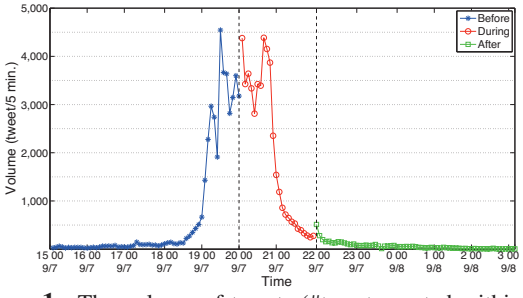


Figure 1: The volume of tweets (#tweets posted within 5 min. time window) during 09/07 15:00-09/08 3:00. The debate was during 09/07 20:00-22:00. All tweets were tagged *#ReaganDebate*.

Based on the graph, we make three observations: (1) The swell of conversation occurred mostly within 1 hour right *before* the debate started, indicating that a large number of people began to pay attention to it then. Since the debate had not started yet, we conjecture their responses were mostly tangential (e.g., posted for presence) or commentaries about the general themes of the debate (which were known in advance). (2) The volume of tweets fluctuated *during* the de-

bate, indicating different levels of involvement of the crowd with the evolving debate. We conjecture these changes were due to the fact that an event is made up of segments in sequence. Each segment covers a set of topics which may be uniquely interesting to the crowd and may influence their responses to be very specific to the content of the event. (3) A much smaller volume of tweets was witnessed right *after* the debate ended, indicating most people quickly lost interest. We conjecture these tweets were of a different nature (e.g. slightly more specific to the content of the event) from the ones posted before the event, as the crowd had just listened to or experienced the event.

In addition to the above observations as reflected by the Twitter volume, we can further understand the crowd's responses from a different angle by analyzing their content. As mentioned earlier, this is nontrivial due to the vast amount of tweets. Hence, we first analyzed a small sample of tweets through manual inspection. We find that a tweet's content can be either weakly or strongly influenced by the debate's content. Tweets with weak correlations used words that were mostly about the general topics of the debate. So they seemed to be steady and less affected by the debate's progress. On the other hand, the words used in tweets with strong correlations were mostly related to specific topics, particularly influenced by the part of the debate that they responded to. Consequently, they seemed to be more episodic. Moreover, we find the pattern of steady versus episodic complies with the timeline of the debate. *Before* (and *after*) debate, most tweets were steady, while the episodic tweets were seen more frequently *during* the debate. According to these findings, our conjectures earlier in this section seem to be verified although the sample is limited.

4 Modeling Topical Influences

The observations mentioned above highlight the importance of developing models that can characterize the crowd's involvement with the event. Since such involvement (tweeting) is topically influenced by the event, which itself is topically evolving, we propose a novel approach based on latent topic modeling to model this complexity.

Our proposed model is called the joint Event and Tweets LDA (ET-LDA), which generalizes LDA (Blei, Ng, and Jordan 2003) by jointly modeling the topic segmentation of an event and two distinct types of topics within associated tweets. ET-LDA assumes that: (1) 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 from the transcript of the event[1]. Creating these segments follows a generative process in ET-LDA: First, we treat each paragraph in an event as being generated from a particular distribution of topics, where each topic is a probability distribution over a vocabulary. Next, we apply the Markov assumption on the distribution over topics covered in the event: with some probability, the topic distribution for paragraph s is the same as the previous paragraph $s-1$; otherwise, a new distribution is sampled over topics for s. This pattern of dependency is produced by associating a binary variable 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.

Furthermore, ET-LDA assumes that: (2) 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 concrete and relate to specific segments of the event. A tweet in which most words belong to general topics is defined as a *steady* tweet, indicating a weak topical influence from the event, whereas a tweet with more words from specific topics is defined as an *episodic* tweet, indicating a strong topical influence from a segment of the event. In other words, an episodic tweet *refers* to a segment of the event. Similar to the event segmentation, composing tweets also follows a generative process in ET-LDA. To begin with, we assume that the distribution of general topics is fixed for a tweet since it is a response tagged with the official hashtag of the event (hence it should be related to the event). On the contrary, the distribution of specific topics keeps varying with respect to the evolution of the event, because it is a more directed and intended response. So, when a person wants to compose a tweet to comment on the on-going event, she has two choices on picking the appropriate words: with some probability, a word w is sampled from the mixture of general topics about the event, otherwise, it is sampled from the mixture of specific topics which occurs "locally" in the parts of the event that w refers to. The hypothesis behind the second case is that, the audience may be influenced by a set of topics that are covered by a particular part (i.e., a segment) of the event. As a result, when she picks a word to respond to that part of the event, its topic is likely to be among the topics that specifically appeared in that segment. For example, consider a tweet which was posted at the beginning of President Obama's Middle East speech: "*Sec Clinton introducing President Obama on #Mideast #StateDept #MESpeech*". It can be viewed as a mixture of general topics "*Middle East*" that was shared across the entire tweets corpus (words: *"#Mideast" and "#MESpeech"*), and specific topic "*Foreign policy*", sensitive to the part of the event when the Secretary of State, Hillary Clinton was introducing President Obama (words: *"Sec"* , *"Clinton"* and *"#StateDept"*). Note that this specific topic only occurred in the tweets that were posted at beginning of the event. Similar to the segmentation of the event, the preference of specific topics versus general topics is controlled by a binary variable associated with each word of a tweet.

Fig. 2 shows the Bayesian graphical model for ET-LDA. Mathematically, an event may choose to draw a word's topic z_s^i from a mixture of topics $\theta^{(s)}$ associated with the paragraph s. $\theta^{(s)}$ is a multinomial distribution over K topics, determined by a binary variable $c^{(s)}$ under the governance of a beta prior $\delta^{(s)}$. If $c^{(s)}=0$, then $\theta^{(s)}=\theta^{(s-1)}$, and s and its preceding paragraph $s-1$ are merged into a segment; other-

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

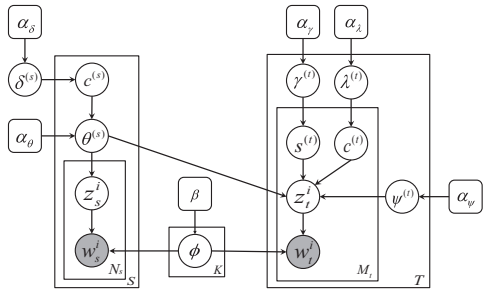


Figure 2: Graph model of ET-LDA. $S(T)$ is a set of paragraphs (tweets). Z_s (Z_t) is the topic for paragraph s (tweet t), which can be drawn either from topic mixture $\theta^{(s)}$ of the event or topic mixture $\psi^{(t)}$ of the tweets corpus. Shaded variables W_s^i and W_t^i are the ith word in s and t and are observed in the dataset.

wise, $\theta^{(s)}$ is drawn from a Dirichlet prior with parameter α_θ for creating a new segment. On the other hand, the topic for a word in a tweet can be sampled from a mixture of specific topics $\theta^{(s)}$ or a mixture of general topics $\psi^{(t)}$ over K topics given a distribution $c^{(t)}$ defining the preference. In the first case, $\theta^{(s)}$ is from a referring segment s of the event, where s is chosen according to a categorical distribution $s^{(t)}$. Although $c^{(t)}$ and $c^{(s)}$ share almost the same functionality, $c^{(t)}$ is controlled by an asymmetrical beta prior, which sets the preference parameter α_{λ_γ} (for specific topics) and α_{λ_ψ} (for general topics) accordingly. Besides, an important property of the categorical distribution $s^{(t)}$ is to allow choosing any segment, which reflects the fact that a person may compose a tweet conveying topics that have been discussed or are being currently discussed or will be discussed after the tweet is posted. Last, ϕ is the word distribution over a vocabulary with corresponding parameter β.

As inference on models in the LDA family is intractable, a variety of approximate algorithms have been developed to estimate the parameters of these models. In this paper, we exploit the Gibbs sampling method for ET-LDA. As a result, the posterior estimates of $\theta^{(s)}$ and $\psi^{(t)}$ given the training set can be calculated using Equation 1. Due to the space limit, the detailed inference is omitted.

$$\theta^{(s)} = \frac{n_k^{\mathcal{S}_i} + nt_k^{\mathcal{S}_i} + \alpha_\theta - 1}{n_{(.)}^{\mathcal{S}_i} + nt_{(.)}^{\mathcal{S}_i} + K\alpha_\theta - 1}, \quad \psi^{(t)} = \frac{n_k^i + \alpha_\psi - 1}{n_{(.)}^i + K\alpha_\psi - 1} \quad (1)$$

where $\mathcal{S}$ is a set of segments of the speech. $n_k^{\mathcal{S}_i}$ is the number of times topic k appears in the segment $\mathcal{S}_i$. $nt_k^{\mathcal{S}_i}$ is the number of times topic k appears in tweets, where these tweets specifically response to the paragraphs in $\mathcal{S}_i$. n_k^i is the number of times topic k appears in the tweets corpus.

5 Experiments

In this section, we study the performance of ET-LDA on two large sets of tweets, each associated with a public televised event. We present: (1) the general and specific topics of both events extracted by ET-LDA, (2) the evolution of episodic tweets over the event's timeline, and (3) the distribution of segments of the events as they were referred to by the episodic tweets. Our experiments are based on quantitative studies and subjective evaluations. In Section 6, we confirm our conjectures presented in Section 3 through the experimental results.

5.1 Experimental Setup

Data collection. To perform the experiments, we crawled tweets for two events using the Twitter API. The first event is President Obama's speech on the Middle East, where we obtained the tweets tagged with "*#MESpeech*". The second is the Republican Primary debate, where the tweets were tagged with "*#ReaganDebate*". Note that we only consider tweets with these hashtags, officially posted by the White House and NBC News, respectively, before the events. We obtained the transcripts of both events from the New York Times[2]. We preprocessed both datasets and the transcripts by removing non-English tweets, retweets, punctuation and stopwords and stemming all terms. Table 1 summarizes the properties of these datasets after preprocessing. We use the hashtags to refer to these events in the rest of this paper.

Table 1: Properties of datasets used in our experiments

Events	**MESpeech**	**ReaganDebate**
Event Air Time	05/19/2011 12:14PM-1:10PM	09/07/2011 8:00PM-10:00PM
Time span of tweets	05/18 - 05/25	09/06 - 09/13
Total #Tweets	11,988	112,414
#Tweets before event	1,916	42,561
#Tweets during event	4,629	46,672
#Tweets after event	5,443	23,181

Expanding tweets. It is known that topic modeling methods behave badly when applied to short documents such as tweets. To remedy this, we need to expand the tweets in some way to augment their context. Current efforts include using Wikipedia to enrich tweets (Hu et al. 2009), grouping tweets by same authors (Zhao et al. 2011), etc. Inspired by (Sahami and Heilman 2006), our approach here treats tweet t as a query and sends 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 $m = 10$.

Model settings. We used the Gibbs sampling algorithm for training ET-LDA on the tweets datasets with the transcript. The sampler was run for 2000 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 $P(W_s, W_t|K)$, a standard approach in Bayesian statistics (Griffiths and Steyvers 2004). As a result, we set $K = 20$. In addition, we set model hyperparameter $\alpha_\delta = 0.1, \alpha_\theta = 0.1, \alpha_\gamma = 0.1, \alpha_{\lambda_\gamma} = \alpha_{\lambda_\psi} = 0.5, \alpha_\psi = 0.1$, and $\beta = 0.01$.

[2] 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

5.2 Quantitative Studies

Topics from the ET-LDA Model. The results of segmentation by ET-LDA are shown in Figure 3a and Figure 3b, for two events. We first study the topics discovered from the two datasets by our proposed model. Table 2 and Table 3 present the highest probability words from two distinct types of topics – in the rest of this paper, we refer to them as **top words**. For the specific topics (under the column **Specific**), we directly pick the top 2 from the distribution of topics for each segment of the event. The topics that are ubiquitously and consistently mentioned in the entire tweets dataset are considered as the general topics (under the column **General**) because their distributions are fixed for the event (recall Section 4). Note that all of the topics have been manually labeled for identification (e.g. "*Arab Spring*") to reflect our interpretation of their meaning from their top words.

(a) Segmentation of MESpeech

(b) Segmentation of ReaganDebate

Figure 3: Segmentation of the Event

For MESpeech (see Table 2), all specific topics in 7 segments seem to correlate with the event well from a reading of the transcript. Furthermore, it is clear that these topics are sensitive to the event's context and keep evolving as the event progresses in the sense that topics from most segments are different. The only exceptions are "*Human rights*" and "*Foreign policy*", which occur in two segments (S1 and S7). This can be explained by the fact that these two segments serve as the opening and ending of the event. Usually, the content of these two parts tends to be similar since they are either related to the outline or the summarization of the event. On the other hand, general topics and their top words capture the overall themes of the event well. But unlike specific topics, these general topics are repeatedly used across the entire event by the crowd, in particular when expressing their views on the themes of the event (e.g., "*Arab spring*") or even on some popular issues that are neither related nor discussed in the event (e.g., "*Obama*").

For ReaganDebate (see Table 3), we show a sample of 7 (out of 14) segments due to the space limit. All specific topics and their top words from these segments look well related to the event. However, compared to MESpeech where the specific topics were discussed in sequence (except for segments (S1 and S7) which we discussed above), we discover that here the specific topics are rather disordered and occur repeatedly. For example, "*Healthcare*" is mentioned in both segments S3 and S10, and "*Immigration*" is mentioned in segments S6 and S13, etc. This interesting observation is mainly due to the structure of the debate.

Table 2: Top words from topics of MESpeech. Top 2 specific topics per segment (S1–S7). Top 3 general topics from the entire tweet corpus.

S	Specific	Top Words
S1	Human rights	Rights Transition People Power
	Foreign policy	Secure Mideast Arab Clinton State
S2	Terrorism	Bin Laden Mass Murderer Cry
	People	Dignity Power Street Square people
S3	Arab democracy	Democracy Yemen Syrian Bahrain
	Egypt revolution	Mubarak Resign Policy Reform
S4	Youth	Promote Untapped Arab talent youth
	Free speech	Open Internet Mind Access Paper
S5	Economics	Aid Trade Debt Billion Dollar
	Reform	Egypt Reform Support Resource
S6	Border problem	Israel Palestine Borders State Jordan
	Peace treaty	Palestine Peace Jewish Agreed treaty
S7	Human rights	Rights Transition People Power
	Foreign policy	Secure Mideast Arab Clinton State

General	Top Words
Arab spring	Arabia Bahrain Iran Mosques Syrian Leader Government Stepped
Israel & Palestine	Israel Palestine Borders Lines Hamas Negotiate Permanent Occupation
Obama	President Job Tough Critique Jews Policies Attacking Weakness

Note that ReaganDebate is a multi-way conversation. Although it was led by two anchors, sometimes a presidential candidate also expounded his claims and attacked the other candidates' records on some topics, resulting in rebuttals among the candidates. Besides, the event partnered with an online medium (Politico.com) through which readers wrote down their questions to the candidates which were then selected by the anchors. Therefore, common concerns such as "*Healthcare*", "*Economics*", and "*Immigration*" were discussed back and forth heavily throughout the entire debate, producing many more segments than MESpeech (14 vs. 7) and the reoccurrence of the specific topics.

Evolution of Episodic Tweets over the Event's timeline. Next, we study the crowd's responses that are strongly influenced by the event. Specifically, we are interested in how these responses evolve over the event's timeline. Determining whether a response is an episodic tweet depends on its associated preference parameter $c^{(t)}$. As defined in ET-LDA, a response is an episodic tweet only if the sampled probability $P(c^{(t)}) > 0.5$, meaning that the majority of its topics are specific topics, influenced by the content of the segment it refers to. Figure 4 and Figure 5 plot the percentage of those episodic tweets, split by 3 periods of the events. The tweets are presented in buckets, and the percentage of the episodic tweets refers to the proportion in a bucket. Note that the tweets in both figures were ordered by their time.

For MESpeech (see Figure 4), only 18% responses were episodic tweets initially, indicating that most responses at the time were either tangential or about the high-level themes of the event. This is because the responses (first 100 to 200 tweets) were contributed almost as early as 1 day before the event started. Then, a rapid increase of episodic tweets (from 18% to 39%) was witnessed just before the event, suggesting that people had gathered more informa-

Table 3: Top words from topics of MESpeech. Top 2 specific topics per segment. Top 3 general topics from the entire tweet corpus.

Specific		Top Words
S1	Campaign	Tonight Republicans Campaign Leadership
	Candidates	Perry Romney Michele Huntsman Governor
S2	Job market	Job Payroll Market Crisis Monstrous
	Taxes	Income Tax Pledges Taxpayer Committed
S3	Healthcare	Obamacare Wrong Unconstitutional Deal
	Economics	Debt Fence Economics Commitment Cured
S6	Candidates	Perry Romney Michele Huntsman Governor
	Immigration	Legal Mexico Immigrants citizen Solution
S9	Debts	Government Financially Failure China
	Regulations	Fed Up Wrong Funding Expenditures
S10	Social Sec.	Social Security Benefits Ponzi Scheme
	Healthcare	Obamacare Wrong Unconstitutional Deal
S13	Immigration	Legal Mexico Immigrants citizens Solution
	Economics	Debt Fence Economics Commitment Cured

General	Top Words
Social security	Perry Social Security Ponzi scheme Check Constitutional Lowest Wage Vote Wrong
Economics	Private Sector Obama Conservative Budget Amendment Growth Employment Taxes Job
Health Care	Legislative Legal Solution Homelessness
	Obamacare Jeopardizes Medicare Doctor

tion about it. We observe that interesting changes occur both when the event begins and as it is ending. In both cases, the percentage of episodic tweets rises up sharply (beginning: from 39% to 52%; ending: from 43% to 50%) and then drops down quickly. We believe this makes sense since people are often very excited when the event starts and ends. Under such circumstances, they tend to respond strongly to both parts. For example, a large number of the responses like "*Obama starts talking*", "*Here we go, Obama finally showed up*" were witnessed in response to the opening of MESpeech, and responses such as "*Obama's speech was finally over*" were seen mostly from the ending of the event. In fact, the beginning (the ending) part is usually determined by ET-LDA as the first (last) segment. More surprising to us was the fact that the percentage of episodic tweets remained mostly stable during the event. This might be because the most audience members had lower interest levels about specific topics about the Middle East, so their responses tended to be general even as the event was airing.

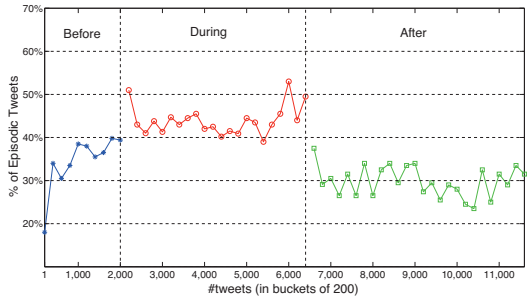


Figure 4: The percentage of episodic tweets to MESpeech over its timeline. Tweets were ordered by their time.

For ReaganDebate (see Figure 5), the graph for the percentage of episodic tweets shows a similar behavior to the one in MESpeech. However, we also discover three key differences through the comparison. First, the responses are much more strongly influenced by the specific topics of the debate when compared to MESpeech, (33% vs. 18% in terms of the lowest percentage). We believe this is because ReaganDebate was about domestic issues that interested more people. Therefore, they tended to follow the debate closely and their responses were more episodic. Second and more interestingly, the crowd was less excited during the opening and ending of the debate. We attribute this to two reasons: (1) MESpeech was significantly delayed by 40 minutes. Therefore, responses were stronger when the event finally began, and (2) before ReaganDebate, there had been 4 Republican Primary debates already, so the crowd might have been less excited at the start. Lastly, we find the percentage of episodic tweets rises significantly during the debate (see the percentage rise around the 66,000th tweet). While looking through the content of the segments that these tweets referred to, we find topics like "*Healthcare*" and "*Economics*" were discussed. We conjecture that, since these topics are controversial and are a strong concern in the Primaries, the responses from the audience were pronounced.

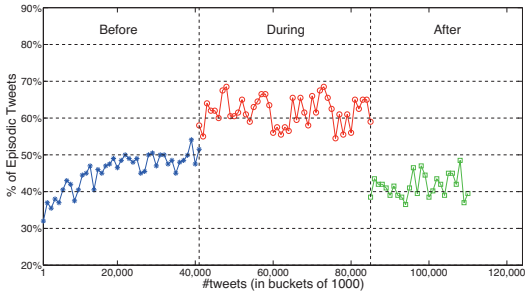
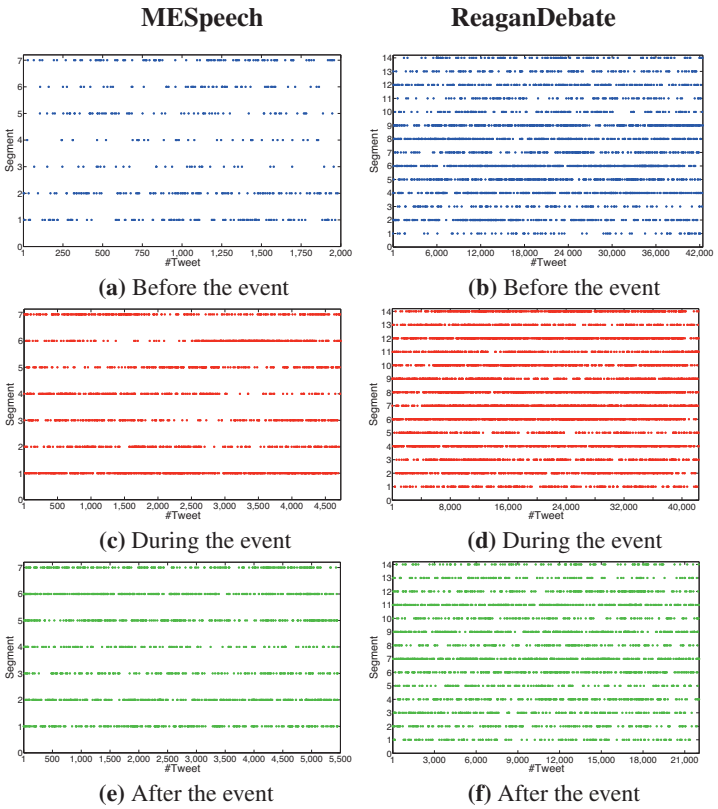


Figure 5: The percentage of episodic tweets to ReaganDebate over its timeline. Tweets were ordered by their time.

Distribution of Segments Referred to by Episodic tweets. We now study how segments in the events were referred to by episodic tweets from the crowd. As defined in ET-LDA, an episodic tweet may refer to any segment of an event based on its associated categorical distribution governed by parameter $s^{(t)}$. We sample the highest probability segment from the distribution and deem it the *referred* segment. Figure 6 plots the results for both events, where each data point denotes a tweet (which is an episodic tweet). Again, all tweets in both figures were ordered by their time.

For MESpeech, Figure 6a shows how segments were referred to before the event started. As expected, the data points to all segments were pretty sparse. Among the segments, Segments 1 and 2 were referred to slightly more by the episodic tweets, since their focused topics (see Table 2) were mostly general (e.g., "*Human rights*") or popular (e.g., "*Terrorism*") so that people could respond specifically without knowing any content of the event. In Figure 6c, the data points seem much denser for all segments. Based on the patterns of the data points in these figures, we make two *key* observations here: (1) Looking horizontally, we find that the

crowd's attention tended to shift from one segment to the next as the event progressed. Our observation is based on the fact that the density of the data points of segments evolved over the event's timeline (see Segments 4-6 in Figure 6c). Initially, a segment was sparse since most people may focus on other segments. Gradually, it becomes dense and stays dense (as more episodic tweets were contributed) during the time that the segment was occurring in the event. Afterwards, the density of the segment turned back to sparse because the audience may have lost interest in these topics. (2) More interestingly, when we look vertically in the graphs, we found the episodic tweets not only refer to the segments whose covered topics had been discussed before or were being discussed currently, but also refer to the segments whose topics were expected to be discussed later on in the event. We believe this is possible as long as the person has a high interest level in these topics. Lastly in Figure 6e, we see the level of overall density of the segments lies between the ones in Figure 6a and Figure 6c. We believe this is because people had gained more information after the event (so they responded more specifically than before the event), but also they lost some interests in the event (so their responses were less specific than during the event).



(a) Before the event **(b)** Before the event

(c) During the event **(d)** During the event

(e) After the event **(f)** After the event

Figure 6: The distribution of referred segments by episodic tweets. Each dot presents a tweet. All tweets were ordered by their posted time.

For ReaganDebate, we observe two major differences from the results in MESpeech. First, there were significantly more episodic tweets regardless of the progress of the event (in Fig. 6b, 6d, 6f, the data points of every segment are much denser than the ones in Fig. 6a, 6c, 6e). Second, nearly all segments drew the crowd's attention (episodic tweets) consistently during and after the event as the segments are continuously dense, as opposed to the ones that have evolved over the timeline of MESpeech (graphically, every line has short periods of high density in Figure 6a, 6c, 6e). We attribute this to the fact that the crowd had a better background in domestic issues and was familiar with the topics covered in the event.

5.3 Subjective Evaluation

To reinforce our quantitative studies, we conducted a user study to evaluate the "goodness" of our proposed method. The quality test involves two parts: (1) evaluating the quality of topics, and (2) the soundness of episodic tweets, both discovered or determined by ET-LDA.

Participants and Method. Participants were 31 graduate students from the engineering school of a university, who were required to follow the news closely and tweet at least three times per week. Median age of participants was 26 years (range: 21-37 years). The procedure of our user study is the following: each participant was presented with a questionnaire, which contained 5 parts: (**i**) 5 samples of segments per event (recall MESpeech has 7 and ReaganDebate has 14 segments), together with short summaries for both events. (**ii**) 5 samples of episodic tweets of each segment. Below each tweet, its top 2 specific topics were listed. (**iii**) 5 samples of steady tweets to the event and its top 2 general topics were listed as well. All topics and segments were generated by ET-LDA during the training time and the ordering of the samples was randomized. For the comparison of the quality of topics and the soundness of episodic tweets, participants were provided with (**iv**) top 2 topics extracted from the episodic tweets (as determined by ET-LDA in advance) using a traditional LDA model trained ($K = 20$) on the tweets corpus only, and (**v**) top 5 tweets per segment measured by the distance (Jensen-Shannon divergence) of their topics to the ones of the referred segment. The JS divergence was calculated as $D_{JS} = \frac{1}{2}D_{KL}(P||R) + \frac{1}{2}D_{KL}(Q||R)$, where $R = (\frac{1}{2}P + \frac{1}{2}Q)$, P is a mixture of topics for tweets and Q is a mixture of topics for the referred segment, both are found by the LDA. Note these tweets are neither episodic nor steady, they are only similar/relevant to the segment of the event. After each sample, the participant was asked to (1) rate the quality of topics, and (2) rate the soundness of episodic tweets as compared to the ones described in (**v**), on a Likert scale of 1 to 5 rating. The duration of the study was 20-30 minutes.

Results. Now, we compare the overall performance of our proposed method against the baseline method (LDA) using the qualitative responses obtained in the user study. In Table 4, we show the measure of the Likert scale for the results of two methods, averaged over the value diversity. We observe that the best ratings are obtained by our proposed method ET-LDA (on an average 18%-41% improvement over the baseline LDA method). Besides, the difference between the methods is more obvious in ReaganDebate

rather than MESpeech, because the crowd was topically influenced by ReaganDebate more (from our observation in Figure 6b, Figure 6d, and Figure 6f) and only our proposed model can characterize such a relationship (while LDA ignores such influences).

Table 4: Performance of methods on the quality of topics (**T**)for each sampled segment (S1-S5) and the soundness of episodic tweets (**ET**) based on Likert scale. The higher values are better.

MESpeech						
		S1	S2	S3	S4	S5
T	ET-LDA	0.51	0.45	0.55	0.62	0.68
	LDA	0.43	0.41	0.47	0.44	0.51
ET	ET-LDA	0.49	0.51	0.56	0.58	0.63
	LDA	0.48	0.49	0.54	0.51	0.57
ReaganDebate						
		S1	S2	S3	S4	S5
T	ET-LDA	0.51	0.61	0.69	0.67	0.68
	LDA	0.48	0.51	0.52	0.54	0.57
ET	ET-LDA	0.51	0.52	0.57	0.62	0.61
	LDA	0.48	0.49	0.51	0.51	0.58

In the light of these observed differences, we study the statistical significance of ET-LDA with respect to LDA. We observe from Table 5 that the comparisons of ET-LDA to LDA yield low p-values, indicating that the improvement in performance of ET-LDA is statistically significant (against significance level of 0.05), particularly for the quality of topics in ReaganDebate. This is in conformity with our observations that ET-LDA outperforms LDA more if there exists a strong influence from the event on the crowd's responses.

Table 5: p-values for LDA against ET-LDA on the quality of topics (**T**) and the soundness of episodic tweets (**ET**)

	MESpeech		**ReaganDebate**	
	T	**ET**	**T**	**ET**
LDA	0.0163	0.0408	0.0092	0.0291

6 Discussions

We now summarize the central findings of this work. The first finding is that the crowd's responses tended to be general and steady before the event and after the event, while during the event, they were more specific and episodic. Such findings confirm our conjectures in Section 3.

Secondly, the crowd showed different levels of engagement in different kinds of events. We attribute this to the reason that people may have greater interest levels about the general topics of certain events (e.g., topics in ReaganDebate). Our final finding is that the topical context of the tweets did not always correlate with the timeline of the event. We have seen that a segment in the event can be referred to by episodic tweets at any time irrespective of whether the segment has already occurred or is occurring currently or will occur later on. This finding is significant in light of the fact that current approaches such as (Shamma, Kennedy, and Churchill 2009) focus on correlating tweets to the event based on their timestamps, however our models enable a richer perspective.

7 Conclusion

We have described a joint statistical model ET-LDA that imposes topical influences between an event and the tweets around it. Depending on such influences, tweets are labeled steady or episodic. Our experimental results also revealed interesting patterns of how users respond to events. Through subjective evaluations on two tweet datasets, our proposed model significantly outperformed the traditional LDA. We believe this paper presents a strong model for understanding complex interactions between events and social media feedback, and reveals a perspective that is useful for tools in event playback and the extraction of a variety of further dimensions such as sentiment and polarity.

References

Bakshy, E.; Hofman, J.; Mason, W.; and Watts, D. 2011. Everyone's an influencer: quantifying influence on twitter. In *WSDM'11*.

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

Chakrabarti, D., and Punera, K. 2011. Event summarization using tweets. In *ICWSM'11*.

Cui, P.; Wang, F.; Liu, S.; Ou, M.; Yang, S.; and Sun, L. 2011. Who should share what? item-level social influence prediction for users and posts ranking. In *SIGIR'11*.

Diakopoulos, N., and Shamma, D. 2010. Characterizing debate performance via aggregated twitter sentiment. In *CHI'10*.

Griffiths, T., and Steyvers, M. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences*.

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'09*.

Hu, Y.; John, A.; and Seligmann, D. 2011. Event analytics via social media. In *SIGMM workshop on Social and Behavioural Networked Media Access (SBNMA'11)*.

Java, A.; Song, X.; Finin, T.; and Tseng, B. 2007. Why we twitter: understanding microblogging usage and communities. In *WebKDD'07*.

Naaman, M.; Boase, J.; and Lai, C. 2010. Is it really about me?: message content in social awareness streams. In *CSCW'10*.

Ramage, D.; Dumais, S.; and Liebling, D. 2010. Characterizing microblogs with topic models. In *ICWSM'10*.

Sahami, M., and Heilman, T. 2006. A web based kernel function for measuring the similarity of short text snippets. In *WWW'06*.

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

Vieweg, S.; Hughes, A.; Starbird, K.; and Palen, L. 2010. Microblogging during two natural hazards events: what twitter may contribute to situational awareness. In *CHI'10*.

Weng, J.; Yao, Y.; Leonardi, E.; and Lee, F. 2011. Event detection in twitter. In *ICWSM'11*.

Zhao, D., and Rosson, M. 2009. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *GROUP'09*.

Zhao, W.; Jiang, J.; Weng, J.; He, J.; Lim, E.; Yan, H.; and Li, X. 2011. Comparing twitter and traditional media using topic models. *ECIR'11*.