Online Spatial Event Forecasting in Microblogs

LIANG ZHAO, George Mason University
FENG CHEN, University of Albany, SUNY
CHANG-TIEN LU and NAREN RAMAKRISHNAN, Virginia Tech

15

Event forecasting from social media data streams has many applications. Existing approaches focus on forecasting temporal events (such as elections and sports) but as yet cannot forecast spatiotemporal events such as civil unrest and influenza outbreaks, which are much more challenging. To achieve spatiotemporal event forecasting, spatial features that evolve with time and their underlying correlations need to be considered and characterized. In this article, we propose novel batch and online approaches for spatiotemporal event forecasting in social media such as Twitter. Our models characterize the underlying development of future events by simultaneously modeling the structural contexts and their spatiotemporal burstiness based on different strategies. Both batch and online-based inference algorithms are developed to optimize the model parameters. Utilizing the trained model, the alignment likelihood of tweet sequences is calculated by dynamic programming. Extensive experimental evaluations on two different domains demonstrate the effectiveness of our proposed approach.

CCS Concepts: • **Computing methodologies** → **Bayesian network models**

Additional Key Words and Phrases: Graphical models, spatiotemporal event forecasting, social media

ACM Reference Format:
Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2016. Online spatial event forecasting in microblogs. ACM Trans. Spatial Algorithms Syst. 2, 4, Article 15 (November 2016), 39 pages.
DOI: http://dx.doi.org/10.1145/2997642

1. INTRODUCTION

Microblogs like Twitter and Weibo are important platforms for ongoing discussions of societal events [Kwak et al. 2010]. As of the end of 2014, 255 million active users were collectively creating 500 million tweets every day, covering a whole variety of content ranging from everyday feelings to comments about social gatherings [Bennett 2014]. Compared to traditional media, Twitter has the following significant characteristics: (i) *Timeliness of messages:* Unlike traditional media, which may take hours or days to publish information, tweets can be posted instantly utilizing portable mobile devices in users' pockets; 2 = (ii) *ubiquity of social sensors:* tweets reflect the public's mood and trends, both of which are potential determinants of future social events; and (iii) *availability of geo-information:* Twitter users provide rich location information within their profiles, texts, and geotags. Recent research has revealed the power of Twitter for event forecasting [Tumasjan et al. 2010; Wang et al. 2012b]; Twitter and other social media have been recognized as playing a key role in events

Authors' addresses: L. Zhao, 12566 Summit Manor Drive # 224, Fairfax, VA 22033; email: lzhao9@gmu.edu; F. Chen, LI-96J, 1400 Washington Avenue, Albany, NY 12222; email: fchen5@albany.edu; C.-T. Lu, 7054 Haycock Road, Falls Church, VA 22043; email: ctlu@vt.edu; N. Ramakrishnan, 7054 Haycock Road, Falls Church, VA 22043; email: naren@cs.vt.edu.

DOI: http://dx.doi.org/10.1145/2997642

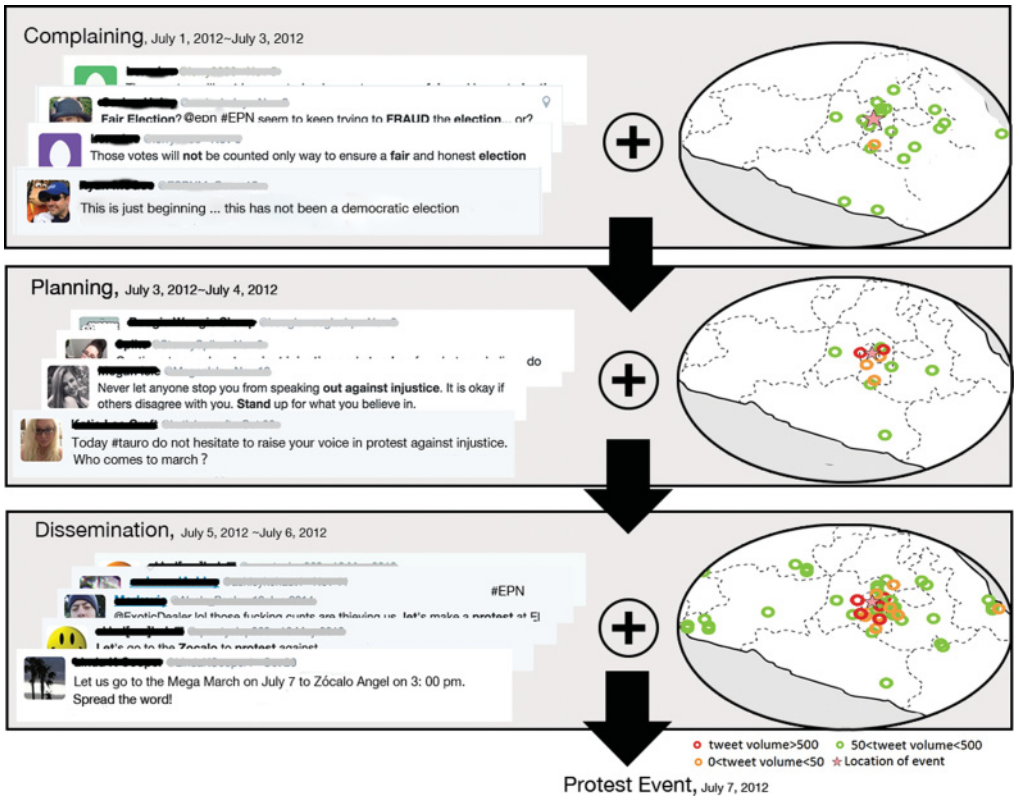


Fig. 1. Twitter predicts a presidential election protest.

such as the "Arab Spring" and the protests surrounding the Mexican presidential election [Ramakrishnan et al. 2014; Wang et al. 2012b]. Figure 1 depicts the activities on Twitter that causally preceded the Mexico City protests. Both the content and the spatiotemporal burstiness of these protest-related tweets reveal the escalation of societal discontent pertaining to this controversial election, with users progressing from complaining through planning and advertising to participating in the final protest event. However, existing event forecasting models in Twitter generally focus on temporal events whose geo-locations are not available or are not considered in the prediction task (e.g., elections [Tumasjan et al. 2010] and sports [Pavlyshenko 2013]). Comparatively little attention has been paid to forecasting spatiotemporal events.

Tweets posted within a certain geographical neighborhood could be able to reflect important spatiotemporal patterns of social event [Zhao et al. 2014]. Thus, the forecasting of spatiotemporal events requires a consideration of spatial features and their correlations in addition to the temporal dimension. This poses the following three challenges: (i) *Capturing spatiotemporal dependencies.* A spatial event may be delineated by not only a specific location and time, but also its geographical and temporal neighborhood. The strength and pattern of the resulting tweets may also vary differently for different development stages for different events. (ii) *Modeling mixed type observations.* When an event occurs, this will involve the temporal evolution of spatially distributed tweets describing the event and its semantics. The joint consideration of these heterogeneous and multidimensional data points is thus crucial. And (iii), *utilizing prior geographical knowledge.* Spatiotemporal events in crucial domains usually have rich historical

records. Different geo-locations may tend to feature their own inherent and distinct event frequencies that can be integrated into a predictive model to improve the forecasting accuracy. For example, the historical crime rates in different areas of a city can help forecast the probability of future crimes occurring in those areas.

This article proposes a new approach to developing spatiotemporal event forecasting models that addresses the above-mentioned issues more effectively. The proposed methodology generatively characterizes the evolutionary development of events, as well as the relationships between the tweet observations both inside and outside the event venue. To uncover the underlying event development mechanics, this approach jointly considers the structural semantics and spatial-temporal burstiness patterns in Twitter streams. Utilizing geographical priors allows the spatial burstiness distributions to be learned for specific corresponding locations, while applying a Gaussian-inverse Wishart prior distribution facilitates event forecasting for unknown locations. The main contributions of this article are:

—**A novel generative framework for spatial event forecasting.** For spatial event forecasting in Twitter, we propose an enhanced Hidden Markov Model (HMM) that characterizes the transitional process of event development by jointly considering the time-evolving context and space-time burstiness of Twitter streams.
—**Effective batch and online algorithms for model parameter inference.** The model inference is formalized as the maximization of a posterior that is analytically tractable. Both Expectation Maximization (EM)-based and stochastic-EM parameter optimization algorithms are proposed to solve this problem effectively and efficiently.
—**A new sequence likelihood calculation method.** To handle the noisy nature of tweet content, words that are exclusive to a single event are identified by a language model that has been optimized by a dynamic programming algorithm to achieve an accurate sequence likelihood calculation.
—**Extensive experimental performance evaluations.** The proposed method outperforms existing methods by 38% and 67% on two different real-world datasets. Sensitivity analyses reveal the impact of the parameters on the new method's performance. Case studies on both datasets are illustrated and elaborated to demonstrate the practical utility of the proposed methods.

The rest of this article is organized as follows. Section 2 reviews existing work in this area, after which Section 3 describes the proposed generative model, Section 4 provides details of the associated parameter estimation, and Section 5 explains the event forecasting function of the proposed model. In Section 6, extensive experiments to evaluate the performance of the new model are conducted and analyzed; the work is summarized and conclusions drawn in Section 7.

2. RELATED WORK

Current research into the analysis of Twitter-based social events can be categorized into two main types: (i) event detection and (ii) event forecasting. These are considered in turn here.

Event detection: A large body of work focuses on the detection of ongoing events [Aggarwal and Subbian 2012; Lappas et al. 2012; Sakaki et al. 2010; Signorini et al. 2011; Weng and Lee 2011]. These papers treat tweets as real-time social sensors to promptly discover new events as they occur. Methods based on spatial bursts use a classifier to extract topic-related tweets and then examine their spatial burstiness and have been tested for applications such as detecting earthquakes [Sakaki et al. 2010] and disease outbreaks [Signorini et al. 2011], while methods based on temporal bursts detect temporal patterns in Twitter streams utilizing techniques such as wavelet analysis [Weng and Lee 2011], temporal clustering [Aggarwal and Subbian 2012], and

query expansion [Zhao et al. 2015a; Jin et al. 2014]. Existing methods developed for flu detection are typically focused on the temporal dimension. For example, Ginsberg et al. proposed monitoring ongoing flu activity using Google search engine data. In contrast, spatiotemporal methods aim to detect bursts in both time and space [Ginsberg et al. 2009]. However, these event detection approaches can only uncover events after they have occurred and are unable to forecast future events because they focus on observations that directly reflect currently occurring events rather than precursor indicators that reveal the causes or development of future events.

Event forecasting: Most research in this area focuses on temporal events and ignores the underlying geographical information that is also available in tweets. A variety of applications have been explored, including predicting election outcomes [O'Connor et al. 2010; Tumasjan et al. 2010], disease outbreaks [Achrekar et al. 2011; Ritterman et al. 2009; Zhao et al. 2016], stock market movements [Arias et al. 2013; Bollen et al. 2011], politics [Marchetti-Bowick and Chambers 2012], box office ticket sales [Arias et al. 2013], the Olympic games [Pavlyshenko 2013], crime [Wang et al. 2012b], and traffic conditions [He et al. 2013]. These papers can be categorized into four types based on the complexity of the models utilized: **(i) Linear regression models.** These methods map simple predictive features such as sentiment score or tweet volume to the occurrence of future events [Arias et al. 2013; Bollen et al. 2011; He et al. 2013; O'Connor et al. 2010]. **(ii) Nonlinear models.** These methods incorporate more informative features such as semantic topics by utilizing methods such as support vector machines and logistic regression [Ritterman et al. 2009; Wang et al. 2012b]. **(iii) Time series-based methods.** These methods consider the temporal correlation of relevant features such as tweet volume by adopting approaches such as autoregressive modeling. For example, Achrekar et al. [2011] utilize an autoregression with exogenous input (ARX) model to forecast flu activity over the next few days. And **(iv) domain-specific approaches.** These methods are designed to solve particular problems and may not be applicable to other application domains. For example, Pavlyshenko [2013] applied an association rule approach to discover the most frequently mentioned players and hence predict the results of sports tournaments, while Marchetti-Bowick and Chambers [2012] focused on improving the performance of sentiment analysis related to political events. As yet, there have been few reports of work specifically on spatiotemporal event forecasting. Gerber [2014] proposed a predictor for spatiotemporal events that utilized historical event counts and topics but did not consider temporal evolution and dependencies, while Wang et al. [2012a] developed a model to characterize and predict spatiotemporal criminal incidents, although their model requires the availability of demographic information. Zhao et al. [2015b] proposed three multitask learning models to forecast civil unrest events utilizing static features and dynamic features. Instead of considering geographic neighborhoods, these models assume all the locations in a country interact equally with each other.

This article proposes a spatiotemporal event forecasting method that is capable of characterizing the evolutionary pattern of both spatial burstiness and structural contexts. By modeling geographical priors more effectively, the new approach proposed here can sufficiently leverage historical prior knowledge for it to be applied to new locations.

3. GENERATIVE PROCESS OF THE PROPOSED MODELS

This section describes the formulation and generative process of the proposed methods. First, the spatiotemporal event forecasting problem is formalized; then our new generative model is described in detail, including the space-time burstiness and structural tweet content modules.

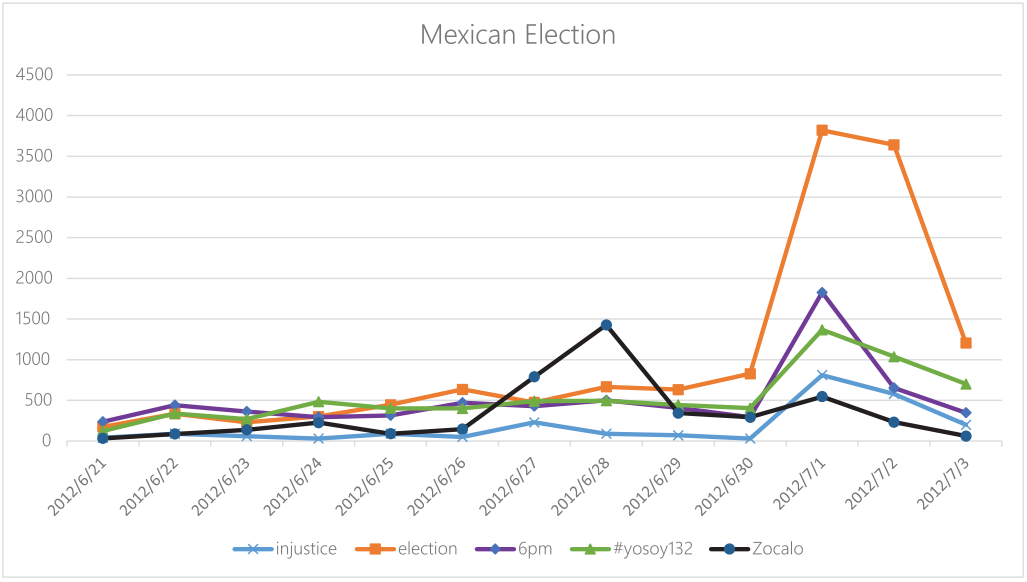


Fig. 2. Keyword tendencies in Twitter during the development of a protest against the presidential election on June 30, 2012, in Mexico.

3.1. Problem Formulation

Indicators for the development of events. Social media has been widely recognized as social sensors and used to detect social events. Beyond that, recently, some researchers utilize microblogs to characterize and track event progress, such as sports [Chakrabarti and Punera 2011] and finance events [Chua and Asur 2013]. More specifically, both the text content and spatial information of social media are regarded as effective social indicators. To see this, some insights on civil unrest and disease outbreak events are provided in the following.

Figure 2 shows the keyword counts during the development of a protest against the presidential election in Mexico, in 2012. The election date was on July 1, which is confirmed by the spike in the curve of the word "election." A protest against it was planned to occur on June 30 and had been planned for a while in social media before it occurred. This can be seen from the slowly increased count of tweets containing the hashtag "#yosoy132," which denotes the name of the organization behind the protest. Another keyword "Zocalo" denotes a plaza in Mexico City, which is the planned location for the protest. And the protesters disseminated this information mainly on June 27, 28, and 29, as revealed by the peak of "Zocalo" in the figure. The word "injustice" shows the people's complaints being expressed before the protest and during the election.

Figure 3 illustrates the keyword trends before and after a protest on June 6 against an increase in the price of metro fares in Sao Paulo. The complaints on metro fare prices had been observed for a while in social media before the protest, as shown in several small outbreaks of the curve "metro fare" in the figure. A protest against the price increase was planned for "Paulista Avenue," a main street in Sao Paulo. This keyword has a spike around June 2, before the protest, when the protest organization was calling for action. Finally, the spike of the keyword "protest" reveals the protest date and the news reporting this afterward.

As shown in Figure 4, there was a large protest on November 9 in Venezuela against its government on some social issues. The development of the protest can be clearly

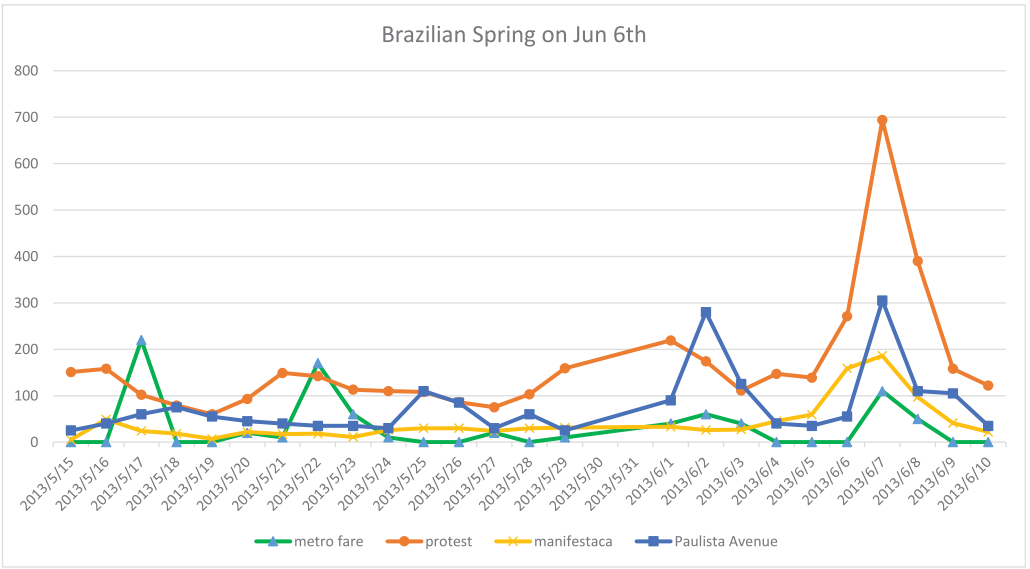


Fig. 3. Keyword tendencies in Twitter during the development of a protest on June 6 during the movement named "Brazilian Spring, 2013" against the government of Brazil.

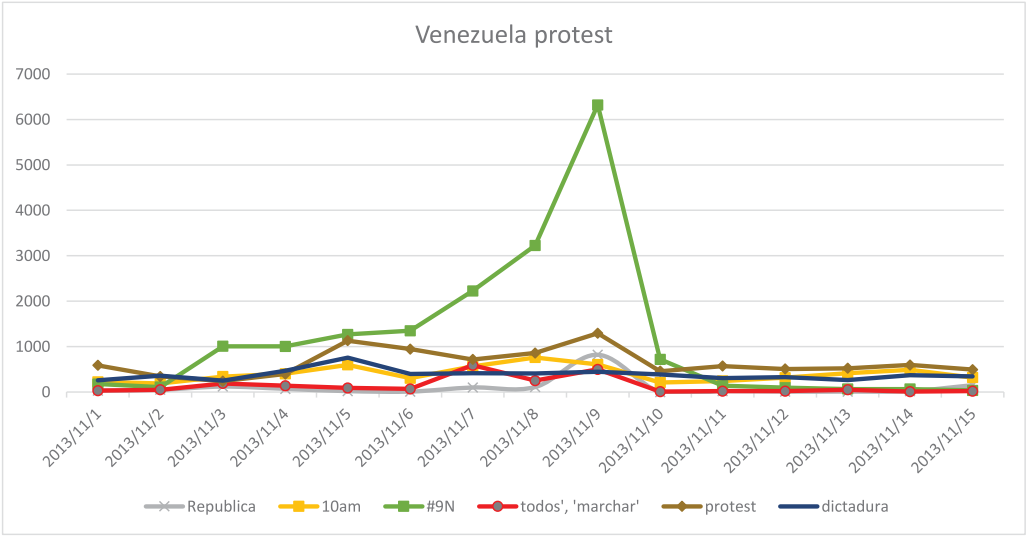


Fig. 4. Keyword tendencies in Twitter during the development of a protest on November 9, 2013, in Venezuela.

seen from the keyword "#9N," which is the name of this protest. It was derived from "9th Nov" which is the date of this protest. The count of tweets containing "#9N" increased from nearly zero to about 1,000 on November 3, more than 6,000 on November 9, and dropped dramatically to below 200 after November 11. The small outbreaks of words like "dictator" revealed the complaints of the people before the protest. The keyword "todos marchar" is a Spanish phrase that means "let us protest" in English, showing that people were calling for protest in social media from November 7. This can also

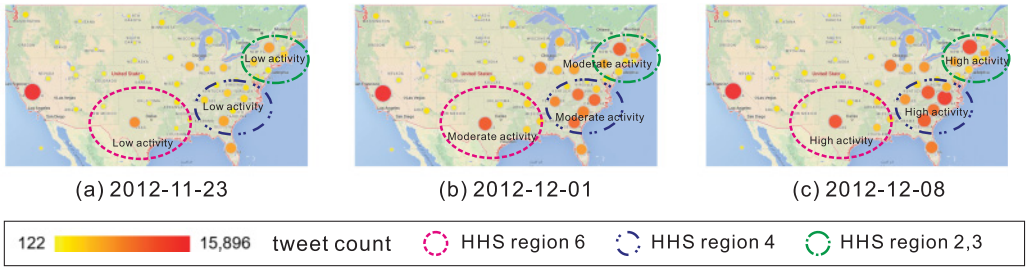


Fig. 5. Spatiotemporal outbreaks of topic-related tweets during the development of influenza epidemics in December 2012 in the United States.

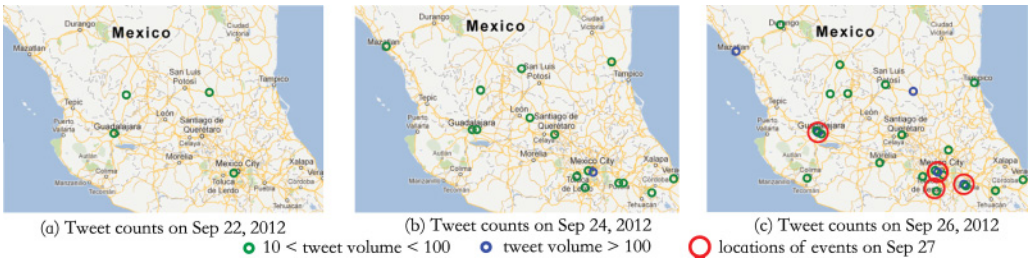


Fig. 6. Spatiotemporal outbreaks of topic-related tweets during the development of civil unrest events in September 2012 in Mexico.

be shown by their poster in the figure. The time "10am" was the planned time for the protest, which also shows the protesters' efforts to organize the protest before its occurrence.

The development of an event can also be indicated by the spatial outbreaks of social media postings. As shown in Figure 5, there were influenza outbreaks in several states, including Texas, North Carolina, and Massachusetts, in the United States in the week around December 15, 2012. These outbreaks had been developing for several weeks before the outbreaks, as shown in the snapshots in Figure 5(a), 5(b), and 5(c). The flu-related tweet counts were increasing within these states. In addition, the counts of the neighbor states (e.g., the other states within the same HHS region[1]) were also increasing, showing the spread of the disease across neighboring states. Similar increases did not happen in the distant states in the central United States, such as Missouri. Similar to Figure 5, Figure 6 illustrates the spatial outbreaks of tweets on several civil unrest events that occurred on September 27, 2012. On September 22, there were a few tweets in cities like Mexico City and Guadalajara. Then the volume of tweets increased in several cities, especially in the locations of the future protests on September 24. On September 26, the spatial outbreaks became more obvious in the cities like Guadalajara, Mexico City, and Puebla. These outbreaks match the civil unrest events that occurred on the next day, namely September 27.

As shown in these figures, there is a development process for events such as influenza epidemics and civil unrest that can be viewed as a Markov process, which is a chain consisting of a sequence of event stages. Although these stages are hidden and cannot be directly observed, they could be identified indirectly by social sensors such as social media. This notion leads to a hidden Markov process where, for example, the hidden states are the underlying stages of an event, while the observations for the underlying

[1]HHS regions are the groups of geo-neighboring states defined by Department of Health and Human Services for epidemic disease prevention. http://www.hhs.gov/about/agencies/regional-offices/.

Table I. Notations and Descriptions

Notations	Descriptions
$Z_{s,t}$	Latent state in sequence s at time t.
$Y_{s,t,n}$	Category-switching variable of the nth word in sequence s at time t.
$X_{s,t,n}$	Topic label of the nth word at time t in sequence s.
$W_{s,t,n}$	The nth word in sequence s at time t.
$r_{s,t}^{in}$	The posting ratio in sequence s's location at time t.
$r_{s,t}^{out}$	The posting ratio outside the location of sequence s at time t.
$N_{s,t,w}$	The frequency of a word w in sequence s at time t.
Ψ	Bernoulli distribution that generates $Y_{s,t,n}$.
Φ	Topic distribution that generates $X_{s,t,n}$.
θ_j^B	Distribution of words under the jth topic.
$\theta s, t^R$	Distribution of words exclusive to sequence s at time step t.
$\mu_{l,k}$	Mean of posting ratios of location l under latent state k.
$\Sigma_{l,k}$	Covariance of posting ratios of location l under latent state k.
$\lambda_{l,k}^{in}$	Mean of posting ratios inside the location l for Poisson distribution for latent state k.
$\lambda_{l,k}^{out}$	Mean of posting ratios outside the location l for Poisson distribution for latent state k.

stages are the postings in social media. More formally, in the following, the detailed formulation is described.

The notation used in this article is introduced in Table I. As demonstrated in Figure 1, in order to accurately forecast spatiotemporal events it is crucial to be able to characterize their underlying development before their occurrence by utilizing relevant tweet observations. An enhanced HMM is proposed here to characterize the underlying development of events.

Given a sequence of observations O, a standard HMM can be denoted as a quadruple (H, Z, A, π), where Z is a set of K latent states. $H_k(O_i)$ denotes the emission probability that a symbol O_i is generated by the kth latent state. A is a $K \times K$ transition probability matrix, where $A_{j,k} = p(Z_j|Z_k)$ is the transitional probability of moving from the jth latent state to the kth latent state, and π is the initial probability vector, where π_k is the probability that the initial state is k. Starting from an initial state k, the HMM generates an observation O_1 according to the emission probability $H_k(O_1)$, and then transitions to a state j with the transitional probability $A_{j,k}$. The training process for an HMM thus entails searching for the set of parameters (H, Z, A, π) that best fits the sequence of observations.

However, a standard HMM is limited to simple symbol observations and will thus face several challenges in our case because the observation does not consist of a single symbol but rather of all the domain-related tweets in each time step. Furthermore, a standard HMM can neither characterize spatial burstiness nor handle structural and noisy observations. Here, both the content and the spatial burstiness of domain-related tweets are the observations, and the underlying stage in the development of social events is characterized as the latent state. A future event is predicted by inferring the underlying development based on tweet observations.

This problem therefore requires several important enhancements to the standard HMM. First, instead of a single symbol, each observation must encompass all the domain-related tweets in each time step. Second, the enhanced HMM treats the spatial burstiness of domain-related tweets as multivariate "posting rates" in the same geographical neighborhood. Third, to address the noisy nature of tweet content, a language model is used to filter out typos and identify proper names that are exclusive to particular events. Fourth, the structural semantics of the filtered tweets is modeled as a mixture of latent topics. The generative process of the new model is described in the following subsections.

Fig. 7. Tweet context evolution during the development of a protest.

More formally, denote $D = \{D_{l,t}\}_{l \in \mathcal{L}, t \in \mathcal{T}}$ as a collection of space-time-indexed Twitter data split into different geographical locations $\mathcal{L}$ and different time intervals $\mathcal{T}$. A sequence of tweets is defined as $s = \{D_{l,t}\}_{t \in T \subseteq \mathcal{T}}$, which contains all the tweets in location l during the time period $T \subseteq \mathcal{T}$. S denotes the number of all such sequences in the data D. Our model characterizes the development of each event as a sequence of latent states $Z = \{1, 2, \ldots, K\}$, with tweet sequence $s \subseteq D_l$ being the observations generated by the latent states.

Structural tweet content modeling. The underlying development of an event is not only reflected by the evolutionary content in tweet texts, but also by the spatial count distribution of event-related tweets. The characterizations of the tweet content and the spatial counts are described in this and the next sections, respectively.

Figure 7 illustrates the temporal evolution of representative tweet content during the development of a protest against the 2012 presidential election in Mexico. It shows how there are generally two categories of words; namely, (i) event-specific words that are specific to a unique event, such as hashtags, hyperlinks, landmarks, and organization names, shown in purple in Figure 7; and (ii) common words that can be common to a number of different events. These mainly belong to two different threads. A thread of common words will contain the background words, which are commonly used but less informative, such as stop-words. Another thread contains the common words that are topical, such as those shown in green, red, and yellow, which encompass information about the action that will be taken during the event development. As shown in Figure 7, unlike event-specific words, topical words are not restricted to specific events and are thus able to indicate the development stages of different events. In the kth latent state, the probability that a word belongs to either of the above two categories is modeled by a Bernoulli distribution:

$$Y_{s,t,n} \sim Bern(Y_{s,t,n}|\Psi_k). \tag{1}$$

If a word $W_{s,t,n}$ in sequence s at time step t belongs to the first category, it is directly generated from a language model $\theta^R_{s,t}$, which identifies the words exclusive to the current event sequence s at current time t:

$$W_{s,t,n} \sim Mult\left(W_{s,t,n}|\theta^R_{s,t}\right). \tag{2}$$

If the word belongs to the second category, then it is selected from one of the latent topics that are shared by all such events.

$$X_{s,t,n} \sim Mult(X_{s,t,n}|\Phi_k). \tag{3}$$

A latent topic j is modeled as a multinomial distribution over words:

$$W_{s,t,n} \sim Mult\left(W_{s,t,n}|\theta^{B_j}\right). \tag{4}$$

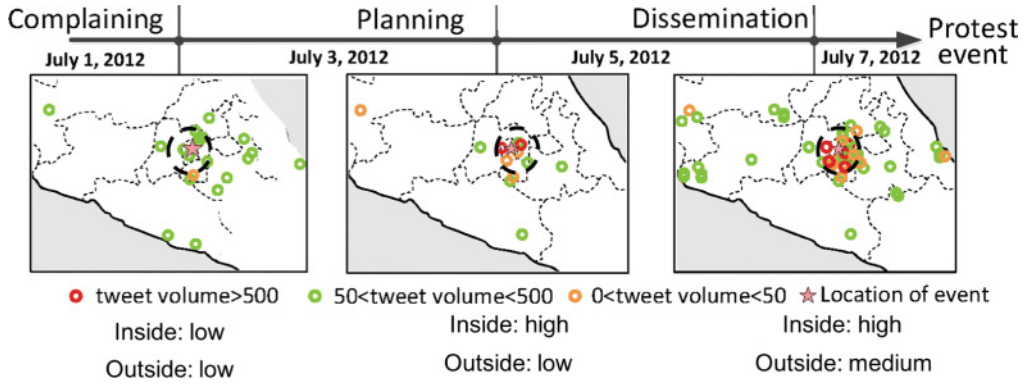


Fig. 8. The evolution of tweet space-time burstiness during the development of a protest.

3.2. Model I: Space-Time Burstiness Modeling with Neighborhood Interactions (STM-I)

The underlying progression of an event's development is reflected by the evolutionary counts of event-related tweets in spatial regions. As shown in Figure 8, on July 1, 2012, when the results of the Mexican presidential election were announced, not only residents of Mexico City, but also people from other regions in the country complained about the so-called fraudulent election. By July 4, people, especially those living in and around Mexico City, started to plan potential protests, leading to an obvious tweet count outbreak inside the city. Finally, the planned protests were advertised to those living in other regions in the country, as shown by a series of outbreaks inside Mexico City. The subsequent nontrivial volume of tweets outside the city revealed that responses were being received to these advertisement from other regions in the country.

The outbreaks or anomalousness of spatial tweet counts in location l are typically characterized by their score functions. To perform spatial outbreak modeling and detection, specialized "spatial scan statistics" have been developed [Kulldorff 1997] that are widely used to model these outbreaks, such as disease outbreaks and bioterrorist attacks [Neill 2012]. Spatial scan statistics characterize the patterns of spatial outbreaks by the counts (e.g., the tweet count) or rates (e.g., the ratio of the disease population) inside and outside the region [Neill 2012]. For example, spatial scan statistics attempt to designate spatial disease outbreaks as occurring in those regions where the underlying disease rates are significantly higher than in other regions. Although the specific forms of the score function vary, the tweet counts (or rates) inside and outside the regions are commonly deemed to follow probabilistic distributions such as Gaussian and Poisson distributions.

Gaussian distributions are commonly used in spatial scan statistics to model the inside and outside counts/rates. In this article, we adopt a bivariate Gaussian because its advantages are twofold. First, its covariance matrix quantifies the different significance of the inside and outside ratios in characterizing the spatial burstiness. Second, the nondiagonal elements of the covariance matrix can also capture the relationship between the inside and outside ratios. The detailed formulation of our bi-Gaussian-based model is as follows.

Given a tweet sequence $s \subseteq D_l$ in location l, denote $c^{in}_{s,t}$ as the count of domain-related tweets inside location l at time t and $c^{out}_{s,t}$ as the count outside this location. Denote $b^{in}_{s,t} = |D_{l,t}|$ as the total tweet count inside location l at time step t and $b^{out}_{s,t}$ as that outside this location. $r^{in}_{s,t} = c^{in}_{s,t}/b^{in}_{s,t}$ and $r^{out}_{s,t} = c^{out}_{s,t}/b^{out}_{s,t}$ are the *inside ratio* and the *outside ratio* and are, respectively, the proportions of the domain-related tweets inside and outside location l. Hence, the spatial burstiness pattern surrounding the location

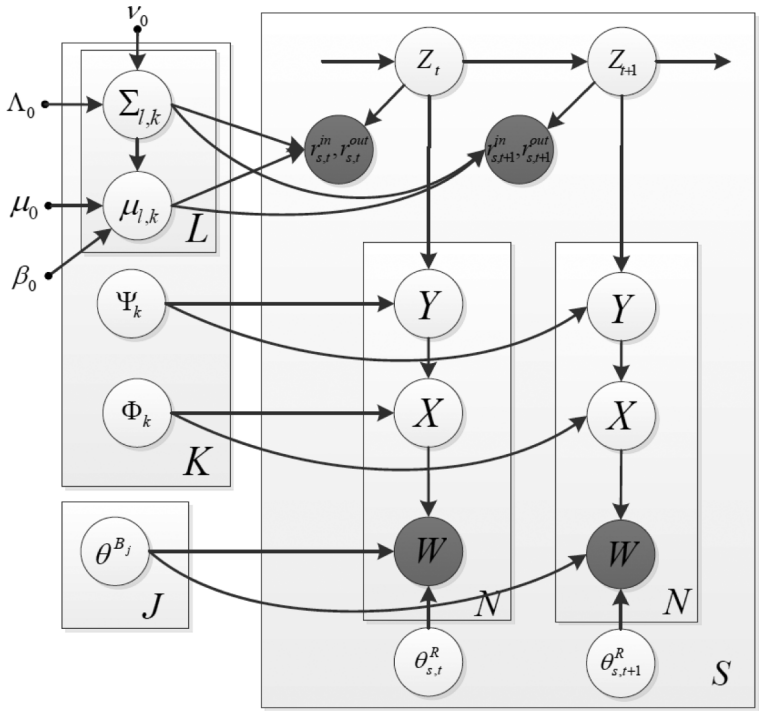


Fig. 9. Plate notation of the proposed STM-I model.

l is jointly characterized by $r^{in}_{s,t}$ and $r^{out}_{s,t}$; spatial burstiness typically occurs when the inside ratio is higher than the outside one. To characterize the spatial burstiness in terms of the inside and outside ratios, a bivariate Gaussian is utilized:

$$r^{in}_{s,t}, r^{out}_{s,t} \sim \mathcal{N}\left(r^{in}_{s,t}, r^{out}_{s,t} | \mu_{l,k}, \Sigma_{l,k}\right). \tag{5}$$

For the kth latent state, draw the mean of the inside and outside ratios $\mu_{l,k}$ from a Gaussian distribution:

$$\mu_{l,k} \sim \mathcal{N}(\mu_{l,k} | \mu_0, \Sigma_{l,k}\, \beta_0), \tag{6}$$

where μ_0 is the historical prior mean of the inside and outside ratios and β_0 is the number of prior measurements. $\Sigma_{l,k}$ is the scale matrix following the inverse Wishart distribution:

$$\Sigma_{l,k} \sim \mathcal{IW}(\Sigma_{l,k} | \Lambda_0^{-1}, \nu_0), \tag{7}$$

where Λ_0 and ν_0 describe the prior scale matrix and the degree of freedom, respectively.

As shown in Figure 9, the generative process of the proposed STM-I, which is the Gaussian-distributed burstiness modeling, is:

—For each sequence s at each time step t,
  —Draw $Z_{s,t} \sim Multi(Z_{s,t} | Z_{s,t-1}, A)$
—For each latent state k in each location l,
  —Draw the mean of the spatial burstiness from a normal distribution $\mu_{l,k} \sim \mathcal{N}(\mu_{l,k} | \mu_0, \Sigma_{l,k}\, \beta_0)$
  —Draw the regional variance from an inverse Wishart distribution $\Sigma_{l,k} \sim \mathcal{IW}(\Sigma_{l,k} | \Lambda_0^{-1}, \mu_0)$
  —For each sequence of tweets s
    —Draw $r^{in}_{s,t}, r^{out}_{s,t} \sim \mathcal{N}(r^{in}_{s,t}, r^{out}_{s,t} | \mu_{l,k}, \Sigma_{l,k})$

—For each word W_n in time step t in tweet sequence s,
—Draw $Y_{s,t,n} \sim Bern(Y_{s,t,n}|\Psi_k)$
—If $Y_{s,t,n} = 0$, draw $W_{s,t,n} \sim Mult(W_{s,t,n}|\theta^R_{s,t})$
—else
—Draw a topic $X_{s,t,n} \sim Mult(X_{s,t,n}|\Phi_k)$.
—Draw a word $W_{s,t,n} \sim Mult(W_{s,t,n}|\theta^{B_j}, j = X_{s,t,n})$.

3.3. Model II: Space-Time Burstiness with Non-negative-Discrete Signals

The model in the previous section, STM-I, characterizes the space-time burstiness by not only considering the signal strengths at different locations, but also the potential correlations among these locations by leveraging the covariance. In spite of these advantages, however, important signals such as the volume of tweets are typically non-negative and discrete and are thus not naturally handled by a Gaussian distribution. To address this issue, in addition to the Gaussian-based version just described, we propose another model, STM-S, to preserve these non-negative and discrete properties based on a Poisson distribution. Specifically, the counts $c^{in}_{s,t}$ and $c^{out}_{s,t}$ are assumed to be Poisson distributed and designated as follows:

$$\begin{aligned} c^{in}_{s,t} &\sim Poisson\big(c^{in}_{s,t}|\lambda^{in}_{k,l} \cdot b^{in}_{s,t}\big), \\ c^{out}_{s,t} &\sim Poisson\big(c^{out}_{s,t}|\lambda^{out}_{k,l} \cdot b^{out}_{s,t}\big), \end{aligned} \tag{8}$$

where $b^{in}_{s,t}$ and $b^{out}_{s,t}$, as previously, represent the total tweet count inside and outside location l at time step t, respectively, and $\lambda^{in}_{k,l}$ and $\lambda^{out}_{k,l}$ denote the means of the inside and outside outbreak ratios, respectively.

Prior knowledge of sufficient statistics to permit a Poisson distribution for different locations and different states follows Gamma distributions:

$$\begin{aligned} \lambda^{in}_{k,l} &\sim Gamma\big(\lambda^{in}_{k,l}|\alpha^{in}, \beta^{in}\big), \\ \lambda^{out}_{k,l} &\sim Gamma\big(\lambda^{out}_{k,l}|\alpha^{out}, \beta^{out}\big), \end{aligned} \tag{9}$$

where α^{in} and β^{in} denote the shape parameter and inverse scale parameter of the Gamma prior for the inside outbreaks distribution. α^{out} and β^{out} denote the shape parameter and inverse scale parameter of the Gamma prior for the outside outbreaks distribution.

As shown in Figure 10, the generative process of the proposed STM-S based on Poisson-distributed burstiness modeling, is:

—For each sequence s at each time step t,
—Draw $Z_{s,t} \sim Multi(Z_{s,t}|Z_{s,t-1}, A)$
—For each latent state k in each location l,
—Draw the mean of the in-location burstiness from a Gamma distribution $\lambda^{in}_{k,l} \sim Gamma(\lambda^{in}_{k,l}|\alpha^{in}, \beta^{in})$
—Draw the mean of the out-location burstiness from a Gamma distribution $\lambda^{out}_{k,l} \sim Gamma(\lambda^{out}_{k,l}|\alpha^{out}, \beta^{out})$
—For each sequence of tweets s
—Draw $c^{in}_{s,t} \sim Poisson(c^{in}_{s,t}|\lambda^{in}_{k,l} \cdot b^{in}_{s,t})$
—Draw $c^{out}_{s,t} \sim Poisson(c^{out}_{s,t}|\lambda^{out}_{k,l} \cdot b^{out}_{s,t})$
—For each word W_n in time step t in tweet sequence s,
—Draw $Y_{s,t,n} \sim Bern(Y_{s,t,n}|\Psi_k)$
—If $Y_{s,t,n} = 0$, draw $W_{s,t,n} \sim Mult(W_{s,t,n}|\theta^R_{s,t})$
—else
—Draw a topic $X_{s,t,n} \sim Mult(X_{s,t,n}|\Phi_k)$.
—Draw a word $W_{s,t,n} \sim Mult(W_{s,t,n}|\theta^{B_j}, j = X_{s,t,n})$.

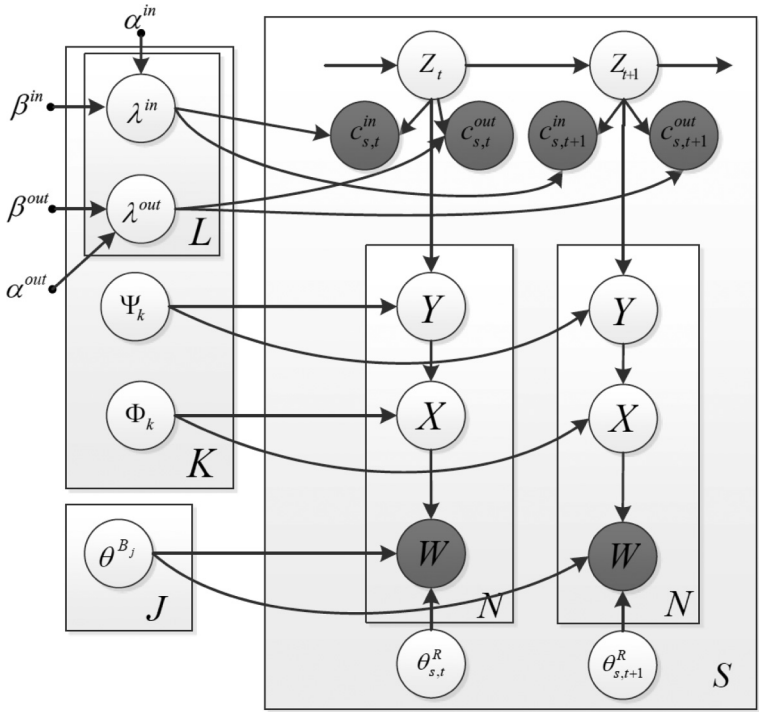


Fig. 10. Plate notation of the proposed STM-S model.

4. PARAMETER ESTIMATION

4.1. Joint Likelihood

Based on the generative process just described, our proposed new model STM-I defines the joint probability of the generation of observed variables, latent variables, and model parameters.

Specifically, the observed variables are the spatial burstiness r^{in}, r^{out}, and words W in the tweet content; the latent variables are topic assignment X, category assignment Y, and latent state assignment Z. The geographical prior is $\Theta_I = \{\mu_0, \beta_0, \Lambda_0, \nu_0\}$. Their joint distribution is expressed as follows:

$$
\begin{aligned}
&p(W, X, Y, Z, \mu, \Sigma, r^{in}, r^{out}|\pi, A, \Psi, \Phi, \theta, \Theta_I) \qquad (10)\\
&= \prod_{s}^{S} p\left(Z_{s,1}|\pi\right) \cdot \prod_{s}^{S}\prod_{t=2}^{T} p\left(Z_{s,t}|Z_{s,t-1}, A\right)\\
&\cdot \prod_{s}^{S}\prod_{t=1}^{T}\prod_{n}^{N} p\left(W_{s,t,n}, Y_{s,t,n}, X_{s,t,n}|Z_{s,t}, \Psi, \Phi, \theta\right)\\
&\cdot \prod_{s}^{S}\prod_{t=1}^{T} p\left(r^{in}_{s,t}, r^{out}_{s,t}|\mu_l, \Sigma_l, Z_{s,t}\right) p(\mu_l, \Sigma_l|\Theta_I),
\end{aligned}
$$

where $\theta = \{\theta^B, \theta^R\}$. Thus, searching for the best settings for the model parameters for STM-I is equivalent to maximizing the logarithm of the joint distribution in

Equation (10). The specific optimization process of this objective function is listed in Equations (18–22) and Equations (29–36) in Appendix A.1.

In STM-S, the Poisson-distributed space-time burstiness modeling approach means that the observed variables are the inside domain-related tweet counts c^{in}, the inside base counts b^{in}, the outside domain-related tweet counts c^{out}, the outside base counts b^{out}, and the words W in the tweet content; the latent variables are once again topic assignment X, category assignment Y, and latent state assignment Z. The geographical prior is $\Theta_{II} = \{\alpha^{in}, \beta^{in}, \alpha^{out}, \beta^{out}\}$. The joint distribution is expressed as follows:

$$
\begin{aligned}
&p(W, X, Y, Z, \mu, \Sigma, r^{in}, r^{out}|\pi, A, \Psi, \Phi, \theta, \Theta_{IIch}) \qquad (11)\\
&= \prod_{s}^{S} p\left(Z_{s,1}|\pi\right) \cdot \prod_{s}^{S}\prod_{t=2}^{T} p\left(Z_{s,t}|Z_{s,t-1}, A\right)\\
&\cdot \prod_{s}^{S}\prod_{t=1}^{T}\prod_{n}^{N} p\left(W_{s,t,n}, Y_{s,t,n}, X_{s,t,n}|Z_{s,t}, \Psi, \Phi, \theta\right)\\
&\cdot \prod_{s}^{S}\prod_{t=1}^{T} p\left(c^{in}_{s,t}|b^{in}_{s,t} \cdot \lambda^{in}_{l}\right) \cdot p\left(c^{out}_{s,t}|b^{in}_{s,t} \cdot \lambda^{in}_{l}\right)\\
&\cdot p\left(\lambda^{in}_{l}, \lambda^{out}_{l}|\Theta_{II}\right).
\end{aligned}
$$

Thus, searching for the best set of model parameters for STM-S is equivalent to maximizing the logarithm of the joint distribution in Equation (11). The specific optimization process of this objective function is listed in Equations (18, 23–36) in Appendix A.1.

Utilizing the prior distributions Θ_I and Θ_{II} enables the model to estimate the spatial burstiness distribution even when there are no spatial outbreak observations. This advantage makes it possible to conduct event predictions even in new locations. More specifically for the model STM-I, according to Equation (19) in Appendix A.1, when $(r^{in}_{s,t}, r^{out}_{s,t})$ is not available, then $\hat{\mu}_{l,k} = 0$. Therefore, $\mu_{l,k} = \mu_0$ according to Equation (21). For STM-S, which utilizes a Poisson distribution, the deduction is the same, using Equations (27) and (28).

The time consumption required by the preceding algorithm is composed of two parts: (i) the computation of the forward-backward algorithm and (ii) the computation of Equations (18)–(36). The time complexity of the first of these is $S \cdot T \cdot K$, where S is the number of sequences, T is the time length of a sequence, and K is the number of latent states. The time complexity of the second is $S \cdot T \cdot V \cdot K + S \cdot T \cdot V \cdot K \cdot J$, where N is the size of the vocabulary and J is the number of latent topics. Combining the two parts and multiplying the result by the number of EM iterations q, the comprehensive time complexity, is $S \cdot T \cdot V \cdot K \cdot J \cdot q$.

Given the large numerical value of S, which includes all the historical sequences, the batch EM algorithm for estimating the model parameters is inevitably quite time-consuming. Moreover, as Twitter streams in real time, the batch-based updating of the model parameter to cope with the constant flow of incoming data requires the continual recalculation of the entire historical training set, which is prohibitively expensive in practice. To address this issue, we propose the use of an online parameter optimization method, which is introduced in the following section.

4.2. Online Parameter Optimization Algorithm

This section proposes online parameter optimization algorithms for STM-I and STM-S.

4.2.1. Parameter Optimization for STM-I. Unlike a standard batch EM algorithm, the capacity to perform online estimation means that the data must be run through only

once [Cappé 2011; Cappé and Moulines 2009]. The basic rationale of an online EM algorithm is to replace the expectation step by a stochastic approximation step while keeping the maximization step unchanged.

For STM-I, we first need to design the corresponding stochastic E-step, including the computation of the conditional expectations. However, unlike batch algorithms where all the event-specific language models θ^R are optimized iteratively, θ^R for each newly arriving event in an online algorithm cannot be known beforehand. Hence the likelihood in Equation (10) is unknown, which prevents the calculation of $\mathrm{E}[p(Z_{s_i,t} = k)]$. To address this problem, we propose to maximize the likelihood in Equation (10) with respect to θ^R, n^R, and n^B, which can be simplified into Equation (14). After calculating θ^R, the conditional expectations of unknown parameters can be obtained by a Stochastic E-step, discussed in more detail in Appendix A.2.

Utilizing the preceding stochastic E-step, the parameters for STM-I are trained on the fly based on the streaming data, as summarized in Algorithm 1. Specifically, the current sequence of social media message s_i is crawled from the data stream and utilized to calculate the conditional expectations for current data points, as in Steps 5, 7, 9, and 11. Then, the conditional expectations are used to update the sufficient statistics $\hat{\mu}_{l,k,i}$, $\hat{\Sigma}_{l,k,i}$, $g_{s,k,w,i}$, and $f_{k,j,w,i}$ in real time, as shown in Steps 6, 8, and 10. Finally, the maximum likelihoods of all the model parameters are calculated in Steps 3–19. This EM iteration is updated continuously while the data are streaming until the end of the stream.

4.2.2. Parameter Optimization for STM-S. As for STM-I, for STM-S the conditional expectations can be obtained through a Stochastic E-step. Here, the model parameters of STM-S θ^B, θ^R, Ψ, and Φ need to be initialized; this initialization follows the same strategy as for STM-I.

Utilizing the above-proposed stochastic E-step, the parameters for STM-S are once again trained on the fly based on the streaming data, as summarized in Algorithm 2. Specifically, the current sequence of social media message s_i is crawled from the data stream and utilized to calculate the conditional expectations for the current data point by performing Steps 7, 9, 13, and 15. Then, these conditional expectations are used to update the sufficient statistics $\hat{\lambda}^{in}_{c,l,k,i}$, $\hat{\lambda}^{in}_{b,l,k,i}$, $\hat{\lambda}^{out}_{c,l,k,i}$, $\hat{\lambda}^{out}_{b,l,k,i}$, $g_{s,k,w,i}$, and $f_{k,j,w,i}$ in real time, as shown in Steps 5, 8, 10, 14, and 16. Finally, the maximum likelihoods of all the model parameters are calculated in Steps 11 and 17–23. This EM iteration is performed continuously while the data are streaming until the end of the stream.

4.2.3. Time Complexity Analysis. As deduced in Section 4.1, for the batch algorithm, the time complexity is $S \cdot T \cdot V \cdot K \cdot J \cdot q$.

For the online algorithm, the time complexity of the E-step is $K \cdot T \cdot V \cdot J \cdot h$, and the time complexity of the M-step is $K \cdot T \cdot (L + J \cdot V + V) + J \cdot W$. Hence, the total time complexity is $(K \cdot T \cdot V \cdot J \cdot h + K \cdot T \cdot (L + J \cdot V)) \cdot q$.

This indicates that the time complexity of the online algorithm is independent of S, the number of sequences in the training set, but is linear in h, the number of iterations used to optimize θ^R, the language model for event-specific expressions.

5. SPATIOTEMPORAL EVENT FORECASTING

In this section, spatiotemporal event forecasting is formalized as a sequence classification problem based on the models proposed earlier, and an effective method for calculating the sequence likelihood is presented.

5.1. Sequence Classification

Given a sequence of tweets, it is first necessary to identify whether the underlying development revealed by this sequence will lead to an event or not. These two possibilities

ALGORITHM 1: Online EM Algorithm for STM-I

Input: D, $\Theta_0 = \{\mu_0, \beta_0, \Lambda_0, \nu_0\}$
Output: A^*, π^*, Ψ^*, Φ^*, θ^*, μ^*, Σ^*.

Set the initial learning rate $\gamma_0 = 0.5$. Initialize θ^B, θ^R, Ψ, and Φ. Set $i = 0$;
repeat
  Get current sequence s_i from Twitter data stream;
  Obtain optimal θ^R by maximizing the likelihood in Equation 10;
  Calculate $E[p(Z_{s_i,t} = k)]$ using forward-backward algorithm;
  for $k \leftarrow 1$ **to** K **do** // iterate over the K latent states
    for $l \leftarrow 1$ **to** L **do** // iterate over the L locations
      $\hat{N}_{l,k,i} \leftarrow (1-\gamma_i) \cdot \hat{N}_{l,k,i-1} + \gamma_i \cdot \sum_t \mathrm{E}[p(Z_{s_i,t} = k)]$;
      $\hat{\mu}_{l,k,i} \leftarrow (1-\gamma_i) \cdot \hat{\mu}_{l,k,i-1} + \gamma_i \frac{\sum_t \mathrm{E}[p(Z_{s_i,t}=k)](r^{in}_{s_i,t}, r^{out}_{s_i,t})}{\hat{N}_{l,k,i}}$;
      $E^{\hat{\Sigma}}_{l,k,i} \leftarrow \sum_t \mathrm{E}[p(Z_{s_i,t}=k)] \left(\hat{\mu}_{l,k} - (r^{in}_{s_i,t}, r^{out}_{s_i,t})\right)^2 / \hat{N}_{l,k,i}$;
      $\hat{\Sigma}_{l,k,i} \leftarrow (1-\gamma_i) \cdot \hat{\Sigma}_{l,k,i-1} + \gamma_i \cdot E^{\hat{\Sigma}}_i$;
      $\mu_{l,k,i} = (\beta_0 \mu_0 + \hat{N}_{l,k,i} \cdot \hat{\mu}_{l,k,i}) / (\beta_0 + \hat{N}_{l,k,i})$;
      $\Sigma_{l,k,i} = \frac{\Lambda_0 + \hat{\Sigma}_{l,k,i}}{\nu_0 + 3} + \frac{\beta_0 \hat{N}_{l,k,i} (\hat{\mu}_{l,k,i} - \mu_0)(\hat{\mu}_{l,k,i} - \mu_0)^T}{(\beta_0 + \hat{N}_{l,k,i})(\nu_0 + 3)}$;
    end
    for $j \leftarrow 1$ **to** J **do** // iterate over the J latent topics
      for $w \leftarrow 1$ **to** V **do** // iterate over the V terms
        $E^f_{k,j,w,i} \leftarrow \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)] \cdot \Psi_{k,2} \cdot \Phi_{k,j} \theta_w^{B_j}}{\Psi_{k,1} \theta^R_{s_i,t,w} + \Psi_{k,2} \sum_j \Phi_{k,j} \theta_w^{B_j}}$;
        $f_{k,j,w,i} \leftarrow (1-\gamma_i) \cdot f_{k,j,w,i-1} + \gamma_i \cdot E^g_i$;
      end
      $\Phi_{k,j,i} = \sum_w f_{k,j,w,i} / \sum_j \sum_w f_{k,j,w,i}$;
    end
    for $w \leftarrow 1$ **to** V **do** // iterate over the V terms
      $E^g_{s_i,k,w,i} \leftarrow \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)] \cdot \Psi_{k,1} \cdot \theta^R_{s_i,t,w}}{\Psi_{k,1} \theta^R_{s_i,t,w} + \Psi_{k,2} \sum_j \Phi_{k,j} \theta_w^{B_j}}$;
      $g_{s_i,k,w,i} \leftarrow (1-\gamma_i) \cdot g_{s,k,w,i-1} + \gamma_i \cdot E^g_i$;
    end
    $\Psi_{k,1,i} = \sum_{s,w} g_{s,k,w,i} / (\sum_{s,w} g_{s,k,w,i} + \sum_{w,j} f_{k,j,w,i})$;
    $\Psi_{k,2,i} = \sum_w \sum_j f_{k,j,w,i} / (\sum_s \sum_w g_{s,k,w,i} + \sum_w \sum_j f_{k,j,w,i})$;
  end
  for $j \leftarrow 1$ **to** J **do** // iterate over the J latent topics
    for $w \leftarrow 1$ **to** V **do** // iterate over the V terms
      $\theta^{B_j}_{w,i} = \sum_k f_{k,j,w,i} / \sum_k \sum_w f_{k,j,w,i}$;
      $\theta^{B_j}_{w,i} = \sum_k f_{k,j,w,i} / \sum_k \sum_w f_{k,j,w,i}$;
    end
  end
  $i \leftarrow i + 1$;
until *the end of data stream*;

each have a corresponding set of sequences, and the two proposed models are trained based on these sequences: one model characterizes the development process leading to an event, while the other characterizes a process that does not lead to an event. For the prediction, an unknown sequence will be aligned with the model in each class. This sequence will be classified into the class corresponding to the higher alignment score.

Denote C_1 as the model trained for the class corresponding to the situation: "future event," while C_2 is the model corresponding to "no event." Denote e_1 as the cost of

ALGORITHM 2: Online EM Algorithm for STM-S

Input: D, $\Theta_0 = \{\mu_0, \beta_0, \Lambda_0, \nu_0\}$
Output: A^*, π^*, Ψ^*, Φ^*, θ^*,λ^{in_*}, λ^{out_*}.

Set the initial learning rate $\gamma_0 = 0.5$. Initialize θ^B, θ^R, Ψ, and Φ. Set $i = 0$;
repeat
  Get current sequence s_i from Twitter data stream;
  Obtain optimal θ^R by maximizing the likelihood in Equation 11;
  Calculate $E[p(Z_{s_i,t} = k)]$ using forward-backward algorithm;
  for $k \leftarrow 1$ **to** K **do** // iterate over the K latent states
    for $l \leftarrow 1$ **to** L **do** // iterate over the L locations
      $\hat{N}_{l,k,i} \leftarrow (1-\gamma_i) \cdot \hat{N}_{l,k,i-1} + \gamma_i \cdot \sum_t \mathrm{E}[p(Z_{s_i,t} = k)]$;
      for $m \in \{in, out\}$ **do** // iterate over the inside and outside ratios
        $E^{\hat{c}^m}_{l,k,i} \leftarrow \sum_t c^m_{s_i,t} \cdot \mathrm{E}[p(Z_{s_i,t} = k)]/\hat{N}_{l,k,i}$;
        $\hat{\lambda}^m_{c,l,k,i} \leftarrow (1-\gamma_i) \cdot \hat{\lambda}^m_{c,l,k,i-1} + \gamma_i \cdot E^{\hat{c}^m}_i$;
        $E^{\hat{b}^m}_{l,k,i} \leftarrow \sum_t b^m_{s_i,t} \cdot \mathrm{E}[p(Z_{s_i,t} = k)]/\hat{N}_{l,k,i}$;
        $\hat{\lambda}^m_{b,l,k,i} \leftarrow (1-\gamma_i) \cdot \hat{\lambda}^m_{b,l,k,i-1} + \gamma_i \cdot E^{\hat{b}^m}_i$;
        $\lambda^m_{k,l,i} = \frac{(\alpha^m - 1) + \hat{\lambda}^m_{c,k,l,i}}{\beta^m + \hat{\lambda}^m_{b,k,l,i}}$;
      end
    end
    for $j \leftarrow 1$ **to** J **do** // iterate over the J latent topics
      for $w \leftarrow 1$ **to** V **do** // iterate over the V terms
        $E^f_{k,j,w,i} \leftarrow \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)] \cdot \Psi_{k,2} \cdot \Phi_{k,j} \theta_w^{B_j}}{\Psi_{k,1} \theta^R_{s_i,t,w} + \Psi_{k,2} \sum_j \Phi_{k,j} \theta_w^{B_j}}$;
        $f_{k,j,w,i} \leftarrow (1-\gamma_i) \cdot f_{k,j,w,i-1} + \gamma_i \cdot E^g_i$;
      end
      $\Phi_{k,j,i} = \sum_w f_{k,j,w,i} / \sum_j \sum_w f_{k,j,w,i}$;
    end
    for $w \leftarrow 1$ **to** V **do** // iterate over the V terms
      $E^g_{s_i,k,w,i} \leftarrow \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)] \cdot \Psi_{k,1} \cdot \theta^R_{s_i,t,w}}{\Psi_{k,1} \theta^R_{s_i,t,w} + \Psi_{k,2} \sum_j \Phi_{k,j} \theta_w^{B_j}}$;
      $g_{s_i,k,w,i} \leftarrow (1-\gamma_i) \cdot g_{s,k,w,i-1} + \gamma_i \cdot E^g_i$;
    end
    $\Psi_{k,1,i} = \sum_{s,w} g_{s,k,w,i} / (\sum_{s,w} g_{s,k,w,i} + \sum_{w,j} f_{k,j,w,i})$;
    $\Psi_{k,2,i} = \sum_w \sum_j f_{k,j,w,i} / (\sum_s \sum_w g_{s_i,k,w,i} + \sum_w \sum_j f_{k,j,w,i})$;
  end
  for $j \leftarrow 1$ **to** J **do** // iterate over the J latent topics
    for $w \leftarrow 1$ **to** V **do** // iterate over the V terms
      $\theta^{B_j}_{w,i} = \sum_k f_{k,j,w,i} / \sum_k \sum_w f_{k,j,w,i}$;
      $\theta^{B_j}_{w,i} = \sum_k f_{k,j,w,i} / \sum_k \sum_w f_{k,j,w,i}$;
    end
  end
  $i \leftarrow i + 1$;
until *the end of data stream*;

misclassifying the first class as the second class, while e_2 is the cost of misclassifying the second class as the first class. The spatiotemporal event forecasting problem can be formalized as follows: Given a newly arriving sequence of tweets s in location l, if $p(C_1|s,l) > \varepsilon \cdot p(C_2|s,l)$, then a future event is deemed likely to happen; $p(C_1|s,l) \leq \varepsilon \cdot p(C_2|s,l)$, where $\varepsilon = e_1/e_2$ is the cost ratio.

According to the Bayesian rule, we have $p(C_i|s,l) = p(s|C_i) \cdot p(C_i|l)/p(s)$, $i = 1, 2$, where $p(C_1|l)$ denotes the prior probability that an event will occur in location l; $p(C_2|l) = 1 - p(C_1|l)$ denotes the prior probability that no event will occur in location l; and $p(s)$ is a constant and thus can be omitted. If the historical record for location l is not available, the preceding Bayesian decision rule is formalized as $p(C_i|s) = p(s|C_i) \cdot p(C_i)/p(s)$, $i = 1, 2$, where $p(C_1)$ is the overall prior probability of event occurrence in any location, while $p(C_2) = 1 - p(C_1)$ denotes the prior probability that no event occurs. Finally, the sequence likelihood $p(s|C_i)$ is calculated based on the method described in the next section.

5.2. Calculation of Sequence Likelihood

In a standard HMM, dynamic programming methods such as the Viterbi algorithm [Chen et al. 2005] are typically utilized to calculate the likelihood of a newly arriving sequence by finding the most likely sequence of latent states. In our model, however, the traditional Viterbi algorithm is not applicable because our model needs to determine the optimal language models $\theta^R = \{\theta^R_{s,t}\}^{S,T}_{s,t}$ that represent the words exclusive to this newly arriving sequence. The calculation of sequence likelihood based on our model involves identifying the most probable latent states and the parameter θ^R that maximizes the probability $p(s|C_i)$:

$$p(s|C_i) = \max_{\{Z_t\}^T_t, \theta^R, n^R, n^B} \ln p(s, Z_1, \ldots, Z_T|C_i), \tag{12}$$

where $n^R = \{n^R_{s,t}\}^{S,T}_{s,t}$ is the number of words explained by the language model θ^R in sequence s at time step t and $n^B = \{n^{B_j}_{s,t}\}^{S,T,J}_{s,t,j}$ is the number of the words explained by different latent topics. By introducing the notation ω_t such that $\omega_t \equiv \ln p(s, Z_1, \ldots, Z_t|C_i)$, Equation (12) can be solved by recursively calculating the following equation:

$$\omega_t = \max_{\theta^R_{s,t}, n^R_{s,t}, n^B_{s,t}} \ln p(s_t|Z_t, C_i) + \max_{Z_{t-1}} \{\ln p(Z_t|Z_{t-1}) + \omega_{t-1}\}, \tag{13}$$

with the initial iteration: $\omega_t = \max_{\theta^R_{s,1}, n^R_{s,1}, n^B_{s,1}} \ln p(s_1|Z_1, C_i) + \ln p(Z_1)$. The variables $\{Z_t\}^T_t$ can be solved via a standard max-sum algorithm.

Next, we address the optimization problem: $\max_{\theta^R_{s,t}, n^R_{s,t}, n^B_{s,t}} \ln p(s_t|Z_t, C_i)$. By referring to Equation (10) and omitting the constant term, the problem can be formalized as the following maximization problem:

$$\max_{\theta^R_{s,t}, n^R_{s,t}, n^B_{s,t}} \sum_i^V n^R_{s,t,i} \cdot \log \theta^R_{s,t,i} + \sum_i^V \sum_j^J n^{B_j}_{s,t,i} \cdot \log \theta^{B_j}_i \tag{14}$$

$$s.t. \sum_i^V \theta^R_{s,t,i} = 1, n^R_{s,t,w} + \sum_j^J n^{B_j}_{s,t,w} = \xi_w, n^R_{s,t,w} \geq 0$$

$$n^{B_j}_{s,t,w} \geq 0, \sum_i^V n^{B_j}_{s,t,i} = \xi \cdot \Psi_{k,2} \Phi_{k,j}, \sum_i^V n^R_{s,t,i} = \xi \cdot \Psi^R_{k,1},$$

where ξ denotes the number of words in sequence s at time step t, $k = Z_t$ is the current latent state in sequence s, and V is the size of the vocabulary. The coupling between the variables $n^R_{s,t,i}$ and $\theta^R_{s,t,i}$ prevents a globally optimal solution to this problem, so Lagrangian multipliers are added to enforce the constraints. Setting the derivative

Table II. Datasets and Event Labels

Dataset	Time Period	# Raw Tweets	# Processed Tweets	#Events
Civil unrest	2013-01-01 - 2013-06-01	32,459,668	57,856	726
Flu	2011-01-01 - 2013-12-31	8,627,664,399	2,252,436	102

with respect to $\theta^R_{s,t,i}$ to 0, we obtain:

$$\frac{n^R_{s,t,i}}{\theta^R_{s,t,i}} + \gamma = 0, \tag{15}$$

where γ is the Lagragian multiplier for the first equality constraint. By utilizing the first two equality constraints in Equation (14), we can derive:

$$\theta^R_{s,t,i} = \frac{n^R_{s,t,i}}{\xi \cdot \Psi^R_{k,1}}. \tag{16}$$

Substituting Equation (16) into Equation (14), we get

$$\max_{n^R_{s,t}, n^B_{s,t}} \sum_i^V n^R_{s,t,i} \cdot \log \frac{n^R_{s,t,i}}{\xi \cdot \Psi^R_{k,1}} + \sum_i^V \sum_j^J n^{B_j}_{s,t,i} \cdot \log \theta_i^{B_j} \tag{17}$$

$$s.t. n^R_{s,t,w} + \sum_j^J n^{B_j}_{s,t,w} = \xi_w, n^R_{s,t,w} \geq 0, n^{B_j}_{s,t,w} \geq 0,$$

$$\sum_i^V n^{B_j}_{s,t,i} = \xi \cdot \Psi_{k,2} \Phi_{k,j}, \sum_i^V n^R_{s,t,i} = \xi \cdot \Psi^R_{k,1.}$$

Here, the objective function in Equation (17) is convex with respect to $n^R_{s,t}$ and $n^{B_j}_{s,t}$. This means that the global solution can be found using a traditional numerical optimization method, such as the interior point method [Mehrotra 1992]. After $n^R_{s,t}$ and $n^{B_j}_{s,t}$ are optimized, $\theta^R_{s,t}$ can be calculated based on Equation (16). Finally, the maximization problem in Equation (12) is solved and thus the sequence likelihood can be calculated.

6. EXPERIMENTAL EVALUATION

This section presents an experimental evaluation of the effectiveness and efficiency of the proposed approach based on comprehensive experiments on two different sets of Twitter data, the first of which seeks to forecast civil unrest events such as protests and strikes in Mexico and the second flu outbreaks in the United States. All the experiments were conducted on a computer with a 2.6GHz Intel i7 CPU and 16GB RAM.

6.1. Experiment Design

This subsection presents the configuration of the datasets, the gold standard report for these event labels (shown in Table II), data processing, comparison methods, parameter settings, and performance metrics.

Datasets: For the analysis of civil unrest event forecasting, we collected 10% of the raw Twitter data for Mexico through Datasift's Twitter collection engine from January 1, 2013, to June 1, 2013. The data from January 1, 2013, to February 28, 2013, were used for training and the remainder for testing. For the analysis of flu forecasting, we collected tweets containing at least one of 124 predefined flu-related keywords (e.g., "cold," "fever," and "cough") during the period from January 1, 2011, to December 31,

2013, from across the United States. The data from January 1, 2011, to January 1, 2013, were used for training and the subsequent tweets for testing.

Gold Standard Report of Event Labels: The civil unrest forecasting results were validated against a labeled set known as the Gold Standard Report (GSR) that was exclusively provided by MITRE (see Ramakrishnan et al. [2014] for more details). The GSR was organized by manually harvesting civil unrest event reports from the 10 most significant news outlets[2] in Mexico and the world, as ranked by International Media and Newspapers[3] There were a total of 726 events during the period January 1, 2013, to Junr 1, 2013. An example of a labeled GSR event is given by the tuple: (CITY = "Hermosillo", STATE = "Sonora", COUNTRY = "Mexico", DATE = "2013-01-20"). The forecasting results for the flu outbreaks were validated against the flu statistics reported by the Centers for Disease Control and Prevention (CDC). CDC publishes the weekly influenza-like illness (ILI) activity level within each state in the United States using the proportion of the outpatient visits to healthcare providers for ILI. There are four ILI activity levels: minimal, low, moderate, and high, where the level "high" corresponds to a salient flu outbreak and is thus considered for forecasting. There were a total of 102 events during the period January 1, 2011, to December 31, 2013. An example of a CDC flu outbreak event i: (STATE = "Michigan", COUNTRY = "United States", WEEK = "01-06-2013 to 01-12-2013").

Data Preprocessing: For the first dataset, three labelers collectively labeled 20,906 tweets in both English and Spanish during June 2012 to February 2013. After two had labeled all the tweets into positive (i.e., relevant to civil unrest) or negative, all the tweets where they disagreed were sent to the third labeler for final determination. Consequently, the tweets were categorized as 6,793 positive and 14,113 negative, and the results used to train a linear SVM classifier. For the second dataset, we utilized the labeled set in Lamb et al. [2013] and used these to train a linear SVM to identify tweets relevant to the flu. Both SVMs were generated based on unigram features containing all the distinct words with frequencies greater than 20 in the individual datasets. The trained SVM classifiers extracted the tweets deemed relevant to civil unrest and flu from the respective datasets. The locations of the tweets were extracted from the geotags (coordinates and places); those tweets without geotags were discarded.

Comparison Methods: There are four proposed approaches evaluated in this article: STM-I, STM-S, and their two online versions, namely STM-I (online) and STM-S (online). Our proposed approaches were compared with four representative methods and one baseline method. The *Autoregressive Exogenous Model (ARX)* [Achrekar et al. 2011] assumes that for each separate location, the count of future events is dependent on both the count of historical events and the tweet volume. In forecasting, an output above "1" indicates that an event has occurred; otherwise, no event is deemed to have occurred. The *Linear Regression (LinReg) model* [Arias et al. 2013; Bollen et al. 2011; He et al. 2013; O'Connor et al. 2010] assumes that for each separate location there is a linear relationship between tweet observations and event occurrences ("0" denotes nonoccurrence, "1" denotes occurrence). The input feature here is the volume of domain-related tweets. When forecasting, an output below 0.5 indicates no event; an output over 0.5 indicates that an event has occurred. In the *Logistic Regression (LogReg) model* [Wang et al. 2012b], event forecasting is treated as a classification problem. Here, the input features are the proportions of latent topics extracted from the tweet texts coming from a specific location based on the latent dirichlet allocation. The output is 0 if there is no event and 1, if there is one. The *Kernel Density Estimation-Based Logistic Regression (KDE LogReg) model* [Gerber 2014] forecasts

[2]These are La Jornada, Reforma, Milenio, The New York Times, The Guardian, The Wall Street Journal, The Washington Post, The International Herald Tribune, The Times of London, and Infolatam.

[3]International Media and Newspapers website. Available: http://www.4imn.com/. Accessed on Oct 1, 2014.

Table III. Event Forecasting Results for Civil Unrest

Metric	Precision	Recall	F-measure	Runtime
Baseline	0.44±0.16	0.59±0.19	0.50±0.07	0.001
ARX	0.26±0.06	0.43±0.05	0.32±0.09	0.001
LinReg	0.70±0.28	0.18±0.04	0.29±0.11	0.001
LDA-LogReg	0.31±0.06	0.70±0.16	0.43±0.14	0.005
KDE-LDA-LogReg	0.42±0.05	0.69±0.21	0.52±0.20	0.005
ST-Burst	0.29±0.19	0.80±0.17	0.42±0.04	0.008
STM-I	0.75±0.29	0.70±0.31	0.72±0.27	0.32
STM-S	0.58±0.26	0.54±0.24	0.56±0.21	0.33
STM-I (online)	0.53±0.11	0.52±0.33	0.53±0.16	0.33
STM-S (online)	0.64±0.12	0.55±0.32	0.60±0.17	0.50

the event occurrence at a location by considering the historical event numbers and the tweet semantics. The set of input features consists of a combination of (i) the historical event numbers spatially smoothed by KDE and (ii) the proportions of the latent topics of tweet content. The *Spatial Temporal Burst Detection (ST-Burst)* [Lappas et al. 2012] is proposed to discover bursts of terms in a specific spatial and temporal neighborhood. The tunable temporal window size was set to 5 in the original work. We also evaluated other values, including 12 and 24, but observed similar results. Finally, the *baseline* method considers the probability of historical event occurrence to be the probability of future event occurrence. More specifically, for each location, it calculates the percentages of positive (i.e., event occurrence) and negative cases (i.e., no occurrence) based on a training set. When making a prediction for each observation, it randomly votes "positive" or "negative" following the preceding empirical percentage for the location of the current observation.

Parameter Settings: Except for the baseline method, which does not require parameters, all the comparison methods were implemented based on the algorithms presented in the original papers. We strictly followed the strategies recommended by the authors to select features and estimated the model parameters via a 10-fold cross-validation. The new method proposed here incorporates several prior parameters and three tunable parameters. The four prior hyperparameters were set as follows: The historical prior ratio mean μ_0 was set as the mean of the domain-related tweet ratios in all the locations and in all the time steps; the prior scale matrix Λ_0 was set as an identity matrix; the number of prior measurements β_0 was set to 1; and the degrees of freedom ν_0 to the dimension of the vector $\mu_{k,l}$. The three tunable parameters were the misclassification cost ratio ε, the number of latent topics J, and the number of latent states K, which were set as 10, 5, and 4, respectively, based on a 10-fold cross-validation.

Performance Metrics: Three main performance metrics were considered here: precision, recall, and F1-score. The reported forecasting alerts were structured as tuples of (date, location), where "location" is defined at the city level for civil unrest events and state level for flu outbreaks. A forecasting alert was matched to a true event if both the date and the location attributes matched; otherwise, it was considered to be a false forecast. Note that because the time granularity of the CDC flu outbreak labels is at week-level, a match in time was deemed to have occurred if the forecast date of an alert fell within a week of a true flu outbreak event.

6.2. Event Forecasting Results

Table III presents the comparison between our four approaches and the six competing methods for the task of forecasting civil unrest events. The test set was split into 20 bins on which the prediction performance and its standard deviations were evaluated.

Table IV. Event Forecasting Results for the Flu Dataset

Metric	Precision	Recall	F-measure	Runtime
Baseline	0.28±0.09	0.39±0.16	0.33±0.13	0.001
ARX	0.14±0.04	0.66±0.22	0.23±0.01	0.001
LinReg	0.64±0.16	0.31±0.07	0.41±0.12	0.01
LDA-LogReg	0.27±0.09	0.55±0.22	0.36±0.11	0.02
KDE-LDA-LogReg	0.78±0.10	0.32±0.08	0.46±0.06	0.03
ST-Burst	0.63±0.15	0.45±0.10	0.53±0.12	0.1
STM-I	0.83±0.19	0.69±0.13	0.75±0.18	2.1
STM-S	0.63±0.23	0.53±0.13	0.58±0.19	2.5
STM-I (online)	0.69±0.10	0.76±0.21	0.72±0.16	2.0
STM-S (online)	0.68±0.07	0.52±0.25	0.59±0.21	2.5

Here, our proposed new approaches achieved the best overall performance in precision, recall, and F1-score, outperforming the five comparison methods by up to 38% in F1-score and 7% in precision. This could be because our approach considers the spatial burstiness as well as the tweet content, which is crucial for the forecasting of civil unrest events. Among our approaches, STM-I, which is the Gaussian-distribution batch-based model, achieved the best performance; the batch-based approaches STM-I and STM-S generally outperformed their online counterparts, but the online versions still outperformed the competing methods by a substantial margin on both precision and recall. The proposed STM-I and STM-S have relatively high standard deviations. The KDE Logistic Regression achieved an F1-score that was 21% higher than those of ARX, LinReg, and LogReg due to its consideration of spatial dependencies. ST-Burst also considers the spatial dependencies and achieves a high recall and smallest standard deviation on F-measure. The poor performances of ARX and LinReg indicate that focusing solely on tweet volume is insufficient for the task of civil unrest event forecasting. Thus, the tweet content and the spatial burstiness are both important factors for this type of application. The baseline method achieved a reasonably good performance, indicating that it captured important historical event counts in different locations.

Table IV demonstrates that our approaches also consistently achieved the best performance in precision, recall, and F1-score for the task of flu outbreak event forecasting. The F1-scores achieved by the proposed models were up to 63% higher than those of the five comparison methods. Among our approaches, the batch version of STM-I again achieved the best performance. Of the existing approaches, ST-Burst and KDE LogReg achieved the highest F1-scores (i.e., F-measures), suggesting the importance of considering spatial burstiness. They also achieved relatively low standard deviations. The F1-score of the baseline was 34% lower than that in the civil unrest dataset, probably because the civil unrest events were clustered in several geographic regions, but the flu outbreak events were scattered across states. As a result, the use of prior information for event location distribution is effective in the civil unrest dataset, but noninformative in the flu dataset. LinReg, on the other hand, achieved a 41% higher F1-score in the flu dataset than in the civil unrest dataset, which indicates that the tweet volume information plays an important role in this scenario. This could also explain why the comparison method LogReg, which only considers tweet semantics, performed less well here than it did in the civil unrest dataset.

The proposed approaches and the five comparison methods all forecast next-day events at the daily level. The running times of our new approaches were on average 0.35 seconds per day on the civil unrest dataset and 2.3 seconds per day on the flu dataset. These were markedly longer than the running time of the comparison methods for both datasets, primarily because our approach considers the characterization of

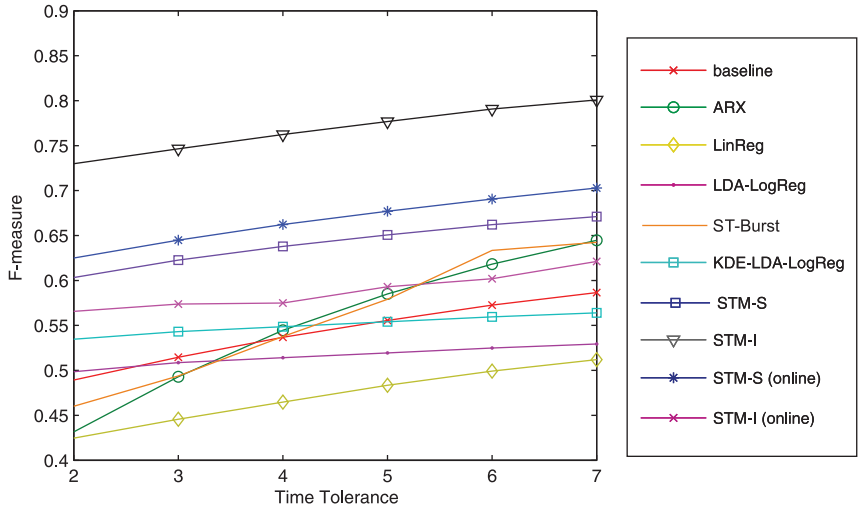


Fig. 11. Prediction performance with respect to the tolerance of predicted time error on the civil unrest dataset. The number of true positives is enlarged when the time tolerance increases.

temporal correlations among tweet contents and the optimization of the language model for event-specific words. However, the running times achieved by our approach were only a maximum of 3 seconds longer than those of the five comparison methods, and the resulting gain in forecast accuracy for next-day events makes this eminently practical for real-world applications.

In many real-world applications, it is not strictly required that the predicted time of events must be accurate within a timestep (e.g., on a precise date), and the next multiple time step is acceptable. For example, when forecasting civil unrest events, users may be interested in predicting whether or not there will be an event occurring within the next so many days. Instead of requiring very accurate predicted times, users may instead emphasize a sufficient lead time for forecasting. Similarly, when forecasting flu outbreaks, people may be interested in forecasting whether or not the influenza activity will be high over the next few weeks. To evaluate the performance of all the methods in this situation, the impact of increasing the correct predictions with respect to a higher tolerance for predicted time error is validated by the data shown in Figures 11 and 12.

In Figure 11, the F-measures of all the methods with respect to increasing the tolerance of the predicted time error are illustrated. As the graph shows, all the F-measures increase when the time tolerance increases. Among the methods, STM-I achieves the highest F-measure, about 0.80, when the time tolerance is 7 days. ARX obtains the largest increase in the rate, from 0.43 at 2 days to 0.64 at 7 days, which indicates a robust prediction performance. Similar to the pattern of ARX, ST-Burst also achieves a significant increase in the performance. STM-I and STM-I (online) also achieve competitive performances of around 0.70 when the tolerance time is almost 7 days.

Figure 12, which shows the equivalent information for the flu dataset, shows a similar pattern to that in Figure 11. Once again, the F-measures of all the methods increase when the tolerances of the predicted time errors increase. Here, the methods STM-I, STM-I (online), STM-S, STM-S(online), ARX, and ST-Burst achieve the best performances. The performance of ARX ramps up fastest when the time tolerance increases, finally achieving an F-measure of 0.76 at 7 days.

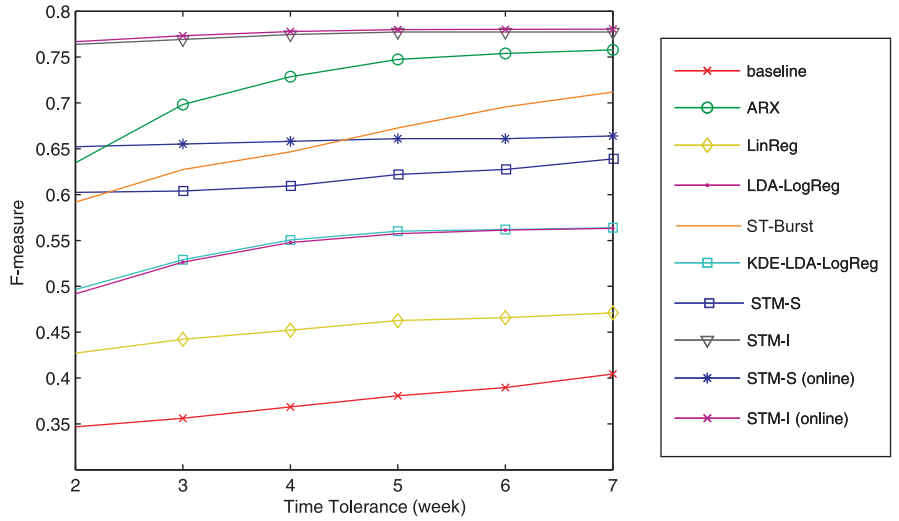


Fig. 12. Prediction performance with respect to the tolerance of predicted time error on the flu dataset. The number of true positives is enlarged when the time tolerance increases.

Table V. Performance Evaluation of Model Components

	only content component			only spatial component		
civil unrest	precision	recall	F-measure	precision	recall	F-measure
STM-I	0.54	0.90	0.68	0.66	0.82	0.73
STM-S	0.34	0.87	0.49	0.35	0.85	0.50
STM-I (online)	0.44	0.79	0.57	0.54	0.72	0.62
STM-S (online)	0.3	0.93	0.45	0.31	0.89	0.46
influenza	precision	recall	F-measure	precision	recall	F-measure
STM-I	0.34	0.81	0.48	0.63	0.68	0.65
STM-S	0.44	1.00	0.61	0.59	0.68	0.63
STM-I (online)	0.34	0.81	0.48	0.66	0.66	0.66
STM-S (online)	0.41	0.78	0.54	0.68	0.55	0.61

The proposed models take into account both structural texts and spatial outbreaks in event forecasting. To examine the respective usefulness of these two components, Table V presents the performance evaluations on the models that only consider a single component; namely, the models only characterize structural texts or spatial outbreaks. By comparing the performance with the complete models in Tables III and IV, it is easy to see that the complete models outperform their "one-component" versions in F-measures. This demonstrates the advantages in considering both components in our models. In addition, by comparing the performance between the models with only a spatial component and that with only a content component, it can be seen that the former outperforms the latter especially on the influenza dataset. This indicates that for the forecasting tasks in influenza outbreaks, the spatial outbreaks information is potentially more important than that in the civil unrest tasks.

6.3. Sensitivity Analysis

This section presents the sensitivity analysis. Here, we will only consider the STM-I model since the experimental results for the other models all follow a similar pattern.

Figures 13 and 14 illustrate the impact of the number of latent states and the number of latent topics on event forecasting performance. Varying the number of latent topics

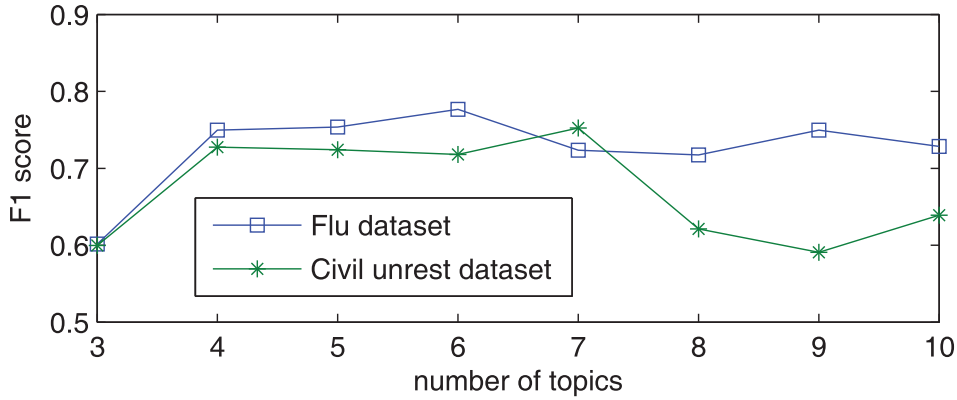


Fig. 13. Sensitivity analysis with respect to number of latent topics.

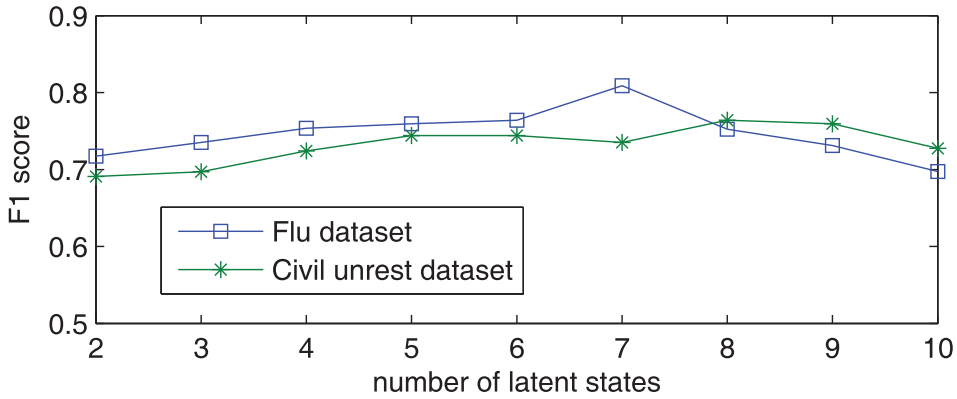


Fig. 14. Sensitivity analysis with respect to number of latent states.

from 3 to 10 caused the F1-score on the civil unrest and the flu datasets to vary between 0.6 and 0.8; even when the number of latent states rose from 2 to 10, the perturbation in the F1-scores remained between 0.7 to 0.8 for both datasets, indicating that the performance is less sensitive to the number of latent states than the latent topics in the given value interval for the parameters. For both parameters, the performance for low values was relatively poor. For the number of latent topics, the range from 4 to 7 achieved the best performance, while for the number of latent states, the range from 4 to 9 corresponded to a good performance.

The precision-recall curves of the new approach and the baseline method are shown in Figures 15(a) and 15(b) for the civil unrest and flu datasets, respectively. To produce these curves, ε, the cost ratio of false positive to false negative was varied from 0.01 to 1 in increments of 0.01, and from 1 to 100 in increments of 1. For both civil unrest and flu forecasting, the performance of our approach clearly outperformed the baseline model.

6.4. Scalability

The training time for the batch-based models is typically sensitive to the size of the training set. Figures 16 and 17 illustrate the impact of scalability on the number of training samples needed by the four proposed approaches for the civil unrest dataset and flu dataset, respectively.

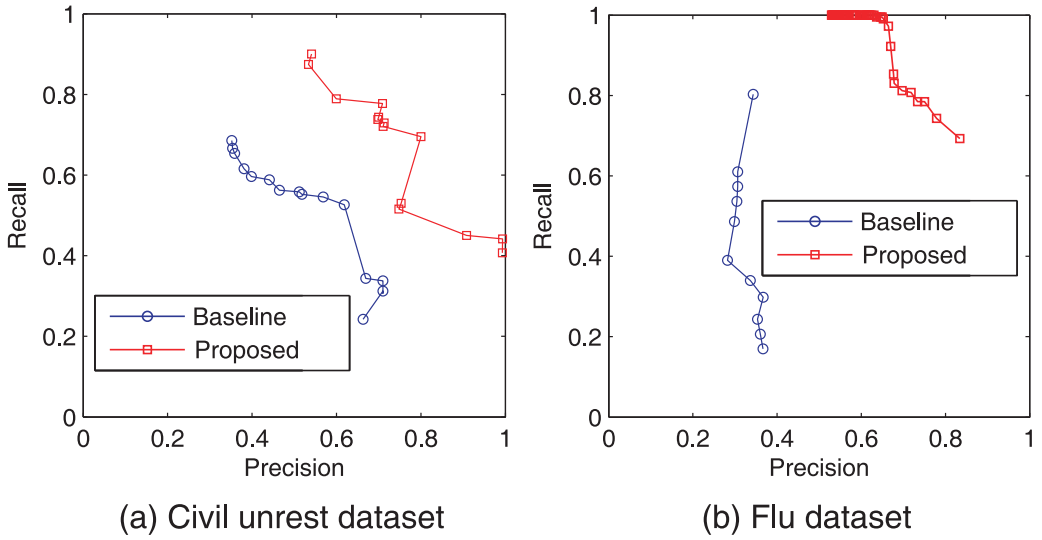


Fig. 15. Precision-recall curves on civil unrest and flu data. The proposed model consistently outperforms the baseline when the cost ratio ε varies.

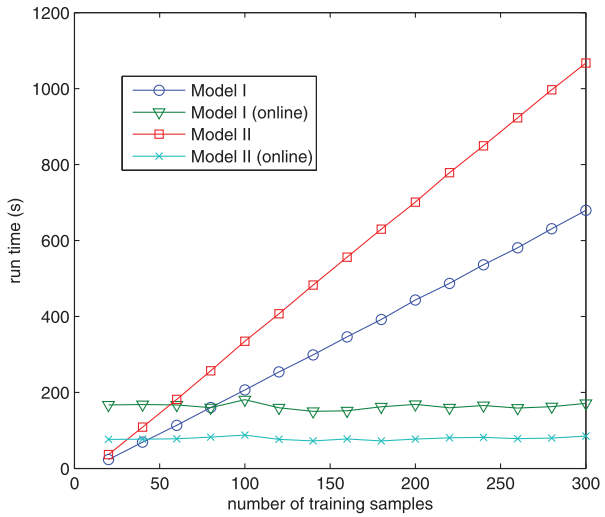


Fig. 16. Scalability of the proposed models for the civil unrest dataset. The runtimes of the batch-based models increase linearly with the size of training set, while the runtimes of the online models are constant.

As shown in Figure 16, for the civil unrest dataset, the runtimes for training STM-I and STM-S are linear in the number of training samples, starting from only 10 seconds with 20 samples and rising to 1,000 seconds with 300 samples. Unlike batch-based models, the training times for the online versions, STM-I (online) and STM-S (online), were not sensitive to the number of training samples utilized, with a relatively constant runtime of around 150 seconds.

On the flu dataset, shown in Figure 17, the runtime for all four approaches was longer than for the civil unrest dataset due to the larger scale of the data. The runtime for training STM-I and STM-S once again increased linearly with the number of training samples, starting from only 10 seconds with 20 samples and rising to 1,600 seconds

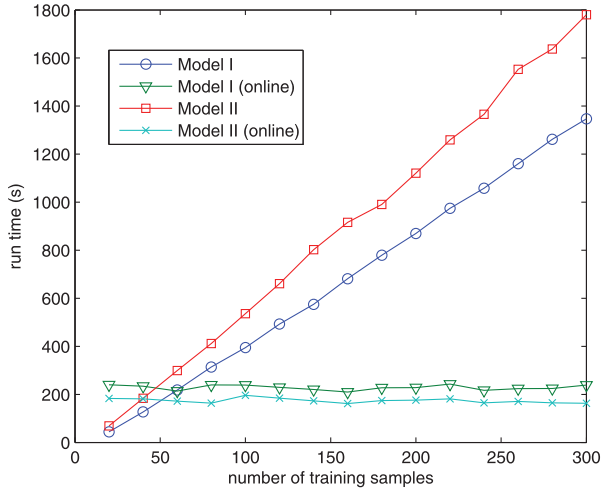
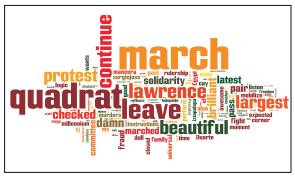


Fig. 17. Scalability of the proposed models for the flu dataset. The runtimes of the batch-based models increase linearly with the size of the training set, while the runtimes of the online models are constant.



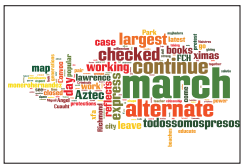
(a) topic 1 (describe protest)

(b) topic 2 (background)

(c) topic 3 (describe protest)

(d) topic 4 (descriptive)

(e) topic 5 (background)

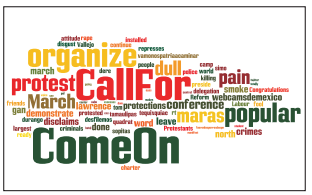


(f) topic 6 (descriptive)

(g) topic 7 (descriptive)

(h) topic 8 (call for protest)

(i) topic 9 (advertise protest)



(j) topic 10 (call for protest)

Fig. 18. Illustration of all 10 topics extracted (translated into English). Topics 2 and 5 contain general background words; Topics 1, 3, 4, 6, and 7 tend to include descriptive words for the protests; Topics 8 and 10 focus on the words specifically calling for a protest. Topic 9 largely contains words related to disseminating information on the planned protests.

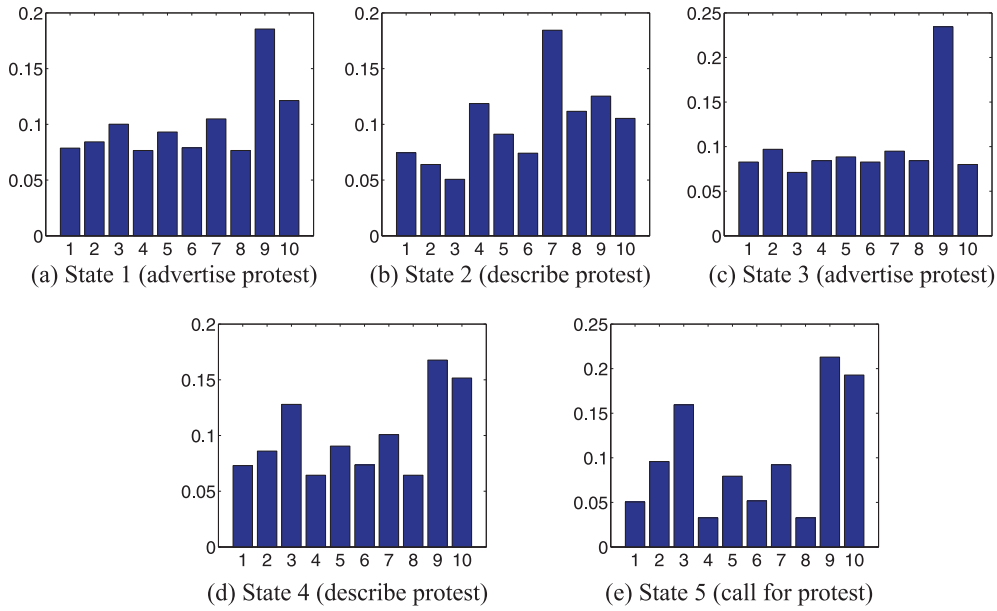


Fig. 19. Contexts of the five latent states indicating the development stages of events. For different stages of the developing protest potential, the distributions of topics changes. States 2 and 4 could indicate the discussion about a potential protest, States 1 and 3 could reveal the spread of propaganda about the planned protests, while State 5 might be related to the organization of the protest.

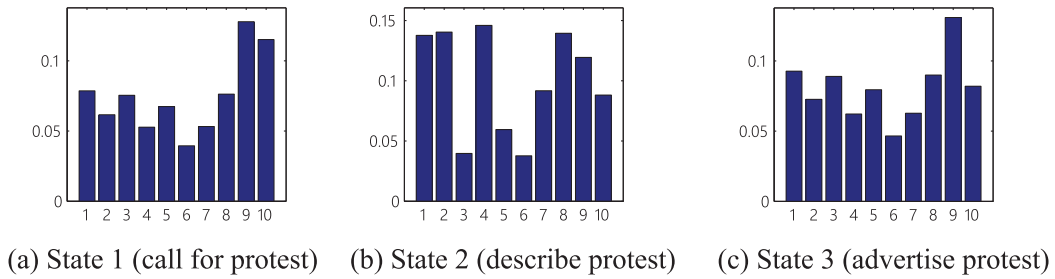


Fig. 20. Contexts of the three latent states indicating the development stages of events. For different stages of the developing potential protest, the distributions of topics change. State 2 could indicate the discussion about a potential protest, State 3 could reveal the spread of propaganda about the planned protests, while State 1 might be related to the organization of the protest.

with 300 samples. The runtimes of the online versions of the proposed models were consistently around 200 seconds when the number of training samples was varied from 20 to 300.

6.5. Case Study

A number of interesting events were predicted by the proposed approaches. In the two examples used for the case study presented in this section, we forecast a civil unrest event that involved Mexican teachers and occurred on March 31, 2013, and a flu outbreak in Texas at the end of November 2013 using STM-I. In the following discussion, we illustrate the topics, states, spatial burstiness, state transitions, and event-specific

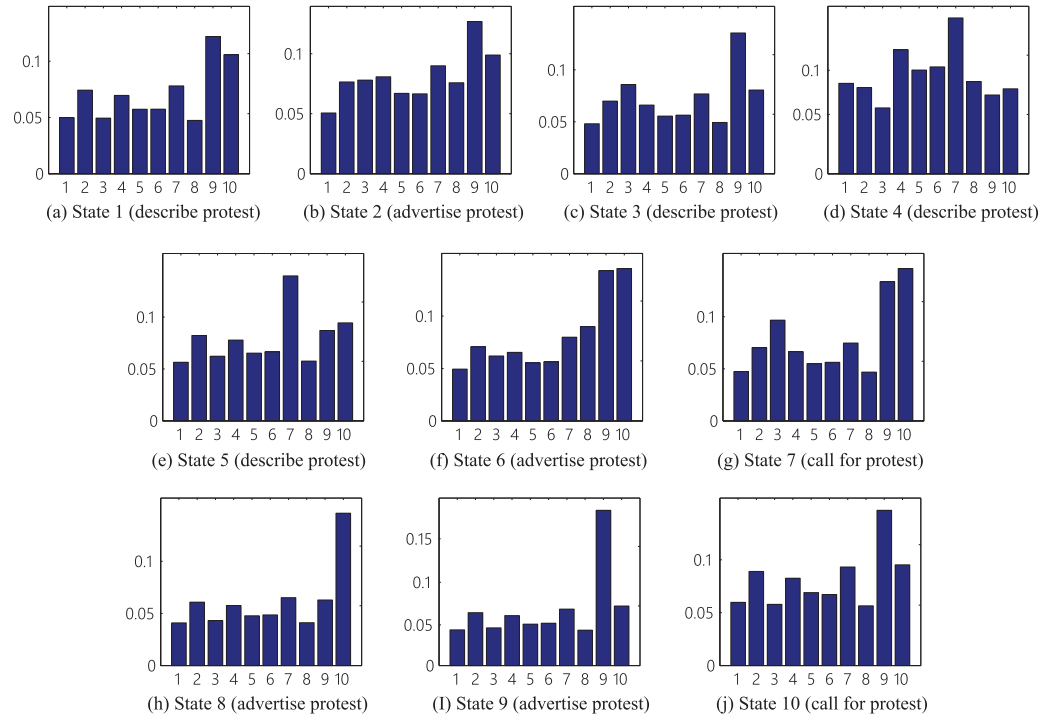


Fig. 21. Contexts of the 10 latent states indicating the development stages of events. For different stages of the developing protest potential, the distributions of topics changes. States 1, 3, 4 and 5 could indicate the discussion about a potential protest, States 2, 6, 8, and 9 could reveal the spread of propaganda about the planned protests, while States 7 and 10 might be related to the organization of the protest.

words identified by STM-I, which were validated by real-world civil unrest and flu outbreak events, verified by authorized news outlets.

6.5.1. Case Study I: Civil Unrest Event Forecasting for Mexico on March 31, 2013. Figure 18 illustrates the extracted topics for civil unrest-related and common words by the proposed new model STM-I. Different topics are clearly used at different stages of the unfolding civil unrest event. For example, Topics 1, 6, and 7, which highlight "march," "move," "plaza," and "takeover," generally refer to the description of a protest that is either happening or planned. Topic 9 tends to concentrate more on advertising a planned protest, with the top keywords here being "largest," "problem," and "protest." Topics 8 and 10 are related to the stage of "calling for protest," with the top keywords used being "please," "call for," and "come on." Topics 2 and 5 are more neutral, mainly consisting of background common words such as "reflects," "continue," and "checked."

Figure 19 demonstrates the distribution of topics in each state of the proposed STM-I. For example, State 1 and State 3 tend to focus on Topic 9 , paying less attention to other topics. This suggests that State 1 is likely to indicate the dissemination of the planned protest. State 5 highlights the call for protest because it leverages both Topics 9 and 10. This contrasts with the most influential topic in State 2, Topic 7, which indicates an emphasis on descriptions of ongoing or past events.

In addition to Figure 19, which illustrates the patterns when number of states $K = 5$, evaluations with more settings when $K = 3$ and $K = 10$ are also provided in Figures 20 and 21, respectively. As shown in Figure 20, when the number of states decreases from $K = 5$ to $K = 3$, the substantial information within the previous five latent states were

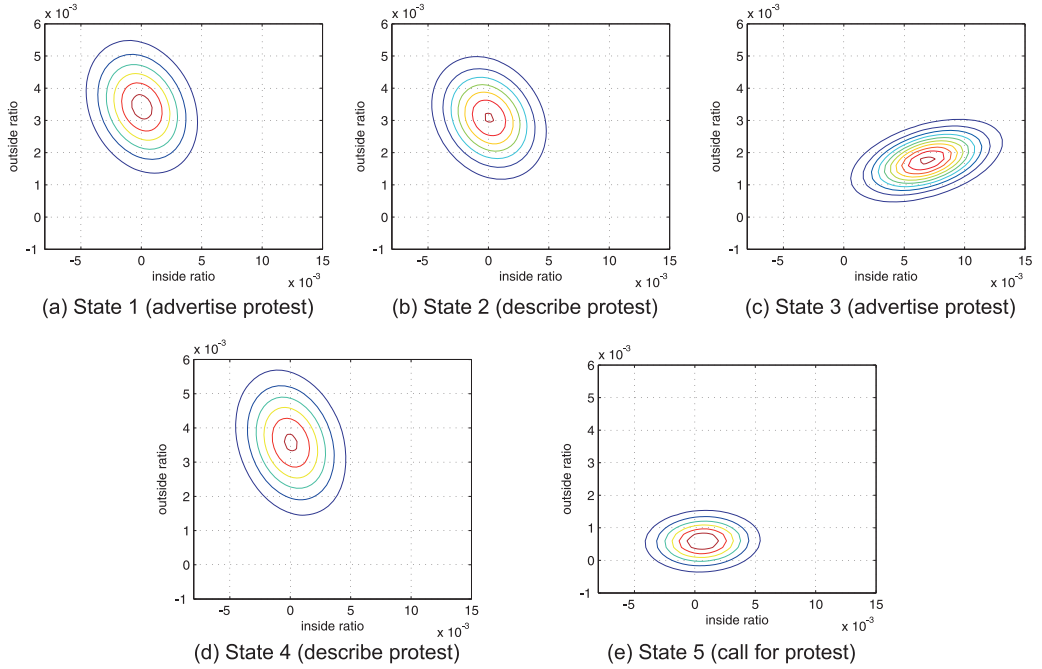


Fig. 22. Spatial burstiness patterns of the five latent states indicating the development stages of a civil unrest event. States 1, 2, and 3 reveal that the event-related tweet percentage inside the location is similar to that outside the location. States 3 and 5 show that the event-related tweet percentage inside the location is much larger than that outside the location, which indicates a potential burstiness in the location.

compressed into these three states. Specifically, State 3 in Figure 20 preserves some pattern of the State 3 in Figure 19; State 2 in Figure 20 is more relevant to State 2 in Figure 19; State 1 in Figure 20 is likely to preserve the major patterns in States 1, 4, and 5 in Figure 19.

In Figure 21, the latent state patterns within the five states are extended and diversified into 10 latent states. Specifically, States 1, 2, 6, 7, and 10 are more relevant to States 4, and 5 in Figure 19, which seems also derived from State 8 in Figure 21; State 5 corresponds to State 2 in Figure 19; States 3 and 9 could be expanded from State 1; State 4 potentially mirrors State 2 in Figure 19.

Figure 22 shows the spatial burstiness in terms of the inside and outside ratios for each state for civil unrest events. In these subplots, each state is illustrated as a bivariate Gaussian whose means are the average inside and outside ratios of the location of the current tweet sequence, and its variance reflects the degree by which the ratios spread out and how the inside and outside ratios are related to each other. For example, States 1, 2, and 4 tend to be similar because the means of their outside ratios are larger than those of their inside ratios, and their inside and outside ratios are likely to also be negatively correlated, as shown in Figures 22(a), 22(b), and 22(d). On the other hand, States 3 and 5 are more likely to have larger inside ratios than outside ratios, and their inside and outside ratios are basically positively correlated. This generally indicates that burstiness occurred inside the location.

The developing progress of an event (as described in Figure 23(c)) is reflected in the transitions among hidden states identified by STM-I, as shown in Figure 23(a). This progress is validated by the ground-truth descriptions from news reports, as shown in Figure 23(b).

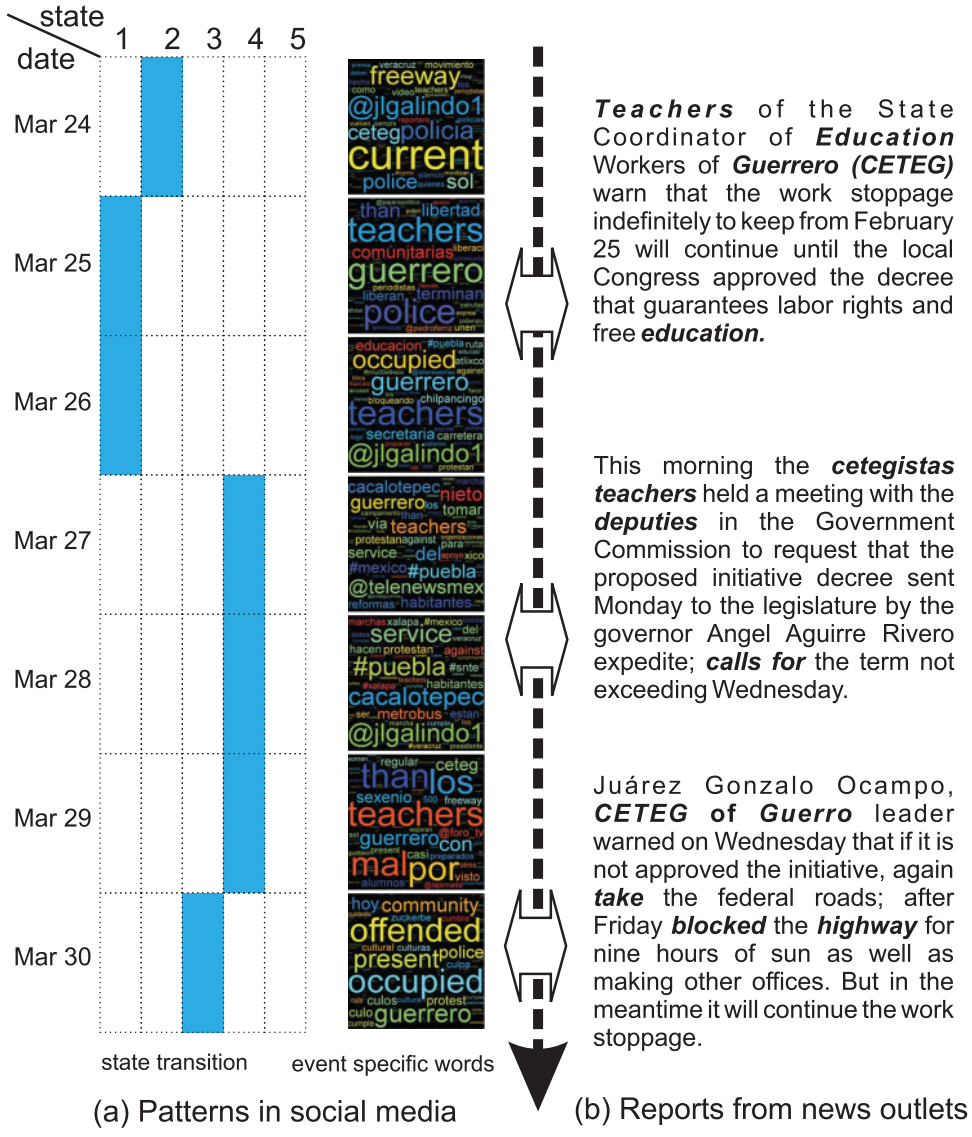


Fig. 23. Comparison of the event development progression discovered on microblogs with the authorized reports by news outlets. The state transition on the left of (a) demonstrates the event stages conceptualized by our model. On the right of (a), the word clouds show that the keywords discovered in the microblogs are a good match for the bold keywords in the news reports in (b). The effective modeling of the development progression finally leads to the accurate prediction of the occurrence of the events described in (c).

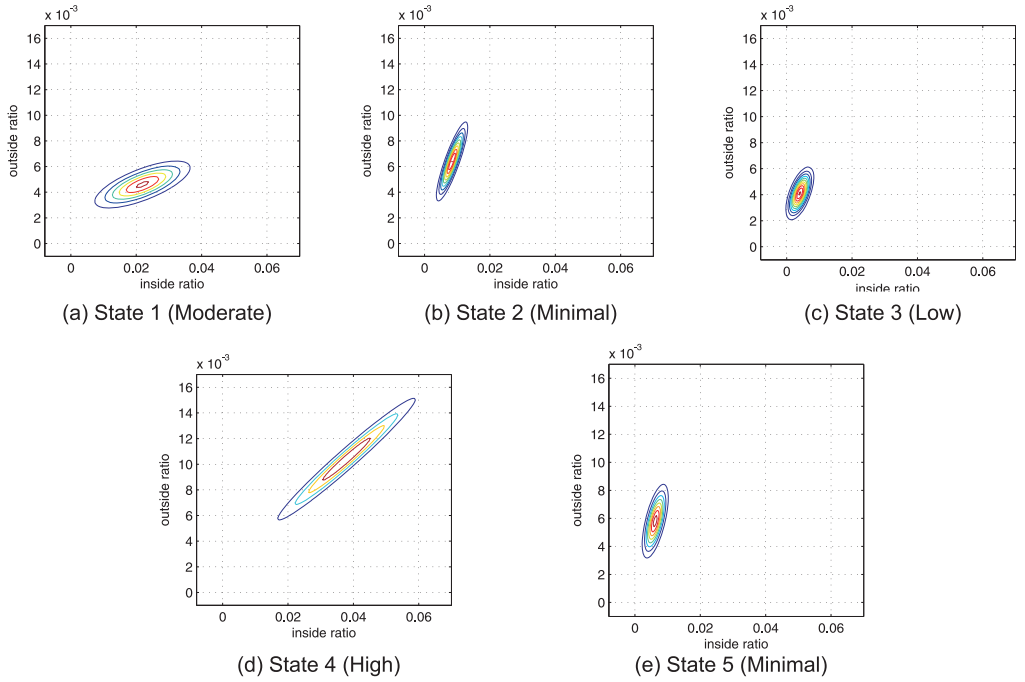


Fig. 24. Spatial burstiness patterns of the five latent states, indicating the development stages of flu events. States 2, 3, and 5 reveal that the event-related tweet percentage inside the location is similar to that outside the location. States 1 and 4 show that the event-related tweet percentage inside the location is much larger than that outside the location, which indicates a potential flu burstiness in that location.

As shown in Figure 23(a), the state transition is State 2 → State 1 → State 1 → State 4 → State 4 → State 4 → State 3. By referring to Figures 18, 19, and 22, this state transition indicates a potential development sequence of "planning → advertising → calling." Figure 23(a) also illustrates the identified event-specific words for each date. Figure 23(b) demonstrates that the identified event-specific words match the ground-truth from the news reports, especially for keywords such as "Guerrero" (protest location and protest target), "teacher" (protest initiator), and "occupy" (protest action). Therefore, the case study confirms that the topics, states, spatial-burstiness, and state transitions identified by our approach are indeed effective and accurate and have practical meanings that match the ground truth obtained from the authorized news outlets.

6.5.2. Case Study II: Flu Outbreak Event Forecasting for Texas, USA, November 24-30, 2013. Figure 24 shows the spatial burstiness in terms of the inside and outside ratios for each latent state. Here, as in Figure 22, in each subplot a state is illustrated as a bivariate Gaussian whose means are the average inside and outside ratios of the location of the current tweet sequence and its variance reflects the degree to which the ratios spread out and the inside and outside ratios are related to each other. For example, States 2, 3, and 5 tend to be more similar because the means of their outside ratios are larger than those of their inside ratios; low inside and outside ratios indicate low influenza activity. The inside and outside ratios are likely to be positively correlated, as shown in Figures 24(b), 24(c), and 24(e). On the other hand, States 1 and 4 are more likely to have larger inside ratios than outside ratios. The much larger inside ratio indicates a strong flu-related signal in social media at the current location, which generally suggests that there is or will be burstiness occurring inside the location.

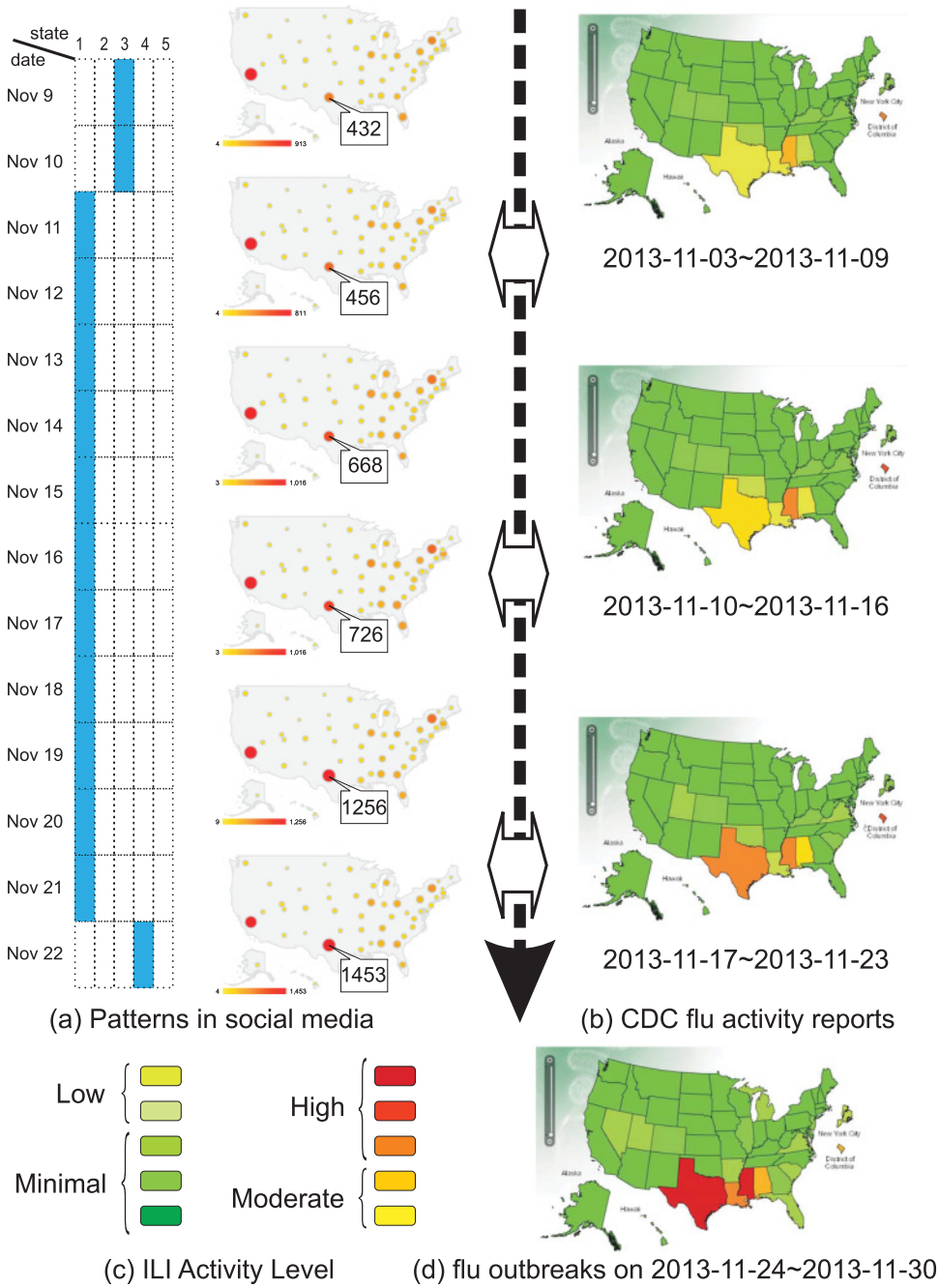


Fig. 25. Comparison of flu event development progression discovered on microblogs with the authorized reports from the CDC. The state transition on the left of (a) demonstrates the event stages conceptualized by our model. On the right of (a), the map shows that the increase of flu-related tweets in Texas is a good match for the rapidly increasing reports of flu activity in Texas, as shown in (b). The effective modeling of the development progression finally leads to the accurate prediction of the occurrence of the flu outbreaks, illustrated in (c).

The left part of Figure 25(a) shows the latent state transition, while 25(b) shows the distribution of flu-related tweets across the whole country. The initial latent state is State 3 on November 9 and 10, which then transfers to State 1 on November 11-21, and finally goes to State 4 on November 22. By referring to Figure 24, we know that State 3 indicates a moderate inside ratio and low outside ratio, while State 1 indicates a high inside ratio and low outside ratio. By modeling this state transition and the flu-related tweets' spatial distribution, our model forecast a potential development of "flu outbreaks" in Texas in the following week. Figure 25(b) shows the flu activity level identified by the authorities, in this case the CDC flu reports. It clearly demonstrates the upgrading of the flu activity in Texas from November 3, 2013 to November 11, 2013, and predicts a flu outbreak for the following week, which is consistent with the pattern identified and modeled through social media.

7. CONCLUSION

This article has presented a novel model for spatiotemporal event forecasting in Twitter. The new generative approach uncovers the underlying development of events by jointly considering the structural semantics and the spatiotemporal burstiness of Twitter streams. Both batch and online-based inference algorithms were developed to optimize the model parameters. Utilizing the trained model, the alignment likelihood of tweet sequences was calculated by dynamic programming. Extensive empirical testing demonstrated the effectiveness of the new approach by comparing it with those of five representative methods. In future work, we plan to extend our approach to other applications, such as forecasting outbreaks of other diseases and local events such as road congestion.

APPENDIX

A.1. Batch Parameter Optimization Algorithm

To find the best parameters for both models, the Expectation Maximization (EM) algorithm can be extended to compute the parameters for modeling structural text and space-time outbreaks.

The standard steps of the Baum-Welch (BW) algorithm [Chen et al. 2005] is applied to calculate the expectation probability $\mathrm{E}[p(Z_{s,t} = k)]$ that the observation of the location s and time t is under the latent state k.

Given the expectations $\mathrm{E}[p(Z_{s,t} = k)]$, the expected count of time intervals that the observations are under the latent state k is calculated based on the following equation:

$$\hat{N}_{l,k} = \sum_{s \subseteq D_l} \sum_{t} \mathrm{E}[p(Z_{s,t} = k)], \tag{18}$$

where $s \in D_l$ signifies that the sequence s belongs to the tweets in location l.

When using the Gaussian distribution to model the space-time burstiness, the maximum likelihoods of the mean and variance of the Gaussian distributed space-time burstiness modeling are computed as:

$$\hat{\mu}_{l,k} = \frac{\sum_{s \subseteq D_l} \sum_{t} \mathrm{E}[p(Z_{s,t} = k)] \cdot (r^{in}_{s,t}, r^{out}_{s,t})}{\hat{N}_{l,k}}, \tag{19}$$

where $(r^{in}_{s,t}, r^{out}_{s,t})$ is the vector observation of the bi-variate Gaussian.

$$\hat{\Sigma}_{l,k} = \frac{\sum_{s \subseteq D_l} \sum_{t} \mathrm{E}[p(Z_{s,t} = k)] \left(\hat{\mu}_{l,k} - (r^{in}_{s,t}, r^{out}_{s,t})\right)^2}{\hat{N}_{l,k}}. \tag{20}$$

The posteriors of the mean and variance of the Gaussian distributed space-time burstiness modeling are computed as:

$$\mu_{l,k} = (\beta_0\mu_0 + \hat{N}_{l,k}\hat{\mu}_{l,k})/(\beta_0 + \hat{N}_{l,k}) \tag{21}$$

$$\Sigma_{l,k} = \frac{\Lambda_0 + \hat{\Sigma}_{l,k}}{\nu_0 + 3} + \frac{\beta_0\hat{N}_{l,k}(\hat{\mu}_{l,k} - \mu_0)(\hat{\mu}_{l,k} - \mu_0)^T}{(\beta_0 + \hat{N}_{l,k})(\nu_0 + 3)}. \tag{22}$$

When using Poisson-distributed space-time burstiness modeling, Equations (19–22) are replaced by Equations (23) and (28), as shown in the following:

The weighted means of the domain-related counts inside location l under latent state k is calculated as:

$$\hat{\lambda}^{in}_{c,k,l} = \sum_{s\in D_l}\sum_{t} c^{in}_{s,t} \cdot \mathrm{E}[p(Z_{s,t} = k)]/\hat{N}_{l,k}. \tag{23}$$

The weighted means of the base counts inside location l under latent state k is calculated as:

$$\hat{\lambda}^{in}_{b,k,l} = \sum_{s\in D_l}\sum_{t} b^{in}_{s,t} \cdot \mathrm{E}[p(Z_{s,t} = k)]/\hat{N}_{l,k.} \tag{24}$$

The weighted means of the domain-related counts outside location l under latent state k is calculated as:

$$\hat{\lambda}^{out}_{c,k,l} = \sum_{s\in D_l}\sum_{t} c^{out}_{s,t} \cdot \mathrm{E}[p(Z_{s,t} = k)]/\hat{N}_{l,k.} \tag{25}$$

The weighted means of the base counts outside location l under latent state k is calculated as:

$$\hat{\lambda}^{out}_{b,k,l} = \sum_{s\in D_l}\sum_{t} b^{out}_{s,t} \cdot \mathrm{E}[p(Z_{s,t} = k)]/\hat{N}_{l,k.} \tag{26}$$

Therefore, the posteriors of the means of the ratios of the domain-related tweets inside and outside, respectively, of the location l under the latent state k are calculated using the following two equations:

$$\lambda^{in}_{k,l} = \frac{(\alpha^{in} - 1) + \hat{\lambda}^{in}_{c,k,l}}{\beta^{in} + \hat{\lambda}^{in}_{b,k,l}} \tag{27}$$

$$\lambda^{out}_{k,l} = \frac{(\alpha^{out} - 1) + \hat{\lambda}^{out}_{c,k,l}}{\beta^{out} + \hat{\lambda}^{out}_{b,k,l}}. \tag{28}$$

In the sequence s and latent state k, the expectations of the count of a word w that has been identified as being specific to the unique event is calculated as:

$$g_{s,k,w} = \sum_{t}^{T} N_{s,t,w}\frac{\mathrm{E}[p(Z_{s,t} = k)] \cdot \Psi_{k,1} \cdot \theta^{R}_{s,t,w}}{\Psi_{k,1}\theta^{R}_{s,t,w} + \Psi_{k,2}\sum_j \Phi_{k,j}\theta^{B_j}_{w}}. \tag{29}$$

In the latent state k, the expectation value of the count of word w that is a common word under latent topic j is calculated as:

$$f_{k,j,w} = \sum_{s}^{S}\sum_{t}^{T} N_{s,t,w}\frac{\mathrm{E}[p(Z_{s,t} = k)] \cdot \Psi_{k,2} \cdot \Phi_{k,j}\theta^{B_j}_{w}}{\Psi_{k,1}\theta^{R}_{s,t,w} + \Psi_{k,2}\sum_j \Phi_{k,j}\theta^{B_j}_{w}}. \tag{30}$$

Among the words corresponding to specific events in the sequence s and latent state k, the likelihood for word w is:

$$\theta^{R}_{s,k,w} = \frac{g_{s,k,w}}{\sum_x g_{s,k,x}}. \tag{31}$$

Among the common words under the topic j in the sequence s and latent state k, the likelihood of word w is:

$$\theta^{B_j}_{w} = \frac{\sum_k f_{k,j,w}}{\sum_k \sum_w f_{k,j,w}}. \tag{32}$$

In the latent state k, the likelihood that a word corresponds to a specific event is:

$$\Psi_{k,1} = \frac{\sum_s \sum_w g_{s,k,w}}{\sum_s \sum_w g_{s,k,w} + \sum_w \sum_j f_{k,j,w.}} \tag{33}$$

In the latent state k, the likelihood that a word is a common word is:

$$\Psi_{k,2} = \frac{\sum_w \sum_j f_{k,j,w}}{\sum_s \sum_w g_{s,k,w} + \sum_w \sum_j f_{k,j,w}}. \tag{34}$$

In the latent state k among all the common words, the likelihood that a word is included under topic j is:

$$\Phi_{k,j} = \frac{\sum_w f_{k,j,w}}{\sum_j \sum_w f_{k,j,w}}. \tag{35}$$

The prior likelihood of the latent state k is:

$$\pi_k = \frac{\sum_s \mathrm{E}[p(Z_{s,1} = k)]}{\sum_s \sum_i \mathrm{E}[p(Z_{s,1} = i)]}. \tag{36}$$

By iteratively executing the E-step and the M-step, the model parameters and the latent variables are continuously updated until convergence is achieved. The model parameters are optimized and the likelihood in Equation (10) is maximized.

A.2. Stochastic E-Step

A.2.1. STM-I. $E^{\hat{\mu}}_i$, $E^{\hat{\Sigma}}_i$, E^{g}_i, and E^{f}_i based on the current sequence s_i can be obtained, as shown in Equation (37).

$$\begin{aligned}
E^{\hat{\mu}}_i &= \frac{\sum_t \mathrm{E}[p(Z_{s_i,t} = k)](r^{in}_{s_i,t}, r^{out}_{s_i,t})}{\hat{N}_{l,k,i}} \\
E^{\hat{\Sigma}}_i &= \frac{\sum_t \mathrm{E}[p(Z_{s_i,t} = k)]\left(\hat{\mu}_{l,k} - (r^{in}_{s_i,t}, r^{out}_{s_i,t})\right)^2}{\hat{N}_{l,k,i}} \\
E^{g}_i &= \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t} = k)] \cdot \Psi_{k,1} \cdot \theta^{R}_{s_i,t,w}}{\Psi_{k,1}\theta^{R}_{s_i,t,w} + \Psi_{k,2} \sum_j \Phi_{k,j}\theta^{B_j}_{w}} \\
E^{f}_i &= \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t} = k)] \cdot \Psi_{k,2} \sum_j \Phi_{k,j}\theta^{B_j}_{w}}{\Psi_{k,1}\theta^{R}_{s_i,t,w} + \Psi_{k,2} \sum_j \Phi_{k,j}\theta^{B_j}_{w}}
\end{aligned} \tag{37}$$

The stochastic approximation of the statistics update is presented in Equation (38):

$$\begin{aligned}
\hat{\mu}_{l,k,i} &= (1-\gamma_i)\cdot \hat{\mu}_{l,k,i-1} + \gamma_i \cdot E_i^{\hat{\mu}} \\
\hat{\Sigma}_{l,k,i} &= (1-\gamma_i)\cdot \hat{\Sigma}_{l,k,i-1} + \gamma_i \cdot E_i^{\hat{\Sigma}} \\
g_{s,k,w,i} &= (1-\gamma_i)\cdot g_{s,k,w,i-1} + \gamma_i \cdot E_i^{g} \\
f_{k,j,w,i} &= (1-\gamma_i)\cdot f_{k,j,w,i-1} + \gamma_i \cdot E_i^{g}
\end{aligned} \tag{38}$$

The model parameters θ^B, θ^R, Ψ, and Φ need to be initialized. For each latent state $k \in K$, the common-word language model $\theta^B \in \mathbb{R}^{K\times J\times N}$ is initialized by maximizing the likelihood of a mixture multinomial model. Specifically,

$$p(w) = \sum_j^J p(j) \prod_n^N p(W_n|j), \tag{39}$$

where $p(W_n|j) = \theta_{n,k}^{B_j}$. By maximizing the log likelihood of $p(w)$, the language model θ^B is determined. Other parameters θ^R, Ψ, and Φ are initialized with uniform distributions.

A.2.2. STM-S. Specifically, the conditional expectations $E_i^{\hat{\mu}}$, $E_i^{\hat{\Sigma}}$, E_i^{g}, and E_i^{f} based on the currency sequence s_i are calculated.

$$\begin{aligned}
E_i^{\hat{\lambda}_c^{in}} &= \sum_t c_{s_i,t}^{in} \cdot \mathrm{E}[p(Z_{s_i,t}=k)]/\hat{N}_{l,k,i} \\
E_i^{\hat{\lambda}_b^{in}} &= \sum_t b_{s_i,t}^{in} \cdot \mathrm{E}[p(Z_{s_i,t}=k)]/\hat{N}_{l,k,i} \\
E_i^{\hat{\lambda}_c^{out}} &= \sum_t c_{s_i,t}^{out} \cdot \mathrm{E}[p(Z_{s_i,t}=k)]/\hat{N}_{l,k,i} \\
E_i^{\hat{\lambda}_b^{out}} &= \sum_t b_{s_i,t}^{out} \cdot \mathrm{E}[p(Z_{s_i,t}=k)]/\hat{N}_{l,k,i} \\
E_i^{g} &= \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)]\cdot \Psi_{k,1}\cdot \theta_{s_i,t,w}^R}{\Psi_{k,1}\theta_{s_i,t,w}^R + \Psi_{k,2}\sum_j \Phi_{k,j}\theta_w^{B_j}} \\
E_i^{f} &= \sum_t^T N_{s_i,t,w} \frac{\mathrm{E}[p(Z_{s_i,t}=k)]\cdot \Psi_{k,2}\sum_j \Phi_{k,j}\theta_w^{B_j}}{\Psi_{k,1}\theta_{s_i,t,w}^R + \Psi_{k,2}\sum_j \Phi_{k,j}\theta_w^{B_j}}
\end{aligned} \tag{40}$$

The stochastic approximation of the statistics update is presented in Equation (41).

$$\begin{aligned}
\hat{\lambda}_{c,l,k,i}^{in} &= (1-\gamma_i)\cdot \hat{\lambda}_{c,l,k,i-1}^{in} + \gamma_i \cdot E_i^{\hat{\lambda}_c^{in}} \\
\hat{\lambda}_{b,l,k,i}^{in} &= (1-\gamma_i)\cdot \hat{\lambda}_{b,l,k,i-1}^{in} + \gamma_i \cdot E_i^{\hat{\lambda}_b^{in}} \\
\hat{\lambda}_{c,l,k,i}^{out} &= (1-\gamma_i)\cdot \hat{\lambda}_{c,l,k,i-1}^{out} + \gamma_i \cdot E_i^{\hat{\lambda}_c^{out}} \\
\hat{\lambda}_{b,l,k,i}^{out} &= (1-\gamma_i)\cdot \hat{\lambda}_{b,l,k,i-1}^{out} + \gamma_i \cdot E_i^{\hat{\lambda}_b^{out}} \\
g_{s,k,w,i} &= (1-\gamma_i)\cdot g_{s,k,w,i-1} + \gamma_i \cdot E_i^{g} \\
f_{k,j,w,i} &= (1-\gamma_i)\cdot f_{k,j,w,i-1} + \gamma_i \cdot E_i^{g}
\end{aligned} \tag{41}$$

ACKNOWLEDGMENTS

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

REFERENCES

Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. 2011. Predicting flu trends using Twitter data. In *Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. 702–707.

Charu C. Aggarwal and Karthik Subbian. 2012. Event detection in social streams. In *SIAM Data Mining* 2012. SIAM, 624–635.

Marta Arias, Argimiro Arratia, and Ramon Xuriguera. 2013. Forecasting with Twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)* 5, 1 (2013), 8.

Shea Bennett. 2014. Facebook, Twitter, Instagram, Pinterest, Vine, snapchat–social media stats 2014 [INFOGRAPHIC]. Retrieved November 8 (2014), 2014.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science* 2, 1 (2011), 1–8.

Olivier Cappé. 2011. Online expectation-maximisation. *Mixtures: Estimation and Applications* (2011), 1–53.

Olivier Cappé and Eric Moulines. 2009. On-line expectation–maximization algorithm for latent data models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71, 3 (2009), 593–613.

Deepayan Chakrabarti and Kunal Punera. 2011. Event summarization using tweets. In *ICWSM*.

Chien Chin Chen, Meng Chang Chen, and Ming-Syan Chen. 2005. LIPED: HMM-based life profiles for adaptive event detection. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2005)*. ACM, 556–561.

Freddy Chong Tat Chua and Sitaram Asur. 2013. Automatic summarization of events from social media. In *International Conference on Web and Social Media (ICWSM 2013)*. AAAI.

Matthew S. Gerber. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* 61 (2014), 115–125.

Jeremy Ginsberg, Matthew H. Mohebbi, Rajan S. Patel, Lynnette Brammer, Mark S. Smolinski, and Larry Brilliant. 2009. Detecting influenza epidemics using search engine query data. *Nature* 457, 7232 (2009), 1012–1014.

Jingrui He, Wei Shen, Phani Divakaruni, Laura Wynter, and Rick Lawrence. 2013. Improving traffic prediction with tweet semantics. In *International Joint Conference on Artificial Intelligence (IJCAI 2013)*. IJCAI, 1387–1393.

Fang Jin, Wei Wang, Liang Zhao, Edward Dougherty, Yang Cao, Chang-Tien Lu, and Naren Ramakrishnan. 2014. Misinformation propagation in the age of Twitter. *Computer* 47, 12 (2014), 90–94.

Martin Kulldorff. 1997. A spatial scan statistic. *Communications in Statistics-Theory and Methods* 26, 6 (1997), 1481–1496.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. 2010. What is Twitter, a social network or a news media? In *World Wide Web Conference (WWW 2010)*. IW3C2, 591–600.

Alex Lamb, Michael J. Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on Twitter. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (HLT-NAACL 2013)*. NAACL, 789–795.

Theodoros Lappas, Marcos R. Vieira, Dimitrios Gunopulos, and Vassilis J. Tsotras. 2012. On the spatiotemporal burstiness of terms. *Very Large Databases (VLDB 2012)* 5, 9 (2012), 836–847.

Micol Marchetti-Bowick and Nathanael Chambers. 2012. Learning for microblogs with distant supervision: Political forecasting with Twitter. In *Conference of the European Chapter of the Association for Computational Linguistics (ECACL 2012)*. ECACL, 603–612.

Sanjay Mehrotra. 1992. On the implementation of a primal-dual interior point method. *SIAM Journal on Optimization* 2, 4 (1992), 575–601.

Daniel B. Neill. 2012. Fast subset scan for spatial pattern detection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74, 2 (2012), 337–360.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *International Conference on Web and Social Media (ICWSM 2010)*. AAAI, 122–129.

Bohdan Pavlyshenko. 2013. Forecasting of events by tweet data mining. *arXiv preprint arXiv:1310.3499* (2013).

Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, Nathan Self, Rupinder Khandpur, Parang Saraf, Wei Wang, Jose Cadena, Anil Vullikanti, Gizem Korkmaz, and others. 2014. "Beating the news" with EMBERS: Forecasting civil unrest using open source indicators. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2014)*. ACM, 1799–1808.

Joshua Ritterman, Miles Osborne, and Ewan Klein. 2009. Using prediction markets and Twitter to predict a swine flu pandemic. In *Proceedings of the 1st International Workshop on Mining Social Media*, Vol. 9.

Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. In *World Wide Web Conference (WWW 2010)*. IW3C2, 851–860.

Alessio Signorini, Alberto Maria Segre, and Philip M. Polgreen. 2011. The use of Twitter to track levels of disease activity and public concern in the US during the influenza an H1N1 pandemic. *PloS One* 6, 5 (2011), e19467.

Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *International Conference on Web and Social Media (ICWSM 2010)*. AAAI, 178–185.

Xiaofeng Wang, Donald E. Brown, and Matthew S. Gerber. 2012a. Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. In *IEEE Intelligence and Security Informatics Conference (ISI 2012)*. IEEE, 36–41.

Xiaofeng Wang, Matthew S. Gerber, and Donald E. Brown. 2012b. Automatic crime prediction using events extracted from Twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 231–238.

Jianshu Weng and Bu-Sung Lee. 2011. Event detection in Twitter. *International Conference on Web and Social Media (ICWSM 2011)*. AAAI, 401–408.

Liang Zhao, Feng Chen, Jing Dai, Ting Hua, Chang-Tien Lu, and Naren Ramakrishnan. 2014. Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. *PloS One* 9, 10 (2014), e110206.

Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2015a. Dynamic theme tracking in Twitter. In *Proceedings of the IEEE International Conference on Big Data*. IEEE, 561–570.

Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2015b. Multi-task learning for spatio-temporal event forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1503–1512.

Liang Zhao, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2016. Hierarchical incomplete multi-source feature learning for spatiotemporal event forecasting. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (SIGKDD 2016)*. ACM, 2085–2094.

Received November 2015; revised May 2016; accepted September 2016