Spatiotemporal Event Forecasting in Social Media

Liang Zhao * Feng Chen † Chang-Tien Lu * Naren Ramakrishnan *

Abstract

Event forecasting in Twitter is an important and challenging problem. Most existing approaches focus on forecasting temporal events (such as elections and sports) and do not consider spatial features and their underlying correlations. In this paper, we propose a generative model for spatiotemporal event forecasting in Twitter. Our model characterizes the underlying development of future events by jointly modeling the structural contexts and spatiotemporal burstiness. An effective inference algorithm is developed to train 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 demonstrated the effectiveness of our proposed approach.

1 Introduction

Microblogs like Twitter and Weibo are experiencing a rapid increase as real-time "sensors" for society [8]. Hundreds of millions of users collectively post millions of tweets every hour, discussing a variety of content ranging from everyday feelings to comments about social events. Compared to traditional media, Twitter has the following significant characteristics: 1) *Timeliness of messages:* Unlike traditional media that take hours or days to publish, tweets can be posted instantly utilizing portable mobile devices; 2) *Ubiquity of social sensors:* Tweets reflect the public's mood and trends, which could be the determinants of future social events; and 3) *Availability of geo-information:* Twitter users provide rich location information in profiles, texts, and geotags. Recent research has revealed the power of Twitter for event forecasting [19, 21]; Twitter and other social media have been recognized for playing a key role in events such as the "Arab Spring" and the Mexican presidential election protests [15, 21]. Figure 1 depicts activities on Twitter that causally preceded the Mexico City protests. Both the content and spatiotemporal burstiness of the protest-related tweets reveal the escalation of societal discontent pertaining to this controversial election, from complaining through planning and advertising, to the final protest event. However, existing event forecasting models in Twitter generally focus on temporal events whose geo-locations are not available or irrelevant to the prediction task (e.g., elections [19] and sports [14]). Comparatively little attention has been paid to forecasting spatiotemporal events.

A spatiotemporal event is mainly relevant to the tweets posted within a certain geographical neighborhood. 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: 1) *Capturing spatiotemporal dependencies.* A spatial event may influence not only the location and time, but also its geographical and temporal neighborhood. The influence strength and pattern may vary in different development stages for different events; 2) *Modeling mixed type observations.* An event involves the temporal evolution of spatially distributed tweets and their semantics. Joint consideration of these heterogeneous and multi-dimensional data is crucial; and 3) *Utilizing prior geographical knowledge.* Spatiotemporal events in crucial domains usually have rich historical records. Different geo-locations may feature their inherent and distinct event frequencies that can be integrated into a predictive model to improve its forecasting accuracy. For example, the historical crime rates in different cities can help forecast the probability of future crime events.

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

- **A novel generative model for spatial event forecasting.** For spatial event forecasting in Twitter, we propose an enhanced hidden Markov model

*Virginia Tech.
†SUNY Albany.

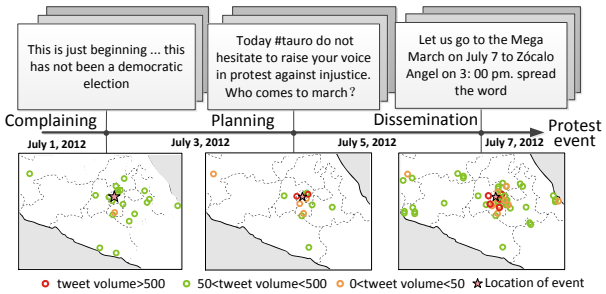


Figure 1: Twitter predicts a presidential election protest.

(HMM) that characterizes the transitional process of event development by jointly considering the time-evolving context and space-time burstiness of Twitter streams.

- **An effective algorithm for model parameter inference.** The model inference is formalized as the maximization of a posterior that is analytically tractable. This problem is effectively solved by our proposed EM-based algorithm.

- **A new sequence likelihood calculation method.** To handle the noisy nature of tweet content, words exclusive to a single event are identified by a language model that is optimized by a dynamic programming algorithm to achieve accurate sequence likelihood calculation.

- **Extensive experimental performance evaluations.** The proposed method outperforms existing methods by 38% and 67% on two different datasets. Sensitivity analyses reveals the impact of the parameters on the new method's performance.

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

2 Related Work

Current researches into the analysis of Twitter-based social events can be categorized into two main types: 1) event detection; and 2) event forecasting. These are considered in turn below.

Event detection: A large body of work focuses on the detection of ongoing events [2, 10, 17, 18, 22]. They utilize tweets as real-time and ubiquitous social sensors to promptly discover new events occurring. Methods based on spatial bursts use a classifier to extract topic-related tweets and then examine their spatial burstiness, in applications such as detecting earthquakes [17] and disease outbreaks [18]. Methods based on temporal bursts detect the temporal patterns of Twitter streams utilizing techniques such as wavelet analysis [22] or temporal clustering [2]. Spatiotemporal methods aim to detect bursts in both time and space [10]. However, these event detection approaches can only uncover events after they have occurred and are unable to forecast future events because they all 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. A variety of applications have been explored, including elections [13, 19], disease outbreaks [1, 16], stock market movements [3, 4], politics [11], box office ticket sales [3], the Olympic games [14], crime [21], and traffic conditions [7]. These papers can be categorized into four types based on the complexity of models utilized: 1) Linear regression model. This thread maps simple predictive features such as sentiment score or tweet volume to the occurrence of future events [3, 4, 7, 13]; 2) Nonlinear models. This thread incorporates more informative features such as semantic topics by utilizing methods such as support vector machines and logistic regression [16, 21]; 3) Time series-based methods. This thread considers the temporal correlation of relevant features such as tweet volume by adopting approaches such as autoregressive modeling [1]; and 4) Domain-specific approaches. This thread is designed to solve particular problems and may not be applicable to other application domains. For example, Pavlysehko [14] 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 [11] 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 [6] proposed a predictor for spatiotemporal events by utilizing historical event counts and topics, but do not consider temporal evolution and dependencies, while Wang et al. [20] developed a model to characterize and predict spatio-temporal criminal incidents, but their model requires the availability of demographic information.

This paper proposes a spatiotemporal event forecasting method that characterizes the evolutionary pattern of both spatial burstiness and structural contexts. By modeling geographical priors effectively, our ap-

Table 1: 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.

proach can sufficiently leverage historical prior knowledge and can be effectively applied to new locations.

3 Model

This section first elaborates upon the generative process of the proposed model, before describing the procedure to estimate its parameters.

3.1 Generative Process First, the spatiotemporal event forecasting problem is formalized. Then, our new generative model is described in detail, including the space-time burstiness module and structural tweet content module.

3.1.1 Problem Formulation The notations used in the paper are introduced in Table 1. As demonstrated in Figure 1, to accurately forecast spatiotemporal events it is crucial to be able to characterize their underlying development before the occurrence by utilizing relevant tweet observations. An enhanced hidden Markov model is proposed here to characterize the underlying development of events.

Given a sequence of observations (i.e. symbols) 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 fit the sequence of observations.

However, a standard HMM is limited to simple symbol observations and will thus face several challenges in our case as the observation does not consist of a single symbol but rather all the domain-related tweets in each time step. Further, 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 with tweet observations.

This problem therefore requires several important enhancements to the standard HMM. First, instead of a single symbol, each observation encompasses 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 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.

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 in 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 $Z = \{1, 2, \cdots, K\}$, with tweet sequence $s \subseteq D_l$ being the observations generated by the latent states.

3.1.2 Space-time burstiness modeling. Given a tweet sequence $s \subseteq D_l$ in location l, denote $c_{s,t}^{in}$ as the count of domain-related tweets inside location l at time t, and $c_{s,t}^{out}$ as the count outside this location; Denote $b_{s,t}^{in} = |D_{l,t}|$ as the total tweet count inside the location l at time step t, and $b_{s,t}^{out}$ as that outside this location. $r_{s,t}^{in} = c_{s,t}^{in}/b_{s,t}^{in}$ and $r_{s,t}^{out} = c_{s,t}^{out}/b_{s,t}^{out}$ are the *inside ratio* and the *outside ratio* and are, respectively, the proportions of the domain-related tweets inside and outside the location l. Hence, the spatial burstiness pattern surrounding the location l is jointly characterized by $r_{s,t}^{in}$ and $r_{s,t}^{out}$. For example, 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:

$$r^{in}_{s,t}, r^{out}_{s,t} \sim \mathcal{N}(r^{in}_{s,t}, r^{out}_{s,t} | \mu_{l,k}, \Sigma_{l,k}) \tag{3.1}$$

The advantages of a bivariate Gaussian are two fold. First, its covariance matrix quantifies the different significance of the inside and outside ratios in characterizing the spatial burstiness. Second, the non-diagonal elements of the covariance matrix can also capture the relationship between the inside and outside ratios.

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

$$\mu_{l,k} \sim \mathcal{N}(\mu_{l,k} | \mu_0, \Sigma_{l,k}/\beta_0) \tag{3.2}$$

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:

$$\Sigma_{l,k} \sim \mathcal{IW}(\Sigma_{l,k} | \Lambda_0^{-1}, \nu_0) \tag{3.3}$$

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

3.1.3 Structural tweet content modeling. In domain-related tweet content, a word is deemed to belong to one of two categories: 1) Specific words: These are specific to a unique event, such as hashtags, hyperlinks, landmarks, and organization names; 2) Common words: These words are commonly used by different events, especially those that reflect the stage of development. In the kth latent state, the probability that a word belongs to either of the above two types is modeled by a Bernoulli distribution:

$$Y_{s,t,n} \sim Bern(Y_{s,t,n} | \Psi_k) \tag{3.4}$$

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 designates the words exclusive to the current observation sequence s at current time t:

$$W_{s,t,n} \sim Mult(W_{s,t,n} | \theta^R_{s,t}) \tag{3.5}$$

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.

$$X_{s,t,n} \sim Mult(X_{s,t,n} | \Phi_k) \tag{3.6}$$

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

$$W_{s,t,n} \sim Mult(W_{s,t,n} | \theta^{B_j}) \tag{3.7}$$

As shown in Figure 2, the generative process of the proposed model is:

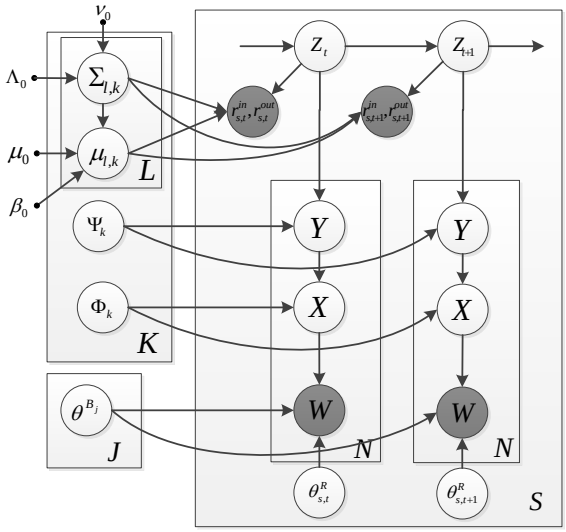


Figure 2: The plate notation of the proposed model.

- 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 $W_{s,t,n} \sim Mult(W_{s,t,n} | \theta^{B_j}, j = X_{s,t,n})$.

3.2 Parameter Estimation. Based on the generative process elaborated above, the proposed model 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 $\Theta_0 = \{\mu_0, \beta_0, \Lambda_0, \nu_0\}$. Their joint

distribution is expressed as follows:

$$
\begin{aligned}
(3.8)\quad & p(W, X, Y, Z, \mu, \Sigma, r^{in}, r^{out} | \pi, A, \Psi, \Phi, \theta, \Theta_0) \\
&= \prod_{s}^{S} p\left(Z_{s,1} | \pi\right) \cdot \prod_{s}^{S} \prod_{t=2}^{T} p\left(Z_{s,t} | Z_{s,t-1}, A\right) \\
&\quad \cdot \prod_{s}^{S} \prod_{t=1}^{T} \prod_{n}^{N} p\left(W_{s,t,n}, Y_{s,t,n}, X_{s,t,n} | Z_{s,t}, \Psi, \Phi, \theta\right) \\
&\quad \cdot \prod_{s}^{S} \prod_{t=1}^{T} p(r_{s,t}^{in}, r_{s,t}^{out} | \mu_l, \Sigma_l, Z_{s,t}) p(\mu_l, \Sigma_l | \Theta_0)
\end{aligned}
$$

where $\theta = \{\theta^B, \theta^R\}$. Thus, searching for the best setting of the model parameters is equivalent to the maximization of the logarithm of the joint distribution in Equation 3.8. This problem can be solved using an EM algorithm[1].

4 Spatiotemporal Event Forecasting

After the learning of model parameters, the spatiotemporal event forecasting is formalized as a sequence classification problem in this section, and an effective method for calculating the sequence likelihood is presented.

4.1 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 each has 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 one characterizes the 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 misclassifying the first class as the second class while e_2 is the cost of for 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 occurrs in location l; $p(C_2|l) = 1 - p(C_1|l)$ denotes the prior probability that no event occurs in location l; $p(s)$ is a constant and thus can be omitted. If the historical record for location l is not available, the above 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.

4.2 Calculation of sequence likelihood. In a standard HMM, dynamic programming methods such as the Viterbi algorithm [5] are typically utilized to calculate the likelihood of the 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 maximize the probability $p(s|C_i)$:

$$
(4.9)\quad p\left(s|C_i\right) = \max_{\{Z_t\}_t^T, \theta^R, n^R, n^B} \ln p(s, Z_1, \cdots, Z_T | C_i)
$$

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. $n^B = \{n^{B_j}_{s,t}\}^{S,T,J}_{s,t,j}$ is the number of the words explained by different latent topics. By introducing the notation ω_t such that $\omega_t \equiv \ln p(s, Z_1, \cdots, Z_t | C_i)$, Equation 4.9 can be solved by recursively calculating the following equation:

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

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 3.8 and omitting the constant term, the problem can be formalized as the following maximization problem:

$$
(4.11)\quad \max_{\theta^R_{s,t}, n^R_{s,t}, n^B_{s,t}} \sum_{i}^{V} n^R_{s,t,i} \cdot \log \theta^R_{s,t,i} + \sum_{i}^{V} \sum_{j}^{J} n^{B_j}_{s,t,i} \cdot \log \theta_i^{B_j}
$$

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

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

[1] Due to the space limitation, the full mathematical formulation of the EM update equations are provided here: https:www.dropbox.comsscxoq6y4fpffa7csupplementary_materials.pdf?dl=0

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 w.r.t. $\theta^R_{s,t,i}$ to 0, we obtain:

$$\frac{n^R_{s,t,i}}{\theta^R_{s,t,i}} + \gamma = 0 \tag{4.12}$$

where γ is the Lagragian multiplier for the first equality constraint. By utilizing the first two equality constraints in Equation 4.11, we can derive:

$$\theta^R_{s,t,i} = \frac{n^R_{s,t,i}}{\xi \cdot \Psi^R_{k,1}} \tag{4.13}$$

Substituting Equation 4.13 into Equation 4.11, we get

$$\max_{n^R_{s,t}, n^B_{s,t}} \sum_i^V n^R_{s,t,i} \cdot \log \frac{n^R_{s,t,i}}{\xi \cdot \Psi^R_{k,1}} + \sum_i^V \sum_j^J n^{B_j}_{s,t,i} \cdot \log \theta^{B_j}_i \tag{4.14}$$

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

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

Here, the objective function in Equation 4.14 is convex with respect to $n^R_{s,t}$ and $n^{B_j}_{s,t}$. Therefore, the global solution can be found by using a traditional numerical optimization method, such as the interior point method [12]. After $n^R_{s,t}$ and $n^{B_j}_{s,t}$ are optimized, $\theta^R_{s,t}$ can be calculated based on Equation 4.13. Finally, the maximization problem in Equation 4.9 is solved and thus the sequence likelihood can be calculated.

5 Experimental Evaluation

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

5.1 Experiment Design. This subsection presents the configuration of the datasets, the gold standard report for these event labels (as shown in Table 2), data processing, comparison methods, parameter settings, and performance metrics.

Datasets: For the analysis of civil unrest events forecasting, we collected 10 percent of raw Twitter data in Mexico through Datasift's Twitter collection engine from Jan 1, 2013 to Jun 1, 2013. The data from Jan 1, 2013 to Feb 28, 2013 was used as training, and the remaining was used 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 Jan 1, 2011 to Dec 31, 2013 in the United States. The data from Jan 1, 2011 to Jan 1, 2013 was used for training, and the rest was used for testing.

Gold Standard Report of Event Labels: The civil unrest forecasting results were validated against a labeled set called Gold Standard Report (GSR) that was exclusively provided by MITRE (see [15] for more details). The GSR was organized by manually harvesting civil unrest events reports from the 10 most significant news outlets[2] in Mexico and the world, as ranked by International Media and Newspapers[3]. There were totally 726 events during Jan 1, 2013 to Jun 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 of flu outbreaks were validated against the flu statistics reported by the Centers for Disease Control and Prevention (CDC). CDC publishes 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 4 ILI activity levels: minimal, low, moderate, and high, where the level "high" corresponds to salient flu outbreak and is considered for forecasting. There were in total 102 events during Jan 1, 2011 to Dec 31, 2013. A example of CDC flu outbreak event is: (STATE = "Michigan", COUNTRY = "United States", WEEK = "01-06-2013 to 01-12-2013").

Data Preprocessing: For the first data set, three labelers collectively labeled 20,906 tweets in both English and Spanish during Jun, 2012 to Feb, 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. Consequentially, the tweets were labeled into 6,793 positive and 14,113 negative, and the results used to train a linear SVM classifier. For the second data set, we utilized the labeled set in [9], and used these to train a linear SVM to identify tweets relevant to the flu. Both SVMs were generated

[2] They 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 2: 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

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). All the tweets without geotags were discarded.

Comparison Methods: Our proposed approach was compared with four representative methods and one baseline method. The *Autoregressive exogenous model (ARX)* [1] assumes that for each separate location, the count of future events is dependent on both the count of historical event and the tweet volume. When forecast, an output above "1" indicates that an event has occurred; otherwise no event is deemed to have occurred. *The linear regression (LinReg) model* [3, 4, 7, 13] 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* [21] event forecasting is treated as a classification problem. The input features are the proportions of latent topics extracted from the tweet texts coming from a specific location based on 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* [6] forecasts the event occurrence at a location by considering the historical event numbers and the tweet semantics. The set of input features is a combination of: 1) the historical event numbers spatially smoothed by KDE; and 2) the proportions of latent topics of tweet content. Finally, the *baseline* method considers the probability of historical event occurrence to be the probability of future event occurrence. Note that this baseline is also used as the prior in our proposed new approach.

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 10-fold cross-validation. The new method proposed here has 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 be 1; and the degrees of freedom ν_0 to the dimension of the vector $\mu_{k,l}$. The three tunable parameters are the misclassification cost ratio ε, the number of latent topics J and the number of latent states K and these were set as 10, 5, and 4, respectively, based on 10-fold cross-validation.

Performance Metrics: Three main performance metrics are considered: precision, recall, and F1-score. The reported forecasting alerts are 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 is matched to a true event if both the date and the location attributes are matched; otherwise, it is considered to be a false forecast. Note that because the time granularity of CDC flu outbreak labels is at week-level, it is considered as a match in time if the forecast date of an alert falls within the week of a true flu outbreak event.

5.2 Event Forecasting Results Table 3 presents the comparison between our approach and the five competing methods for the task of forecasting civil unrest and flu outbreak events.

For the civil unrest dataset, our approach achieved the best overall performance in precision, recall, and F1-score, outperforming the five comparison methods by 38% in F1-score and 7% in precision. This is likely because our approach considers the spatial burstiness as well as the tweet content, which is crucial for the forecasting of civil unrest events. KDE Logistic Regression achieved a F1-score that was 21% higher than those of ARX, LinReg, and LogReg due to its consideration of spatial dependencies. 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 as well as the spatial burstiness are important factors. The baseline method achieved the third best performance, indicating that it captured important historical event counts in different locations.

Table 3 demonstrates that our approach also consistently achieved the best performance in precision, recall, and F1-score, for the task of flu outbreak event forecasting. The F1-score of our approach was 63% higher

Table 3: Event forecasting results for the civil unrest and flu datasets

Dataset	Metric	Baseline	ARX	LR	LDA-LR	KDE-LDA-LR	Proposed algorithm
civil unrest data	precision	0.44	0.26	0.7	0.31	0.42	**0.75**
	recall	0.59	0.43	0.18	**0.7**	0.69	**0.7**
	f1-score	0.5	0.32	0.29	0.43	0.52	**0.72**
	runtime per day (sec)	$\mathbf{10^{-3}}$	$\mathbf{10^{-3}}$	0.001	0.005	0.005	0.32
flu data	precision	0.28	0.14	0.64	0.27	0.78	**0.83**
	recall	0.39	0.66	0.31	0.55	0.32	**0.69**
	f1-score	0.33	0.23	0.41	0.36	0.46	**0.75**
	runtime per day (sec)	$\mathbf{10^{-3}}$	0.001	0.01	0.02	0.03	2.1

than those of the five comparison methods. KDE LogReg achieved the second highest F1-score, suggesting the importance of considering spatial burstiness. 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 data set. LinReg, on the other hand, achieved a 41% higher F1-score in the flu data set than in the civil unrest data set, 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, achieved a poorer performance than in the civil unrest data set.

Our new approach and the five comparison methods all forecast next day events at the daily level. The running times of our approach were 0.32 seconds per day on the civil unrest dataset, and 2.1 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 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 forecasting accuracy of next day events makes this eminently practical for real-world applications.

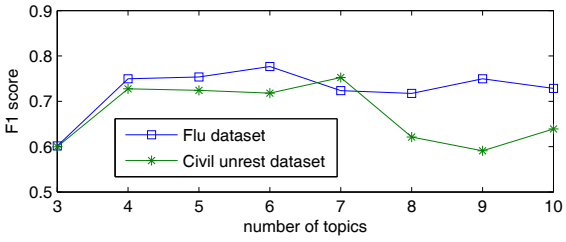


Figure 3: Sensitivity analysis on number of latent topics.

5.3 Sensitivity Analysis Figures 3 and 4 illustrate the impact of the number of latent states and the number of latent topics on the event forecasting performance. By varying the number of latent topics from 3 to 10, the F1-score on the civil unrest and the flu data sets varies between 0.6 and 0.8. When the number of latent states was raised from 2 to 10, the perturbation in the F1-scores remained between 0.7 to 0.8 for both datasets. This indicates that the performance is less sensitive to the number of latent states than the latent topics in the given value interval of parameters. For both parameters, the performance for low values is 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.

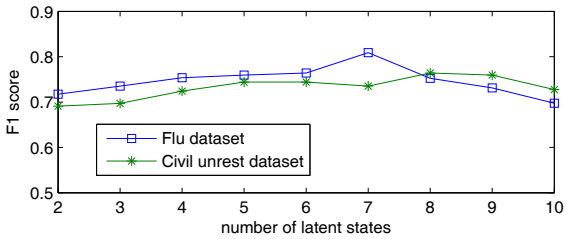
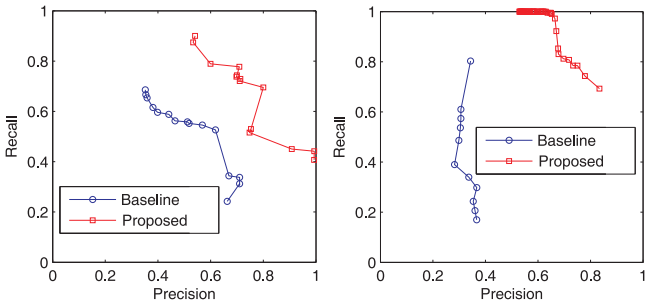


Figure 4: Sensitivity analysis on number of latent states.

For both the civil unrest and flu datasets, the precision-recall curves of the new approach and the baseline method are shown in Figure 5(a) and Figure 5(b). 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.

6 Conclusion

This paper presents 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. Exten-



(a) Civil unrest dataset (b) Flu dataset

Figure 5: Precision-recall curves on civil unrest and flu data.

sive empirical testing demonstrated the effectiveness of the new approach by comparing it with five representative methods. For future work, we plan to extend our approach to other applications, such as forecasting other disease outbreaks and local events such as road congestion.

Acknowledgements

This work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337; the US Government is authorized to reproduce and distribute reprints 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

[1] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu. Predicting flu trends using Twitter data. In *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 702–707, 2011.

[2] C. C. Aggarwal and K. Subbian. Event detection in social streams. In *SDM*, pages 624–635, 2012.

[3] M. Arias, A. Arratia, and R. Xuriguera. Forecasting with Twitter data. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(1):8, 2013.

[4] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[5] C. C. Chen, M. C. Chen, and M.-S. Chen. Liped: HMM-based life profiles for adaptive event detection. In *SIGKDD*, pages 556–561, 2005.

[6] M. S. Gerber. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems*, 61:115–125, 2014.

[7] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence. Improving traffic prediction with tweet semantics. In *IJCAI*, pages 1387–1393, 2013.

[8] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? In *WWW*, pages 591–600, 2010.

[9] A. Lamb, M. J. Paul, and M. Dredze. Separating fact from fear: Tracking flu infections on Twitter. In *HLT-NAACL*, pages 789–795, 2013.

[10] T. Lappas, M. R. Vieira, D. Gunopulos, and V. J. Tsotras. On the spatiotemporal burstiness of terms. *VLDB*, 5(9):836–847, 2012.

[11] M. Marchetti-Bowick and N. Chambers. Learning for microblogs with distant supervision: Political forecasting with Twitter. In *ECACL*, pages 603–612, 2012.

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

[13] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM*, 11:122–129, 2010.

[14] B. Pavlyshenko. Forecasting of events by tweet data mining. *arXiv preprint arXiv:1310.3499*, 2013.

[15] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, R. Khandpur, P. Saraf, W. Wang, J. Cadena, A. Vullikanti, G. Korkmaz, et al. 'beating the news' with embers: Forecasting civil unrest using open source indicators. *arXiv preprint arXiv:1402.7035*, 2014.

[16] J. Ritterman, M. Osborne, and E. Klein. Using prediction markets and Twitter to predict a swine flu pandemic. In *1st international workshop on mining social media*, volume 9, 2009.

[17] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW*, pages 851–860, 2010.

[18] A. Signorini, A. M. Segre, and P. M. Polgreen. 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):e19467, 2011.

[19] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[20] X. Wang, D. E. Brown, and M. S. Gerber. Spatio-temporal modeling of criminal incidents using geographic, demographic, and Twitter-derived information. In *ISI*, pages 36–41, 2012.

[21] X. Wang, M. S. Gerber, and D. E. Brown. Automatic crime prediction using events extracted from Twitter posts. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 231–238. Springer, 2012.

[22] J. Weng and B.-S. Lee. Event detection in Twitter. In *ICWSM*, 2011.