INVITED PAPER

Spatial Event Forecasting in Social Media With Geographically Hierarchical Regularization

This paper proposes a novel multiresolution framework that can jointly optimize the forecasting accuracy and discernibility utilizing the spatial hierarchy, correlation, and heterogeneity.

By Liang Zhao, Junxiang Wang, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan

ABSTRACT | Social media has been utilized as a significant surrogate for spatial societal event forecasting. The accuracy and discernibility of a spatial event forecasting model are two key concerns, as they determine how accurate and how detailed the model's predictions will be. Existing research focuses almost exclusively on the accuracy alone, seldom considering the accuracy and discernibility simultaneously because this would require a considerably more sophisticated model while suffering from several challenges, namely: 1) the precise formulation of the tradeoff between accuracy and discernibility; 2) the scarcity of social media data with a high spatial resolution; and 3) the characterization of spatial correlation and heterogeneity. This paper proposes a novel feature learning framework that concurrently addresses all the above challenges by formulating prediction tasks for different locations with different spatial resolutions, allowing the heterogeneous relationships among the tasks to be characterized. This characterization is then integrated into our new models based on multitask learning, with parameters optimized by our proposed algorithm based on the alternative direction method of multipliers (ADMM) and dynamic programming. Extensive experimental evaluations performed on several data sets from different domains demonstrated the effectiveness of our proposed approach.

KEYWORDS | Alternating direction method of multipliers; multiresolution; spatial event forecasting

I. INTRODUCTION

Social media like Twitter and Weibo have become popular platforms, serving as real-time "sensors" for social trends and incidents [1]. Millions of Twitter users around the globe broadcast their daily observations and sentiments on an enormous variety of topics, e.g., crime, sports, and politics. The collection of these observations and sentiments could provide a useful window into emerging social trends. For instance, expressions of discontent about gas price increases could be potential precursors to a more widespread protest about government policies in general. Moreover, people use social media to plan, advertise, and organize future social events, such as the planned protests in the "Arab Spring" and "Brazilian Spring" [2]. Numerous recent research works have widely explored and demonstrated the power of social media for spatial event forecasting for various topics. A majority of them focus on temporal events like elections [3], stock market movements [4], box office ticket sales [5], and crime ratios [6]–[8]. To achieve temporal event forecasting, a number of strategies have been proposed based on various

Manuscript received November 27, 2016; revised June 13, 2017; accepted June 16, 2017. Date of publication August 16, 2017; date of current version September 18, 2017. This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337; the U.S. Government is authorized to reproduce and distribute reprints of this paper 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 U.S. Government.

(Corresponding author: Liang Zhao.)

L. Zhao and **J. Wang** are with the Faculty of the Department of Information Science and Technology, George Mason University, Fairfax, VA 22030 USA (e-mail: lzhao9@gmu.edu; junweihan2010@gmail.com).

F. Chen is with the Department of Computer Science, University at Albany–SUNY, Albany, NY 12222 USA (e-mail: fchen5@albany.edu).

C.-T. Lu and **N. Ramakrishnan** are with the Department of Computer Science, Virginia Tech, Blacksburg, VA 24061 USA (e-mail: ctlu@vt.edu; naren@cs.vt.edu).

Digital Object Identifier: 10.1109/JPROC.2017.2719039

techniques, such as time series analysis [9], supervised predictive modeling [8], [10]–[12], and causality analysis [13]–[15]. In addition, a few methods are able to forecast spatiotemporal events by further considering the information in spatial dimension, such as geographical priors [16], spatial correlation [17], [18], and demographic models [19], [20].

In spatial event forecasting, the accuracy and discernibility of the forecasting model are the two core concerns that determine how accurate and detailed the predictions will be. There is typically a tradeoff between the two: the finer the granularity for discernment, the lower the accuracy for prediction. For example, a civil unrest event could be discerned at a number of different spatial granularities ranging from country level down through state level and city level to block level. Suppose we know there will be an event on a given day in a specific country, say Mexico, which has 31 states and over 2000 cities. To predict the event location with a random predictor, we can achieve an expected accuracy of 1/31 at the state level but less than 1/2000 at the city level. The discernibility is also influenced by the capabilities of the sensors and labels. For instance, we could not make a prediction at the street level if we only possessed country-level observations or train a city-level prediction model effectively if only have state-level labels are available. Social media are composed of such noisy data that they provide social sensors with different geographical discernibilities. For example, geotagged tweets provide pinpoint geographical coordinates if their users enable this function on their mobile device, but this is uncommon; many users provide only their city information while others provide information on their state, country, or nothing at all, leaving their postings with spatial resolutions of city, state, country, or the planet Earth, respectively.

Existing work on spatial event forecasting in social media typically only zeroes in on the prediction accuracy, although the joint consideration of both discernibility and accuracy is actually a crucial issue in practice [2], [16], [21]. Until now, few related works have been published in this research area. Instead of comprehensively characterizing and utilizing the tradeoff between accuracy and discernibility, most researchers have tended to focus on the following aspects.

1) *Evaluation metrics.* Ramakrishnan *et al.* [2] proposed a new metric to evaluate the location quality of the predicted events with different spatial granularities. But they presented no analysis or suggested ways to utilize it for modeling both accuracy and discernibility for event forecasting.
2) *Multiresolution inputs.* Several studies have utilized multisource data with different geogranularities and proposed different strategies to fuse the input data, including discretization [16], clustering [22], and multilevel models [18]. However, none of these considered the effect of multiresolution on the outputs, and thus they only assess the accuracy but not the discernibility of the resulting predictions.

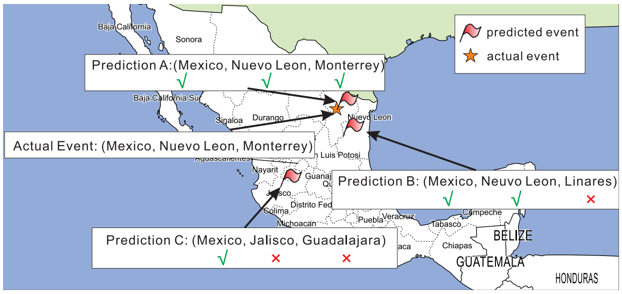


Fig. 1. ***Spatial event forecasting performance. Each prediction location is shown by a location tuple: ([country],[state],[city]), which denotes the names of country, state, and city. The qualities of these three predictions are different because they achieve correct predictions at different discernibility levels.***

3) *Multiscale event detection.* Alsaedi [23] proposed a method to jointly detect large-scale and small-scale events by using unsupervised clustering techniques to extract spatial outbreaks, but this model can only detect historical or ongoing events instead of forecasting future events.

Overall, this research area remains wide open with a distinct lack of research that comprehensively characterizes both the predictive accuracy and discernibility. The major challenges can be summarized as follows.

1) *Tradeoff between accuracy and discernibility.* Traditionally, researchers have focused on evaluating whether a prediction is correct rather than "how correct" it actually needs to be for practical purposes. For example, Fig. 1 shows three event predictions, *A*, *B*, and *C*, for a future date "August 13, 2016." Each prediction location is denoted by a tuple: ([country],[state],[city]), which denotes the names of country, state, and city, respectively. Using traditional metrics, both *B* and *C* are identified as "incorrect" and punished equally in training. However, for real-world applications, it is more reasonable to evaluate *B* as a better prediction than *C* since *B* is correct at the state level.
2) *Insufficient location information in social media data.* Existing methods typically discard large amounts of data that contain geoinformation that is insufficient for the forecasting task. For example, when performing street-level event forecasting, tweets without street-level geocodes are discarded. However, taking the Mexican Twitter data as an instance, although only around 3% of all the data possesses spatial coordinates that include street-level geoinformation, 30%–50% do contain city-level or state-level information. As only 3% of all the data would be used, it is not surprising that the model performance is then limited by insufficient data.
3) *Interaction and heterogeneity of geographical locations.* Nearby locations could benefit from regional correlations as they are likely to be influenced by the same

epidemics, natural disasters, and social events. At the same time, locations such as cities also have their own characteristics, including population, climate, and culture. Hence, it is difficult to impute basal levels of occurrence uniformly. Considering civil unrest as an example, finding 1000 tweets mentioning the keyword "protest" is not likely to mean much in a city with a population of a few million users but could be a strong indicator of an upcoming civil unrest event in a much smaller city with a population of only 10 000. In addition, it is difficult to dynamically adjust such thresholds effectively because of the data sparsity problem, especially in the latter case.

In order to simultaneously overcome all the aforementioned challenges, we propose a novel framework, for multiresolution spatial event forecasting. Being based on multitask learning, this framework jointly reinforces both the accuracy and discernibility of event forecasting with each task being treated as a model for each location with each spatial resolution. Thus, when we minimize the model's empirical loss, not only the accuracy but also the granularity of the prediction is evaluated and optimized. Moreover, by letting the models (tasks) with different spatial resolutions learn from each other, our framework provides better estimates at the finest spatial resolution by learning knowledge from coarser spatial resolutions. This capability is extremely beneficial because usually in social media the great majority of the data contain merely coarse-grained spatial information. In order to characterize the geographical neighborhood relationship among tasks, a tree-structure geographical hierarchy is also developed. The major contributions of this paper are as follows.

1) *Formulating a framework for multiresolution spatial event forecasting.* Here, multiresolution spatial event forecasting is formulated as a multitask learning problem, where a task is the model for each location in each spatial resolution. The proposed framework jointly optimizes the accuracy and discernibility of forecasting, and is enhanced by utilizing the task relatedness across different spatial resolutions and neighboring locations.
2) *Proposing a multitask model with heterogeneous task relationships.* In the proposed multitask model, three types of task relationships are considered, namely the spatial neighborhood, spatial resolution, and spatial parent–child relationships. All are characterized by different regularization terms and constraints.
3) *Developing an efficient algorithm for a new variant of overlapping group lasso problem.* The optimization of the proposed multitask models involves overlapping group lasso problems with nonsmooth equality and inequality constraints, which is challenging to solve. By introducing auxiliary variables and proposing a new dynamic programming-based method, we develop an effective alternative direction method of multipliers (ADMM)-based algorithm to ensure an effective solution for this problem.
4) *Comprehensive experiments to validate the performance of the proposed techniques.* We conducted extensive evaluations of the proposed methods using several data sets and compared the results obtained with those from seven existing event forecasting methods. The new methods proposed here consistently outperformed all the competing methods. We also performed sensitivity analyses to investigate the impact of various parameters on the performance of the proposed methods.

The rest of this paper is organized as follows. Section II reviews existing work. Section III introduces the problem setup. Section IV provides details of our proposed models and their parameter optimization algorithms. In Section VI, extensive experiments to evaluate the performance of the proposed models are conducted and analyzed; the work is summarized and conclusions are drawn in Section VII.

II. RELATED WORK

The related work is presented and summarized in this section.

A. Event Detection

A considerable amount of work has been done on the identification of ongoing social events [24]. Generally, for event detection, either classification or clustering is utilized to extract contents of interest as "signals" of social indicators, which are then examined to identify the potential occurrence of ongoing events by considering different types of signal burstiness, as described in the following.

1) *Temporal burstiness* [25]–[30]. Focusing on temporal events like elections and stock market movements, this type of methods tracks the time-evolving signals and identifies their spikes in the temporal dimension. For example, Schubert *et al.* [25] proposed to detect the popularity of the trending topics in the web using a "hotness" metric and an efficient algorithm for selecting the topic surrogates. To detect the streaming events like sports and elections, Adedoyin *et al.* [26] proposed a transaction-based method for temporal change pattern mining.
2) *Spatiotemporal burstiness* [31]–[39]. This type of methods examines the patterns of signals in both temporal and spatial dimensions, and identifies the events that are typically spatiotemporal outbreaks. For example, Krumm and Horvitz [35] proposed Eyewitness, a system that identifies the anomalous spatiotemporal spikes based on time series analysis of geotagged tweet volumes. Similarly, Zhang *et al.* [36] proposed a spatial event detection system that ingests, processes, summarizes, and monitors Twitter

streams in real time. Utilizing retrospective analysis on tweets, Dong *et al.* proposed a wavelet-based clustering method to extract those related to the same historical events but with different time durations and spatial sizes [34]. However, rather than forecasting events in the future, these approaches typically uncover them only after they have occurred.

B. Event Forecasting

Social event forecasting methods can be classified into three categories according to the problem formulations.

1) *Causality-based methods* [13]–[15], [40], [41]. This type of models predicts future events directly based on their causal relations with other relevant ongoing or historical events, without sophisticated considerations on temporal or spatial patterns. This type of methods relies heavily on the data quality and causal assumptions. For example, Muthiah *et al.* [41] first utilized a list of phrases to retrieve the "propaganda events" in news articles, from which they directly extracted the future events planned in those propaganda. Furthermore, the model proposed by Hu *et al.* [14] is able to automatically identify the conditional probabilities among different events in news utilizing deep models.
2) *Temporal signal-based methods* [3], [4], [4]–[12]. Social events are usually too complex for us to find clear causal relations among them, and hence comprehensive candidate signals typically need to be taken into account [2]. To harness these multivariate (sometimes even high-dimensional) signals, supervised learning techniques were commonly used to formulate forecasting tasks into classification or regression problems. Here, the historical signals are inputs while the event occurrences at future time points are model responses. For example, Wang *et al.* [6] extracted latent topics from tweet messages and used them as inputs to forecast the crime ratio based on logistic regression.
3) *Spatiotemporal signal-based methods* [17]–[21]. Beyond temporal dimension, many social events typically exhibit spatial properties such as disease outbreaks and civil unrest [2]. To forecast spatial events, some existing approaches are able to further explore and exploit spatial priors and patterns. Similar to Zhao *et al.* [19], Zhang *et al.* [20] proposed a system for influenza outbreaks forecasting that relies on a domain-specific mechanistic model and demographic information with the enhancement of social media data mining. Zhao *et al.* [21] proposed a multitask learning framework for event forecasting that jointly learns multiple related spatial locations. Relying on a special healthcare metadata, the model proposed by Rekatsinas *et al.* [17] predicted rare disease outbreaks based on spatiotemporal anomaly detection over autoregressive indicators. However, existing methods typically only consider events using a single geographical granularity and do not jointly optimize the discernibility and accuracy.

C. Multitask Learning

In multitask learning (MTL), multiple related tasks are learned simultaneously to improve generalization performance [1]. Many MTL approaches have been proposed [42]. In [43], Kim *et al.* proposed a regularized MTL which constrained the models of all tasks to be close to each other. This task relatedness can also be characterized by constraining multiple tasks to share a common underlying structure, such as a common set of features [44] or a common subspace [45], or by using a tree-structured model [43]. For example, Kim and Xing [43] proposed a multitask learning model which leverages a tree-structured relationship among the tasks. MTL approaches have been applied in many domains, including computer vision and biomedical informatics. Cheng *et al.* [46] proposed to impose regularization on convolutional neural networks to enforce the feature representations of objectives with different rotations in the images. Lin *et al.* [47] proposed an efficient model to consider the interaction features in multitask learning. To the best of our knowledge, however, we are the first to apply MTL for multiresolution spatial civil unrest forecasting.

D. Multiresolution Detection

Multiresolution detection approaches have been widely applied in domains such as computer vision and satellite remote sensing [48] for object detection. To analyze the response rates for website advertisements, Agarwal *et al.* [49] developed a method for estimating and predicting fine-grained geolocations. Aiming at a retrospective analysis of historical events, Jiang *et al.* [48] designed a framework to extract and summarize events from different views with different resolutions. As of yet, however, few researchers are utilizing multiresolution in spatial event forecasting. In addition to the features typically extracted by a deep Boltzmann machine, Han *et al.* [50] utilized higher level features from weakly labeled images to enhance the object detection performance in the images. Cheng *et al.* [51] developed a new type of higher level image feature that they dubbed "partlets" to better represent the images and used l_0-norm to enforce the sparsity. High-level features were also used by Yao *et al.* [52] to develop automatic semantic annotation for satellite images.

Multiresolution approaches have been used for temporal event detection. For example, to address the problem of worm detection, Sekar *et al.* [53] applied temporal multiresolution using different sliding-window sizes to handle high-rate and low-rate attacks. Moon *et al.* [54] took the concept further by applying it to handle the challenges in memory efficiency when implementing the temporal multiresolution concept. In order to mine the hidden patterns in the data more effectively, Doucoure *et al.* [55] utilized wavelet decomposition and artificial neural networks for their multiresolution analysis of wind speed time series, while Cooper *et al.* [56] proposed a multiresolutions temporal clustering method that allow users to organize their photo collections at different time scales. Jiang *et al.* [57] performed wavelet transforms on multiscale network traffic time series and then utilized the principal components to classify the anomalousness of the network traffic.

There has been far less work on spatial-resolution event analytics, however, with existing works typically focusing on the following aspects.

1) *Evaluation metrics.* Ramakrishnan *et al.* [2] proposed a new metric to evaluate the location quality of the predicted events with different spatial granularities but conducted no analysis or examples of its utilization for modeling accuracy and discernibility for event forecasting.
2) *Multiresolution inputs.* Several studies have utilized multisource data with different geogranularities and proposed different strategies to fuse the input data, including discretization [16], clustering [22], and multilevel models [18]. However, none has considered the multiresolution in the outputs, and thus they focus sorely on the accuracy and ignore the discernibility in prediction.
3) *Multiscale event detection.* Alsaedi *et al.* [23] proposed a method to jointly detect large-scale and small-scale events by using unsupervised clustering techniques to extract the spatial outbreaks, while Dong [34] presented a concrete clustering algorithm capable of discovering temporal and spatial outbreaks to different scales. Unfortunately, at present, these models for multiscale event detection can only detect historical or ongoing events and are not able to forecast future events. To the best of our knowledge, we are the first to apply multiple geographical resolutions to civil unrest forecasting.

III. PROBLEM SETUP

The problem setup for this paper is presented in this section.

Denote $X=\{X_t\}_t^T$ as a collection of time-indexed Twitter data, where $X_t \in X$ represents the subcollection of tweets at tth time interval and T is the set of time intervals. According to the granularity of the geoinformation they contain, tweets can be geocoded into different spatial resolutions corresponding to different levels of administrative divisions, such as country level, state level, and city level. Before formally stating the problem, we first introduce two definitions related to geographical hierarchy.

Definition 1 (Spatial Subregion): Given two locations q_i and s_j under ith and jth ($i > j$) spatial resolutions, respectively, if the entire spatial area of location q_i is included within location s_j, q_i is a spatial subregion of s_j, denoted as $q_i \sqsubseteq s_j$ or equally $s_j \sqsupseteq q_i$ ($i > j$).

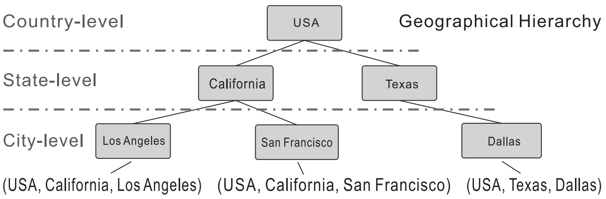


Fig. 2. ***The location tuples based on geographical hierarchy.***

Definition 2 (Location Tuple): As shown in Fig. 2, the location of a tweet or an event is denoted by a location tuple $s = (s_1, s_2, \ldots, s_N)$, which is an array that configures each location s_n in each spatial resolution n by a parent–child hierarchy such that $s_n \sqsubseteq s_{n-1}$ ($n = 2, \ldots, N$); s_{n-1} is the parent of s_n while s_n is the child of s_{n-1}.

A tweet subcollection X_t can be spatially distributed in N different ways based on the N different spatial resolutions such that $\{X_{t,s_n}\}_{s_n}^{S_n} \subseteq X_t$, where S_n is the location set under the nth spatial resolution, $n = 1, \ldots, N$. $X_{t,s_n} \in \mathbb{N}^{K\times 1}$ is a feature vector for the tweets in location $s_n \in S_n$ at time t, where the elements could be, for instance, the keyword counts and the number of retweets. K is the number of features. Also, define $S = \bigcup_n^N S_n$ as the set of all the locations in different spatial resolutions. In addition, for each location s_n with spatial resolution n at time τ, we denote the actual occurrence ("yes" = 1 or "no" = 0) of a future event as a binary variable $Y_{\tau,s_n} \in \{0, 1\}$, where $Y_{\tau,s_n} = 0$ means no event occurs; otherwise $Y_{\tau,s_n} = 1$. According to the definition of the location tuple, we also have $Y_{\tau,s} = (Y_{\tau,s_1}, \ldots, Y_{\tau,s_N})$. The problem addressed in this paper can thus be formulated as follows.

Problem Formulation: Given the tweet data X_t in N different spatial resolutions, the goal is to predict the occurrence of a future event for location $s = (s_1, \ldots, s_N)$ within time interval τ, where s_n ($n = 1, \ldots, N$) is the location name for the nth spatial resolution. In addition, $\tau = t + p$, where $p > 0$ is the lead time. Formally, this problem is formulated as learning the mapping from tweets data to future event predictions $f: X_{t,s} \rightarrow \{Y_{\tau,s_1}, \ldots, Y_{\tau,s_N}\}$ for locations s at N spatial resolutions.

Definition 3 (Multiresolution Event Forecasting Error): The multiresolution event forecasting error $\mathcal{L}(W)$ is defined as the summation of errors in all the spatial resolutions against the labels of actual event occurrence

$$\mathcal{L}(W) = \frac{1}{|S|} \Sigma_n^N \Sigma_{s_n}^{S_n} \mathcal{L}(W_{s_n})$$

where $W = \{\{W_{s_n}\}_{s_n}^{S_n}\}_n^N$ is the parameter of the forecasting model and $W_{s_n} \in \mathbb{R}^{1\times K}$. $\mathcal{L}(W_{s_n})$ is the sum of the empirical errors of the prediction $f(X_{t,s_n} \cdot W_{s_n})$ against the labels Y_{τ,s_n} for all the time intervals T. $\mathcal{L}(W_{s_n})$ can be a logistic loss [58] where $f(x) = 1/(1 + e^{-x})$.

Due to the different characteristics of the various locations and spatial resolutions involved, it is unfeasible to build a single model to characterize them all simultaneously. To address this issue, a simple approach is to learn corresponding models for different locations and different spatial resolutions. However, this creates several challenges.

1) *Data scarcity.* Geographically small locations typically lack sufficient data to train models adequately. Moreover, due to the scarcity of tweet data with high spatial resolution, prediction tasks that involve high resolution are also more challenging.
2) *Spatial neighborhood.* Forecasting tasks have regional relatedness such that nearby locations could be influenced by interrelated events.
3) *Multiresolution event forecasting paradox.* Contradictory predictions at different spatial resolutions can also happen. For example, a model that predicts there will be an event in "Los Angeles" could also predict that there will be no event in "California."

To address these three challenges, in the next section, we propose a novel multitask learning model that we have named MREF based on mixed-structured task relatedness and a nonsmooth constraint.

IV. MULTIRESOLUTION SPATIAL EVENT FORECASTING

In this section, we describe a new multitask learning framework for multiresolution spatial event forecasting. In Section IV-A, the multiple types of task relatedness are characterized mathematically. Section IV-B then proposes two novel models, MREF-I and MREF-II, based on different assumptions for the task relatedness.

A. Heterogeneous Relatedness of Tasks

The forecasting models for all the locations are built simultaneously by characterizing the structural relationships and utilizing appropriate shared information across tasks. Fig. 3 illustrates the proposed multitask learning framework that characterizes all three major aspects of relatedness among all the locations (tasks) for the problem of multiresolution spatial event forecasting: 1) spatial neighborhood relationships; 2) spatial resolution relationships; and 3) parent–child relationships.

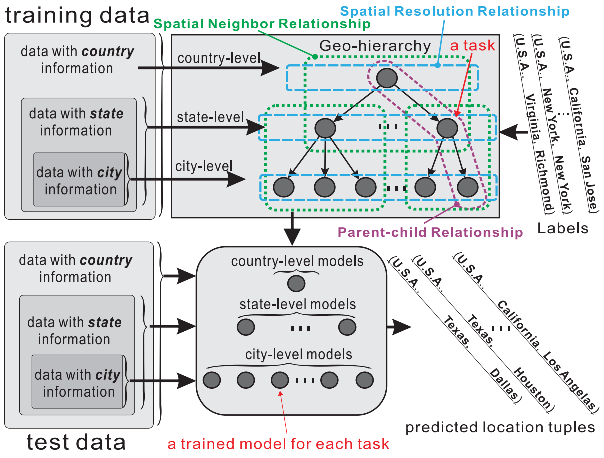

training data
Spatial Neighbor Relationship
Spatial Resolution Relationship
Geo-hierarchy
a task
data with **country** information
data with **state** information
data with **city** information
country-level
state-level
city-level
(U.S.A., Virginia, Richmond)
(U.S.A., New York, New York)
(U.S.A., California, San Jose)
Labels
Parent-child Relationship
data with **country** information
data with **state** information
data with **city** information
country-level models
state-level models
city-level models
(U.S.A., Texas, Dallas)
(U.S.A., Texas, Houston)
(U.S.A., California, Los Angelas)
predicted location tuples
test data
a trained model for each task

Fig. 3. ***A schematic view of the proposed model.***

1) *Spatial Neighborhood Relationships:* Events that occur at neighboring locations at around the same time could well involve similar topics, so the tweets from different locations may share a number of common keywords that are related to the events. To take this into account, the geographical hierarchy among locations is leveraged, which is shown as a tree in the top right of Fig. 3; the location (task) nodes in a subtree are within a spatial neighborhood.

As shown in Fig. 3, a geographical hierarchy is a tree whose nodes consist of all the possible spatial locations and the links are the parent–child relationships among them. In this geographical hierarchy tree, denote $\mathcal{T}=\{\mathcal{T}_i\}_i$ as the set of subtrees that are defined as $\mathcal{T}_i=\{s_n\}\cup\{s'_{n+1}\,|\,s'_{n+1}\sqsubseteq s_n, n<N\}$, which means a subtree contains a location s_n and all of its children. Denote $P_{s_n,k}$ as the spatial neighborhood relationship model parameter for location s_n and feature k. Define $P_{\mathcal{T}_i,k}$ as the set of model parameters for the subtree $\mathcal{T}_i$ for feature k such that

$$P_{\mathcal{T}_i,k}=\bigcup\nolimits_{s_n\in\mathcal{T}_i} P_{s_n,k}. \tag{1}$$

To incorporate the spatial neighborhood relationship, the model needs to enforce a similar feature selection pattern across the prediction tasks for locations in the subtree $\mathcal{T}_i$.

2) *Spatial Resolution Relationships:* Tasks for locations with the same spatial resolution have a closer spatial scale, so tweets from these locations may share a closer scale of keyword counts and retweet counts. To encompass this notion, we denote $Q_{s_n,k}$ as the spatial resolution relationship model parameter for location s_n and feature k. $Q^{(n)}_{\cdot,k}$ represents the model parameters for feature k for all the locations at n spatial resolution such that

$$Q^{(n)}_{\cdot,k}=\bigcup\nolimits_{s_n\in S_n} Q_{s_n,k} \tag{2}$$

where S_n is the set of all the locations at the nth spatial resolution. When considering spatial resolution relationships, the model needs to enforce a similar feature selection pattern across the prediction tasks for locations with the same spatial resolution.

3) *Parent–Child Relationships:* The situation of an event occurrence in a location indicates and constrains the possible situations for its child locations, and *vice versa*. When we learn a model for a specific location with a specific spatial resolution, we also "borrow" information from the other locations with different spatial resolutions, so learning multiple related tasks simultaneously increases the sample size for each location. The parent–child relationship among locations within different spatial resolutions can be characterized as follows.

Lemma 1: If there is no event in a location, then there is no event in all of its subregions. Formally, without loss of generality, assume $i > j$, then $\forall q_i \sqsubseteq s_j \wedge Y_{\tau,s_j} = 0 : Y_{\tau,q_i} = 0$, which is equal to $\exists s_j \sqsupseteq q_i \wedge Y_{\tau,q_i} = 1 : Y_{\tau,s_j} = 1$.

Theorem 1: According to the definition of Y such that $Y_{\tau,s} \in \{0,1\}$, the sufficient and necessary condition of Lemma 1 is $Y_{\tau,s_j} \geq \max(\{ Y_{\tau,q_i} | q_i \sqsubseteq s_j, i > j\})$.

Proof: Sufficiency. Given $Y_{\tau,s_j} \geq \max(\{ Y_{\tau,q_i} | q_i \sqsubseteq s_j, i > j\})$ and $Y_{\tau,s_j} \in \{0,1\}$, if $Y_{\tau,s_j} = 0$, it is clear that $\max(Y_{\tau,q_i}) = 0$, which is equal to $Y_{\tau,q_i} = 0$ for any $q_i \sqsubseteq s_j$. The sufficiency is proved.

Necessity. If $Y_{\tau,s_j} = 1$, then $Y_{\tau,s_j} \geq \max\ (\{ Y^{(i)}_{\tau,q_i} | q_i \sqsubseteq s_j \})$ is satisfied based on the definition of Y. On the other hand, when $Y_{\tau,s_j} = 0$, according to Lemma 1, we know $\max(Y_{\tau,q_i}) = 0$. Thus, the necessity is proved.

Lemma 1 and Theorem 1 specify how to utilize the event occurrence information in the parent location to regulate the event occurrence identification in the child location. This is a necessary condition of the parent–child relationship. In fact, the event occurrence information in the child location can also constrain the event occurrence pattern in the parent location, which is discussed in the following lemma.

Lemma 2: There is no event in a location, if there is no event in any of its subregions. Formally, without loss of generality, assume $i > j$, then $\forall q_i \sqsubseteq s_j \wedge Y_{\tau,q_i} = 0 : Y_{\tau,s_j} = 0$. Thus, we have $Y_{\tau,s_j} \leq \max(\{ Y_{\tau,q_i} | q_i \sqsubseteq s_j, i > j\})$

Combining Theorem 1 and Lemma 2, we get the sufficient and necessary condition of the parent–child relationship.

Corollary 1: There is no event in a location if, and only if, there is no event in any of its subregions. On the other hand, there is at least one event in a location if, an only if, there is event in at least one location among its subregions. Formally, $Y_{\tau,s_j} = \max(\{ Y_{\tau,q_i} | q_i \sqsubseteq s_j,\ i > j\})$.

B. Models

The above consideration of the heterogeneous relatedness of forecasting tasks leads to a new multitask feature learning framework developed based on a general paradigm of multitask learning, namely minimizing the penalized empirical loss

$$\min_{W} \mathcal{L}(W) + \Omega(W) \quad s.t.\ W \in \mathcal{C} \tag{3}$$

where $\mathcal{L}(W)$ is the forecasting error on the training set, as defined in Definition 3, and $\Omega(W)$ is the regularization term that encodes structured task relatedness for both spatial neighborhood relationships and spatial resolution relationships. To achieve this, we decompose the model parameter W into two components: a tree-structured component P for spatial neighborhood relationships and a grouping-structured component Q for spatial resolution relationships such that $W = P + Q$. To take into account the parent–child relationship, the feasible set is constrained to a convex set $\mathcal{C}$. When only the necessary condition of the parent–child relationship is considered as the constraint, Theorem 1 is leveraged to instantiate the parent–child relation. The resulting model is called the multiresolution spatial event forecasting model I (MREF-I)

$$\begin{aligned} &\min_{W} \mathcal{L}(W) + \gamma_P \textstyle\sum_{k,\mathcal{T}_i}^{K,\mathcal{T}} \| P_{\mathcal{T}_i,k} \|_F + \gamma_Q \textstyle\sum_{k,n}^{K,N} \| Q^{(n)}_{\cdot,k} \|_F \\ &\qquad s.t. \qquad W = P + Q, \\ &f(X_{t,s_n} \cdot W_{s_n}) \geq \max(\{f(X_{t,s'_{n+1}} \cdot W_{s'_{n+1}}) | s'_{n+1} \sqsubseteq s_n\}) \end{aligned} \tag{4}$$

where the Frobenius norm $\|\cdot\|_F$ is utilized in $\sum_k^K \sum_{\mathcal{T}_i}^{\mathcal{T}} \| P_{\mathcal{T}_i,k} \|_F^2$ to enforce a similar feature selection among tasks with the spatial neighborhood. $\sum_k \sum_i^n \| Q^{(i)}_{\cdot,k} \|_F^2$ enforces similar feature selection among tasks in the same spatial resolution. The inequality constraint is introduced from Theorem 1 by considering the mapping $f : X_{t,s_n} \rightarrow Y_{\tau,s_n}$. γ_P and γ_Q are regularization parameters such that $\gamma_P = \gamma / \sum_{\mathcal{T}_i}^{\mathcal{T}} \sqrt{|\mathcal{T}_i|}$ and $\gamma_Q = \gamma / \sum_n^N \sqrt{|S_n|}$, where γ is the regularization parameter that balances the tradeoff between the loss function $\mathcal{L}(W)$ and the regularization terms.

We leverage the sufficient and necessary condition of the parent–child relationship, then the constraint is instantiated based on Corollary 1. The modified model is called the multiresolution spatial event forecasting model II (MREF-II)

$$\begin{aligned} &\min_{W} \mathcal{L}(W) + \gamma_P \textstyle\sum_{k,\mathcal{T}_i}^{K,\mathcal{T}} \| P_{\mathcal{T}_i,k} \|_F + \gamma_Q \textstyle\sum_{k,n}^{K,N} \| Q^{(n)}_{\cdot,k} \|_F \\ &\qquad s.t. \qquad W = P + Q, \\ &f(X_{t,s_n} \cdot W_{s_n}) = \max(\{f(X_{t,s'_{n+1}} \cdot W_{s'_{n+1}}) | s'_{n+1} \sqsubseteq s_n\}). \end{aligned} \tag{5}$$

The parameter optimizations of both MREF-I and MREF-II are challenging due to the existence of nonlinear and nonsmooth terms in the constraint. In the following, we propose novel algorithms to handle these challenges.

V. OPTIMIZATION

This section describes the optimization algorithms for the proposed MREF-I and MREF-II. In Section V-A, a parameter optimization algorithm for MREF-I is proposed to handle the nonsmooth inequality constraint. Section V-B elaborates on the new algorithm for MREF-II, in which a new dynamic programming-based algorithm is proposed to address the nonsmooth equality constraint.

A. Parameter Optimization of MREF-I

1) Objective Function of MREF-I: The objective function in (4) encompasses the joint consideration of the heterogeneous task relationships. However, to solve this objective function two challenges must first be overcome: 1) nonsmooth inequality constraint; and 2) overlapping among the coupled subtrees, discussed and addressed as follows.

1) *Nonsmooth inequality constraint:* The nonsmooth function *max*$(\cdot)$ in the inequality constraint in (4) makes the objective function difficult to solve. To address this challenge, we replace this term with an

alternative constraint that applies a sufficiency condition to the original constraint

$$f(X_{s_n,t}\cdot W_{s_n}) \geq f(X_{s'_{n+1},t}\cdot W_{s'_{n+1}}),\ s'_{n+1} \sqsubseteq s_n. \tag{6}$$

This is equal to the following equation due to the strictly monotonic increase of the function $f(\cdot)$:

$$X_{s_n,t}\cdot W_{s_n} \geq X_{s'_{n+1},t}\cdot W_{s'_{n+1}},\ s'_{n+1} \sqsubseteq s_n \tag{7}$$

which is both linear and smooth and thus ensures the accurate solution of the original objective function.

2) *Overlapping among the coupled subtrees:* As Fig. 3 demonstrates, a node in the geographical hierarchy tree can belong to two different subtrees. For example, state-level nodes belong to a subtree whose root is a country-level node, but they can also be the root of another subtree whose leaves are city-level nodes. This issue prevents an easy solution because a model parameter could be regularized by different Frobenius norm terms. To solve this, we propose the use of an efficient optimization solution based on the introduction of two auxiliary variables U and V. U is the model parameter set for the set of subtrees $\mathcal{T}_\mathcal{O}$ with roots in odd-number (i.e., $n = 1, 3, 5, \ldots$) spatial resolutions, while V represents the set of subtrees $\mathcal{T}_\mathcal{E}$ with roots in even-number (i.e., $n = 2, 4, 6, \ldots$) levels. Thus, neither U nor V contains overlapping subtrees. We also know that $\mathcal{T}_\mathcal{O} \cup \mathcal{T}_\mathcal{E} = \mathcal{T}$and $\mathcal{T}_\mathcal{O} \cap \mathcal{T}_\mathcal{E} = \varnothing$.

Therefore, the objective function becomes

$$\begin{aligned}\min_W\ & \mathcal{L}(W) + \gamma_0 \textstyle\sum_k^K \sum_{\mathcal{T}_i}^{\mathcal{T}_\mathcal{O}} \| U_{\mathcal{T}_i,k}\|_F + \gamma_1 \sum_k^K \sum_{\mathcal{T}_T}^{\mathcal{T}^\mathcal{E}} \|V_{\mathcal{T}_i,k}\|_F \\ &\quad + \gamma_2 \textstyle\sum_k^K \sum_n^N \| Q_{\cdot,k}^{(n)}\|_F \\ s.t.\ & W = P + Q,\ P = U,\ P = V, \\ & g(X,W) + \beta = 0, \beta = \beta_+,\ \beta_+ \geq 0\end{aligned} \tag{8}$$

where $g(X\cdot W) = \{g(X_{s_n,t}, W_{s_n})\}_{s_n,t}^{S',T}$ is a matrix composed of the set of elements $g(X_{s_n,t}, W_{s_n}) = X_{s'_{n+1},t}\cdot W_{s'_{n+1}} - X_{s_n,t}\cdot W_{s_n}$ and $S' = \bigcup_{n=1}^{N-1} S_n$. Two auxiliary matrix variables β and β_+ are added, which have the same size of $g(X,W)$. γ_0, γ_1, and γ_2 are regularization parameters such that $\gamma_0 = \gamma / \sum_{\mathcal{T}i}^{\mathcal{T}_\mathcal{O}} \sqrt{|\mathcal{T}_i|}$, $\gamma_1 = \gamma / \sum_{\mathcal{T}i}^{\mathcal{T}_\mathcal{E}} \sqrt{|\mathcal{T}_i|}$, and $\gamma_2 = \gamma / \sum_n^N \sqrt{|S_n|}$ here is the regularization parameter that balances the tradeoff between the loss function $\mathcal{L}(W)$ and the regularization terms.

2) *Optimization Algorithm of MREF-I:* The objective function in (8) is convex because the loss function, inequality constraints, and regularization terms are convex, while the equality constraints are affine. To solve the convex and nonsmooth objective function with constraints, the alternating direction method of multipliers (ADMM) has begun to be widely utilized as an efficient algorithm. ADMM first breaks the original large problem into smaller subproblems that can be solved easily and fast, and then iteratively solves the subproblems in turn until convergence is achieved. Here we propose an ADMM-based framework that solves (8) by first obtaining its augmented Lagrangian format as follows:

$$\begin{aligned}\min_\Theta \mathcal{L}(W) &+ \gamma_0 \textstyle\sum_{k,\mathcal{T}_i}^{K,\mathcal{T}_\mathcal{O}} \| U_{\mathcal{T}_i,k}\|_F + \gamma_1 + \gamma_1 \sum_{k,\mathcal{T}_i}^{K,\mathcal{T}_\mathcal{E}} \|V_{\mathcal{T}_i,k}\|_F \\ &+ \gamma_2 \textstyle\sum_k \sum_n^N \|Q_{\cdot,k}^{(n)}\|_F + \langle \alpha_1, W - P - Q\rangle \\ &+ \frac{\rho}{2}\|W - P - Q\|_F^2 + \langle \alpha_2, P - U\rangle + \frac{\rho}{2}\|P - U\|_F^2 \\ &+ \langle \alpha_3, P - V\rangle + \frac{\rho}{2}\|P - V\|_F^2 + \langle \alpha_4, g(X,W) + \beta\rangle \\ &+ \frac{\rho}{2}\| g(X,W) + \beta\|_F^2 + \langle \alpha_5, \beta - \beta_+\rangle + \frac{\rho}{2}\|\beta - \beta_+\|_F^2\end{aligned} \tag{9}$$

where $\Theta = \{W, P, U, V, \alpha, \beta, \beta_+\}$ are the parameters to be optimized. $\alpha = \{\alpha_i\}_{i=1}^5$ is the set of Lagrangian muliplier that are the dual variables of ADMM and ρ is the step size of the dual step. The parameters $\Theta = \{W, P, U, V, \alpha, \beta, \beta_+\}$ are alternately solved by the proposed algorithm, referred to as mixed-structured multitask learning, as shown in Algorithm 1. It alternately optimizes each of the parameters in Θ until an acceptable residual is achieved. Lines 4 and 5 show the alternating optimization of each of the parameters. The calculation of the primal and dual residuals are illustrated in line 6. Lines 7–13 describe the updating of the penalty parameter ρ, which follows the updating strategy proposed by Boyd *et al.* [59]. The detailed optimization steps are described in more detail below.

1) *Update W, fix others:* The optimization of the parameter W is a generalized linear regression with square loss functions

$$\begin{aligned}W \leftarrow \operatorname*{argmin}_W \mathcal{L}(W) &+ \langle \alpha_2, g(X,W) + \beta\rangle + \frac{\rho}{2}\|g(X,W) + \beta\|_F^2 \\ &+ \langle \alpha_1, W - P - Q\rangle + \frac{\rho}{2}\|W - P - Q\|_F^2.\end{aligned} \tag{10}$$

In order to solve this problem, a second-order Taylor expansion is performed, where we approximate the Hessian using a multiple of the identity with an upper bound of $(1/4)I$.

2) *Update P, fix others:* The optimization of P can be formulated as the following least squares problem:

$$\begin{aligned}P \leftarrow \operatorname*{argmin}_P\ &\langle \alpha_1, W - P - Q\rangle + \langle \alpha_2, P - U\rangle + \frac{\rho}{2}\|P - U\|_F^2 \\ &+ \frac{\rho}{2}\|W - P - Q\|_F^2 + \langle \alpha_3, P - V\rangle + \frac{\rho}{2}\|P - V\|_F^2\end{aligned} \tag{11}$$

where the solution is $(1/3)(W + U + V - Q) + (1/(3\rho))(\alpha_1 - \alpha_2 - \alpha_3)$.

3) *Update U, V, Q, fix others:* The optimizations of U, V, and Q are all problems of least squares loss functions with $\ell_{2,1}$-norms

$$\begin{aligned}U &\leftarrow \operatorname*{argmin}_U \gamma_0 \textstyle\sum_{k,\mathcal{T}_i}^{K,\mathcal{T}_\mathcal{O}} \| U_{\mathcal{T}_i,k}\|_F + \langle \alpha_2, P - U\rangle + \frac{\rho}{2}\|P - U\|_F^2 \\ V &\leftarrow \operatorname*{argmin}_V \gamma_0 \textstyle\sum_{k,\mathcal{T}_i}^{K,\mathcal{T}_\mathcal{E}} \| V_{\mathcal{T}_i,k}\|_F + \langle \alpha_2, P - V\rangle + \frac{\rho}{2}\|P - V\|_F^2 \\ Q &\leftarrow \operatorname*{argmin}_Q \gamma_2 \textstyle\sum_k \sum_n^N \| Q_{\cdot,k}^{(n)}\|_F + \langle \alpha_1, W - P - Q\rangle \\ &\qquad + \frac{\rho}{2}\|W - P - Q\|_F^2\end{aligned} \tag{12}$$

where all three problems can be efficiently solved by using proximal operators [60].

4) Update β, fix others: The optimization of β can be formulated as the following least squares problem:

$$\begin{aligned}\beta \leftarrow & \operatorname*{argmin}_{\beta} \langle \alpha_4, g(X,W) + \beta \rangle + \frac{\rho}{2} \| g(X,W) + \beta \|_F^2 \\ & + \langle \alpha_5, \beta - \beta_+ \rangle + \frac{\rho}{2} \| \beta - \beta_+ \|_F^2 \qquad (13)\end{aligned}$$

where the solution is $\beta = (1/2)(\beta_+ - g(X,W)) - (1/(2\rho))(\alpha_4 + \alpha_5)$.

5) Update β_+, fix others: The optimization of β_+ can be formulated as a least squares problem with linear inequality constraint

$$\beta_+ \leftarrow \operatorname*{argmin}_{\beta_+ \geq 0} \langle \alpha_5, \beta - \beta_+ \rangle + \frac{\rho}{2} \| \beta - \beta_+ \|_F^2. \qquad (14)$$

To eliminate the inequality constraint, first let $c^2 = \beta_+$, $c \in \mathbb{R}$, yielding the following equivalent problem:

$$\beta_+ \leftarrow \operatorname*{argmin}_{c^2} \langle \alpha_5, \beta - c^2 \rangle + \frac{\rho}{2} \| \beta - c^2 \|_F^2.$$

It can be easily seen that β_+ has two solutions: $\beta_+ = c^2 = \beta + \alpha_5/\rho$ and $\beta_+ = c^2 = 0$. Therefore, the solution is $\beta_+ = \max(\beta + \alpha_5/\rho, 0)$.

6) Update $\alpha_i (i = 1, \ldots, 5)$: The updating of the dual variables α_i is as follows:

$$\begin{aligned} \alpha_1 \leftarrow & \quad \alpha_1 + \rho \cdot (W - P - Q) \\ \alpha_2 \leftarrow & \quad \alpha_2 + \rho \cdot (P - U), \quad \alpha_3 \leftarrow \alpha_3 + \rho \cdot (P - V) \\ \alpha_4 \leftarrow & \quad \alpha_4 + \rho \cdot (F + \beta), \quad \alpha_5 \leftarrow \alpha_5 + \rho \cdot (\beta - \beta_+). \qquad (15)\end{aligned}$$

7) Calculate residuals: The primal and dual residuals of the $(k+1)$th iteration are calculated based on the following theorem, where the parameters with superscript k (e.g., P^k) correspond to their values in the kth iteration.

Algorithm 1 Optimization of MREF-I

```
Input: $X$, $Y$, $\gamma$
Output: solution $W$
Initialize $\rho$=1, $W$, $U$, $V$, $P$, $Q$, $\alpha_i$, $\beta$, $\beta^+$=0, $i$ = 1,⋯,5.
Choose $\varepsilon_p > 0, \varepsilon_d > 0$.
repeat
    Update $W$, $U$, $V$, $P$, $Q$ by Equations (10), (11), and (12).
    Update $\{\alpha_i\}_{i=1}^5$, $\beta$, $\beta^+$ by Equations (13) and (15).
    Update primal and dual residuals $p$ and $d$ based on Theorem 2.
    if $p$ > 10$d$ then
        $\rho \leftarrow 2\rho$    # Update penalty parameter
    else if 10$p$ < $d$ then
        $\rho \leftarrow \rho/2$ # Update penalty parameter
    else
        $\rho \leftarrow \rho$   # Update penalty parameter
    end if
until $p < \varepsilon^p$ and $d < \varepsilon^d$ #  Convergence criterion
```

Theorem 2: The primal residual and the dual residual of the algorithm are as follows.

- The primal residual: $p = \| W^{k+1} - P^{k+1} - Q^{k+1} \|_F + \| P^{k+1} - U^{k+1} \|_F + \| P^{k+1} - V^{k+1} \|_F + \| g(X, W^{k+1}) + \beta^{k+1} \|_F + \| \beta^{k+1} - \beta_+^{k+1} \|_F$.
- The dual residual: $d = \rho \cdot (\| (P^{k+1} - P^k) + (Q^{k+1} - Q^k) + \partial g(X, W^{k+1})(\beta^k - \beta^{k+1}) \|_F + \| (U^{k+1} - U^k) + (V^{k+1} - V^k) + (Q^{k+1} - Q^k) \|_F + \| \beta_+^{k+1} - \beta_+^k \|_F)$.

Proof: The primal residual is easily deduced from the primal feasibility according to (22) directly. The deduction of the dual residual is elaborated in the following. The dual feasibility of the objective function is

$$\begin{aligned} 0 \in & \quad \partial \mathcal{L}(W^*) + \alpha_1^* + \alpha_4^* \partial g(X, W^*) \\ 0 \in & \quad \partial \gamma_0 \textstyle\sum_{v,k} \| U_{G(v),k}^* \|_F^2 + \alpha_2^* \qquad (17) \\ 0 \in & \quad \partial \gamma_1 \textstyle\sum_{v,k} \| V_{G(v),k}^* \|_F^2 + \alpha_3^* \qquad (16) \\ 0 \in & \quad \partial \gamma_2 \textstyle\sum_{v,k} \| (Q_{\cdot,k}^{(i)})^* \|_F^2 + \alpha_4^* \\ & -\alpha_1^* + \alpha_2^* + \alpha_3^* = 0; \quad \alpha_4^* + \alpha_5^* = 0 \end{aligned}$$

where the variables with superscript * denote the optimal solutions. According to (10), we know that

$$\begin{aligned} 0 \in & \quad \partial \mathcal{L}(W^{k+1}) + \rho(W^{k+1} - P^k - Q^k) + \alpha_4^k \cdot \partial g(X, W^{k+1}) \\ & + \alpha_1^k + \rho(\partial g(W^{k+1}) \cdot g(X, W^{k+1}) + \beta^k \partial g(X, W^{k+1})). \end{aligned}$$

According to (15), the above equation becomes

$$\begin{aligned} 0 \in & \, \partial \mathcal{L}(W^{k+1}) + \alpha_1^{k+1} + \alpha_4^{k+1} \partial g(X, W^{k+1}) + \rho(P^{k+1} - P^k) \\ & + \rho(Q^{k+1} - Q^k) + \rho(\beta^k - \beta^{k+1}) \cdot \partial g(X, W^{k+1}) \\ = & \, \partial \mathcal{L}(W^{k+1}) + \alpha_1^{k+1} + \alpha_4^{k+1} \partial g(X, W^{k+1}) + d_W \qquad (18) \end{aligned}$$

where the dual residual according to W is $d_W = \rho\ (P^{k+1} - P^k) + \rho(Q^{k+1} - Q^k) + \rho(\beta^k - \beta^{k+1}) \cdot \partial g(X, W^{k+1})$.

Similarly, the dual residuals with respect to other parameters are $d_P = \rho(U^{k+1} - U^k) + \rho(V^{k+1} - V^k) + \rho(Q^{k+1} - Q^k)$ and $d_\beta = \rho(\beta_+^{k+1} - \beta_+^k)$.

B. Parameter Optimization of MREF-II

1) Objective Function of MREF-II: The objective function in (5) formulates the heterogeneous task relationships into regularization terms and equality constraints based on Corollary 1. Ideally, this constraint should be satisfied as follows:

$$f(\tilde{X}_{t,s_n} \tilde{W}_{s_n}) = \max(\{ f(\tilde{X}_{t,s'_{n+1}} \tilde{W}_{s'_{n+1}}) | s'_{n+1} \sqsubseteq s_n \}) \qquad (19)$$

where $\tilde{X}_{t,s_n}$ and $\tilde{X}_{t,s'_{n+1}}$ are the input values in ideal situation, of which the corresponding parameters are $\tilde{W}_{s_n}$ and $\tilde{W}_{s_{n+1}}$. However, the actual data set contains noise, and this type of strict equality constraint will be very sensitive to the existence of noise in the input data. We thus assume the relationship between the ideal and actual value is $X_{t,s_n} \cdot W_{s_n} = \tilde{X}_{t,s_n} \cdot \tilde{W}_{s_n} + R_{t,s_n}$ and $X_{t,s'_{n+1}} \cdot W_{t,s'_{n+1}} = \tilde{X}_{t,s'_{n+1}} \cdot \tilde{W}_{t,s'_{n+1}} + R_{t,s'_{n+1}}$, where R_{t,s_n} and $R_{t,s'_{n+1}}$ are the noise terms. Because $f(\cdot)$ is a strictly monotonic increase, it is easy to see that (19) is equivalent to the following equation:

$$\tilde{X}_{t,s_n} \cdot \tilde{W}_{s_n} = \max(\{\tilde{X}_{t,s'_{n+1}} \cdot \tilde{W}_{s'_{n+1}} \,|\, s'_{n+1} \sqsubseteq s_n\}). \quad (20)$$

Considering (20) and the problem of overlapping coupled subtrees introduced in Section V-A1, (5) can be transformed to the following equation, in order to be solved by ADMM:

$$\begin{aligned} \min_{W} \ \mathcal{L}(W) + \gamma_0 \textstyle\sum_k^K \sum_{\mathcal{T}_i}^{\mathcal{T}_\mathcal{O}} \| U_{\mathcal{T}_i,k} \|_F + \gamma_1 \sum_k^K \sum_{\mathcal{T}_i}^{\mathcal{T}_\mathcal{E}} \| V_{\mathcal{T}_i,k} \|_F \\ + \gamma_2 \textstyle\sum_k^K \sum_n^N \| Q_{\cdot,k}^{(n)} \|_F + \gamma_3 \| R \|_F^2 \\ s.t.\ W = P + Q,\ P = U,\ P = V, \\ X_{t,s'_{n+1}} \cdot W_{t,s'_{n+1}} = Z_{t,s'_{n+1}} + R_{t,s'_{n+1}} \\ Z_{t,s_n} = \max(\{ D_{t,s'_{n+1}} \,|\, s'_{n+1} \sqsubseteq s_n \}),\ Z = D \quad (21) \end{aligned}$$

where $Z \in \mathbb{R}^{T\times S}$, of which $Z_{t,s_n} = \tilde{X}_{t,s_n} \cdot \tilde{W}_{s_n}, (t \in T, s_n \in S)$ and thus $Z_{t,s'_{n+1}} = \tilde{X}_{t,s'_{n+1}} \cdot \tilde{W}_{s'_{n+1}}$. Similarly, we denote $D \in \mathbb{R}^{T\times S}$ and $R \in \mathbb{R}^{T\times S}$, where D_{t,s_n} and R_{t,s_n} are their elements corresponding to time t and location s_n . $\gamma_3 = 1/\big(T \cdot \sum_{n=2}^{N} | S_n |\big)$. The corresponding augmented Lagrangian format is as follows:

$$\begin{aligned} \min_{\Theta'} \ \mathcal{L}(W) &+ \gamma_0 \textstyle\sum_{k,\mathcal{T}_i}^{K,\mathcal{T}_\mathcal{O}} \| U_{\mathcal{T}_i,k} \|_F + \gamma_1 \sum_{k,\mathcal{T}_i}^{K,\mathcal{T}_\mathcal{E}} \| V_{\mathcal{T}_i,k} \|_F \\ &+ \gamma_2 \textstyle\sum_k \sum_n^N \| Q_{\cdot,k}^{(n)} \|_F + \gamma_3 \| R \|_F^2 \\ &+ \frac{\rho}{2} \| W - P - Q + \Gamma_1 \|_F^2 \\ &+ \frac{\rho}{2} \| P - U + \Gamma_2 \|_F^2 + \frac{\rho}{2} \| P - V + \Gamma_3 \|_F^2 \\ &+ \frac{\rho}{2} \| Z - D + \Gamma_6 \|_F^2 \\ &+ \textstyle\sum_{t,s_n}^{T,S} \frac{\rho}{2} \| Z_{t,s_n} - X_{t,s_n} \cdot W_{t,s_n} + R_{t,s_n} + [\Gamma_4]_{t,s_n} \|_F^2 \\ &+ \textstyle\sum_{t,s_n}^{T,S} \frac{\rho}{2} \| Z_{t,s_n} - \max(\{ D_{t,s'_{n+1}} \,|\, s_{n+1} \sqsubseteq s_n \}) \\ &+ \Gamma_5 \|_F^2 \qquad (22) \end{aligned}$$

where $\Theta' = \{W, P, U, V, Q, Z, R, D, \Gamma\}$ are the parameters to be optimized. $\Gamma = \{\Gamma_i\}_{i=1}^6$ is the set of Lagrangian multipliers that are the dual variables of ADMM and ρ is the step size of the dual step.

2) Optimization Algorithm of MREF-II: The parameters of Θ' are alternately solved by the proposed algorithm, known as nonsmooth constrained multitask learning. It alternately optimizes each of the parameters in Θ' until the residuals of the variables are smaller than a predefined value. The detailed optimization steps are described in detail in the following. Note that the updates of P, Q, U, and V are the same as those in Algorithm 1 and thus here we only focus on the remaining parameters, namely W, Z, R, and D.

1) Update W: The optimization of parameter W is a generalized linear regression with least squares loss functions

$$\begin{aligned} W \leftarrow \ \operatorname*{argmin}_{W} \ & \mathcal{L}(W) + \frac{\rho}{2} \| W - P - Q + \Gamma_1 \|_F^2 \\ & + \textstyle\sum_{t,s_n}^{T,S} \frac{\rho}{2} \| Z_{t,s_n} - X_{t,s_n} \cdot W_{t,s_n} + R_{t,s_n} \\ & + [\Gamma_4]_{t,s_n} \|_F^2. \qquad (23) \end{aligned}$$

Similar to (10), a second-order Taylor expansion is performed, where we approximate the Hessian using a multiple of the identity with an upper bound of $(1/4)I$.

2) Update Z and R

$$\begin{aligned} Z \leftarrow \ \operatorname*{argmin}_{Z} \ & \frac{\rho}{2} \| Z - D + \Gamma_6 \|_F^2 \\ & + \textstyle\sum_{t,s_n}^{T,S} \frac{\rho}{2} \| Z_{t,s_n} - X_{t,s_n} \cdot W_{t,s_n} + R_{t,s_n} + [\Gamma_4]_{t,s_n} \|_F^2 \\ & + \textstyle\sum_{t,s_n}^{T,S} \frac{\rho}{2} \| Z_{t,s_n} - \max(\{ D_{t,s'_{n+1}} \,|\, s'_{n+1} \sqsubseteq s_n \}) + \Gamma_5 \|_F^2 \end{aligned} \qquad (24)$$

where the analytical solution is

$$Z_{t,s_n} = \begin{cases} [D - \Gamma_6 - R - \Gamma_4]_{t,s_n} + X_{t,s_n} W_{t,s_n}, & n = N \\ [D - \Gamma_6 - R - \Gamma_4 - \Gamma_5]_{t,s_n} + X_{t,s_n} W_{t,s_n} & \\ \quad + \max(\{ D_{t,s'_{n+1}} \,|\, s'_{n+1} \sqsubseteq s_n \}) & n = \leq N - 1. \end{cases}$$

R can be easily updated as follows:

$$\begin{aligned} R \leftarrow \ \operatorname*{argmin}_{R} \ & \textstyle\sum_{t,s_n}^{T,S} \frac{\rho}{2} \| Z_{t,s_n} - X_{t,s_n} \cdot W_{t,s_n} + R_{t,s_n} \\ & + [\Gamma_4]_{t,s_n} \|_F^2 + \gamma_3 \| R \|_F^2 \qquad (25) \end{aligned}$$

where the analytical solution is $R_{t,s_n} = (\rho/\rho + \gamma_3)(X_{t,s_n} W_{t,s_n} - Z_{t,s_n} - [\Gamma_4]_{t,s_n})$.

3) Update D: To update D_{t,s_n}, when $n = 1$ we need to solve the following optimization problems:

$$D_{t,s_1}^{*} = \operatorname*{argmin}_{D_{t,s_1}} \frac{\rho}{2} \| Z_{t,s_1} - D_{t,s_1} + [\Gamma_6]_{t,s_1} \|_F^2 \qquad (26)$$

where the analytical solution can be easily obtained $D_{t,s_1} = Z_{t,s_1} + [\Gamma_6]_{t,s_1}$

When $n > 1$, we need to solve the following optimization problems:

$$\begin{aligned} D_{t,r(n)}^{*} = \ \operatorname*{argmin}_{D_{t,r(n)}} \ & \frac{\rho}{2} \sum_{s'_{n+1} \in r(n)} \| Z_{t,s'_{n+1}} - D_{t,s'_{n+1}} + [\Gamma_6]_{t,s'_{n+1}} \|_F^2 \\ & + \textstyle\sum_{t,s_n}^{T,S} \frac{\rho}{2} \| Z_{t,s_n} - \max(D_{t,r(n)}) + [\Gamma_5]_{t,s_n} \|_F^2 \qquad (27) \end{aligned}$$

where $r(n) = \{ s'_{n+1} \,|\, s'_{n+1} \sqsubseteq s_n \}$ denotes all the subregions of the location s_n, and $D_{t,r(n)} \in \mathbb{R}^{1\times|r(n)|}$ is a vector such that $D_{t,r(n)} = \{ D_{t,s'_{n+1}} \,|\, s'_{n+1} \sqsubseteq s_n \}$, where $|r(n)|$ is the number of all the subregions of s_n.

It can be seen that $D_{t,r(n)}$ has no simple closed-form solution due to its existence inside the "max" function. It is also difficult to solve efficiently using traditional solutions such as subgradient methods due to the difficulty in selecting an appropriate step size to ensure convergence. Even though convergence can be achieved, the process is very slow near the optimal point. Therefore, here we propose a new dynamic programming-based algorithm to solve this problem that finds the optimal solution efficiently. The method is as follows.

Denote $D^{\max}_{t,r(n)} = \max(D_{t,r(n)})$. For each element $D_{t,{s'}_{n+1}} \in D^{*}_{t,r(n)}$, there are only two possible situations: 1) less than the max value $D_{t,{s'}_{n+1}} < D^{\max}_{t,r(n)}$; and 2) equal to the max value $D_{t,{s'}_{n+1}} = D^{\max}_{t,r(n)}$. Therefore, solving $D_{t,{s'}_{n+1}}$ is equal to the following two sequential subroutines: Subroutine 1) find the solution for each situation; and Subroutine 2) identify which situation each element $D_{t,{s'}_{n+1}}$ belongs to.

Subroutine 1: According to (27), if $D_{t,{s'}_{n+1}}$ belongs to the first situation, namely $D_{t,{s'}_{n+1}} < D^{\max}_{t,r(n)}$, it is easy to see its closed-form solution $D_{t,{s'}_{n+1}} = Z_{t,{s'}_{n+1}} + [\,\Gamma_6\,]_{t,{s'}_{n+1}}$. After substituting $D_{t,{s'}_{n+1}}$ with this solution, the objective function in (27) is equivalent to the following:

$$\begin{aligned} h(M, D^{\max}_{t,r(n)}) = \quad & |M|(Z_{t,s_n} - D^{\max}_{t,r(n)} + [\,\Gamma_5\,]_{t,s_n})^2 \\ & + \textstyle\sum_{k\in M}(D^{\max}_{t,r(n)} - \zeta_{t,k})^2 \end{aligned} \tag{28}$$

where $\zeta_{t,k} = Z_{t,[r(n)]_k} + [\,\Gamma_6\,]_{t,[r(n)]_k}$ and we also define $\zeta_t = Z_{t,r(n)} + [\,\Gamma_6\,]_{t,r(n)}$. M consists of the elements among $r(n)$ that belong to the second situation and thus the complementary set $\bar{M} = r(n) - M$ represents the first situation. Therefore, solving the second situation is equivalent to finding the optimal $D^{\max}_{t,r(n)}$ that minimizes $h(M, D^{\max}_{t,r(n)})$. By applying the least squares method to (28), the closed-from solution of $D^{\max}_{t,r(n)}$ that minimizes $h(M, D^{\max}_{t,r(n)})$ becomes the following:

$$D^{\max}_{t,r(n)} = \sum_{i\in M}(\zeta_{t,i} + Z_{t,s_n} + [\,\Gamma_5\,]_{t,s_n})/(|M| + 1) \tag{29}$$

Subroutine 2: Solving Subroutine 2 is equivalent to searching for the best subset M that minimizes the objective function $h(M, D^{\max}_{t,r(n)})$ defined in (28). Note that we have $D^{\max}_{t,r(n)} > D_{t,[r(n)]_j} = \zeta_{t,j}$, where $j \in \bar{M}$. Hence, it is easy to see from (28) that in order to minimize $h(M, D^{\max}_{t,r(n)})$, the elements in M should be those with the largest $\zeta_{t,j}$'s.

So now the remaining problem is to determine how many elements in $r(n)$ with the largest $\zeta_{t,j}$ should be selected, namely determining $|M|$. Since M only contains the largest $\zeta_{t,j}$'s, $h(M, D^{\max}_{t,r(n)})$ increases monotonously with $|M|$. This means we need to search for the smallest $|M|$ that satisfies the condition of the second situation. Thus, the best size $|M|$ that minimizes the objective function $h(M, D^{\max}_{t,r(n)})$ is equal to the optimal solution of the following objective:

$$\begin{aligned} & \arg\min_{k} k \\ s.t.\ \ & \textstyle\sum_{i=1}^{k}(\tilde{\zeta}_{t,i} + Z_{t,s_n} + [\,\Gamma_5\,]_{t,s_n})/(k+1) \ > \ \tilde{\zeta}_{k-1} \end{aligned} \tag{30}$$

where $\tilde{\zeta}_t \in \mathbb{R}^{1\times|\zeta_t|}$ is defined as an ordered list which is just ζ_t being sorted in decreasing order. The objective function in (30) can thus be efficiently solved by a simple linear search.

4) Update $\Gamma_i (i = 1,\ldots,6)$: The updating of the dual variables $\Gamma_i, (i = 1,\ldots,6)$ are as follows:

$$\begin{aligned} \Gamma_1 \quad & \leftarrow \quad \Gamma_1 + (W - P - Q), \quad \Gamma_6 \leftarrow \Gamma_6 + (Z - D), \\ \Gamma_2 \quad & \leftarrow \quad \Gamma_2 + (P - U), \quad \Gamma_3 \leftarrow \Gamma_3 + (P - V), \\ [\,\Gamma_4\,]_{t,s_n} \quad & \leftarrow \quad [\,\Gamma_4\,]_{t,s_n} + (Z_{t,s_n} - X_{t,s_n} \cdot W_{t,s_n} + R_{t,s_n}), \\ [\,\Gamma_5\,]_{t,s_n} \quad & \leftarrow \quad [\,\Gamma_5\,]_{t,s_n} + (Z_{t,s_n} - \max(\{\,D_{t,{s'}_{n+1}} \,|\, {s'}_{n+1} \sqsubseteq s_n\,\})). \end{aligned} \tag{31}$$

C. Algorithm Analysis

1) Time Complexity Analysis: For the optimization of the model MREF-I, the time complexities for updating W, P, and Q are $O(|S| \cdot |T| \cdot K)$, $O(|S| \cdot K)$, and $O(K \cdot |S|)$, respectively. The time complexity for updating U is $|\mathcal{T}_{\mathcal{O}}| \cdot K$ and for V is $|\mathcal{T}_{\mathcal{E}}| \cdot K$. Combining these we get the summation $O(\cdot (2|S| - 1 - |S_N|) \cdot K)$. The time complexity for updating β and β_+ are both $O(|S| \cdot |T|)$. In all, the total time complexity is $O(L \cdot |S| \cdot |T| \cdot K)$, where L is the number of iterations of the ADMM iterations.

For the optimization of the model MREF-II, the time complexities for updating W, P, and Q are $O(|S| \cdot |T| \cdot K)$, $O(|S| \cdot K)$, and $O(\gamma K \cdot |S|)$, respectively. The time complexity for updating both U and V is $O(\gamma \cdot (2|S| - 1 - |S_N|) \cdot K)$. The time complexity for updating R and Z are both $O(|S| \cdot |T|)$. Finally, the time complexity for D-update is $O(|S| \log n_r + |T| \cdot |S|)$, where $|S| \log n_r$ is for sorting the variable array ζ_t. n_r is the average number of the children of a location. In all, the total time complexity is $O(L \cdot |S| \cdot |T| \cdot K + |S| \log n_r)$, where L is the number of iterations of the ADMM iterations.

This means that the major difference in the computational time between the proposed two models lies in the different ways in which the parent–child constraint is treated. According to Theorem 1 and Corollary 1, it is obvious that MREF-II's constraint is tighter than that of MREF-I. Due to the additional constraint imposed by Corollary 1, the constraint of MREF-II also contains a nonsmooth term, which is the max function and is more challenging to solve. However, due to our efficient dynamic programming-based algorithm, the optimization to this max function (i.e., the update of the variable D) is accelerated. Compared to MREF-I, there is only an additional computation time of $O(\log n_r)$ for MREF-II. Typically, in practice, $\log n_r \ll |S| \cdot K$. This indicates that the computational time of MREF-II is generally as efficient as MREF-I, even though it solves a much harder problem.

2) Applicability Analysis: As mentioned above, the difference between the models MREF-I and MREF-II lies in their respective assumptions for the parent–child relationship; MREF-I is based on Theorem 1 while MREF-II is based on Corollary 1. Specifically, MREF-I assumes that if there is no event for a location, then there is no event in any of its subregions; MREF-II, in addition to this assumption, also expects that if there is no event in any subregion of a location, then there is no event in this location.

Therefore, if every social media message (e.g., tweet) contains spatial information for all the geographical levels (including city level, state level, country level), then both of these assumptions are applicable. In this case, the use of model MREF-II is preferable because it effectively considers the prior knowledge based on a more comprehensive assumption to ensure good model generalization. However, in social media like Twitter, messages seldom contain complete spatial information that includes city-level or state-level locations. Due to this scarcity of spatial information at the finer grained geographical level, sometimes even there is an event happening in a city, especially a small city, it is hard to detect it in city level but more likely to show up at the state or country level because the tweets from that city might only contain state or country-level spatial information. This means the discernible georesolution is restricted to state or country level. Therefore, in such a situation, the assumption inherent in MREF-II is not preferable because even though there is no event detected from any city-level model, there still might be an event detected using the state or country-level models. Hence, if such situations appear very frequently in a data set, then the MREF-I model is preferable. In applications where the events usually tale place in more populous locations (e.g., bigger cities), however, this situation arises less frequently because such locations generally have sufficient social media messages geocoded in finer granularity. Then, MREF-II might still be advantageous.

VI. EXPERIMENTS

In this section, the proposed models, MREF-I and MREF-II, are evaluated on several real-world data sets from two different domains. After the experimental setup has been introduced in Section VI-A, the effectiveness of the methods is evaluated against several existing methods for different spatial resolutions, along with an analysis of the models' performance on precision-recall curves for all the comparison methods in Section VI-B. The parameter sensitivity analysis is presented in Section VI-C.

A. Experimental Setup

1) Data Sets and Labels: The experimental evaluations in this study are based on nine data sets on different domains. Of these, eight data sets are used for event forecasting under the civil unrest domain while the other is applied to the influenza outbreaks domain. For the civil unrest domain data sets, Table 2 shows the specific country from which the Twitter data was gathered for each data set. The raw Twitter data are collected from the Datasift Twitter collection engine and divided into periods for the training and test sets, as shown in Table 2. The data collection is partitioned into a sequence of date-interval bins for forecasting day by day. The event forecasting results are validated against a labeled events set, known as the gold standard report (GSR), which is publicly available from the Harvard Dataverse: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/EN8FUW. GSR is a collection of civil unrest news reports manually labeled by social science domain experts from the most influential newspaper outlets in Latin America [2], as shown in Table 1. For civil unrest forecasting, three spatial resolutions are considered, namely country level, state level, and city level. An example of a labeled GSR event is given by the tuple: (CITY="Hermosillo," STATE = "Sonora," COUNTRY = "Mexico," DATE = "2013-01-20").

Table 1 Data Sets and Labels

Dataset	#Tweets	Label sources*	#Events
Brazil	185,286,958	O Globo; O Estado de São Paulo; Jornal do Brasil	3417
Colombia	158,332,002	El Espectador; El Tiempo; El Colombiano	1287
Ecuador	50,289,195	El Universo; El Comercio; Hoy	511
El Salvador	21,992,962	El Diáro de Hoy; La Prensa Gráfica; El Mundo	730
Mexico	197,550,208	La Jornada; Reforma; Milenio	5907
Paraguay	30,891,602	ABC Color; Ultima Hora; La Nacíon	2114
Uruguay	10,310,514	El Paí; El Observador	664
Venezuela	167,411,358	El Universal; El Nacional; Ultimas Notícias	3320
U.S.	11,993,211,616	CDC Flu Activity Map	1027

*In addition to the top three domestic news outlets, the following news outlets are included: *The New York Times; The Guardian; The Wall Street Journal; The Washington Post; The International Herald Tribune; The Times of London; Infolatam.*

For the data set applied to the influenza outbreaks domain, we collected tweets containing at least one of 124 predefined flu-related keywords (e.g., "cold," "fever," and "cough") provided by Paul and Dredze [61]; the time period covered by this data set is also shown in Table 2. The data collection for the influenza data set are partitioned into a sequence of week-interval bins for week-wise forecasting. The predictions were validated against the flu statistics reported by the Centers for Disease Control and Prevention (CDC), downloadable via the link: https://gis.cdc.gov/grasp/fluview/main.html. CDC typically organizes the influenza surveillance data by HHS regions,[1] which groups U.S. states into ten regions. CDC publishes the weekly influenza-like illness (ILI) activity level within each state in the United States based on the proportion of 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 considered the target for forecasting. In forecasting influenza outbreaks, three spatial resolutions are considered, namely country level, HHS-region level, and state level. An example of a CDC flu outbreak event is: (STATE = "California," HHS_REGION = "Region 9," COUNTRY = "United States," WEEK = "01-09-2013 to 01-15-2013").

[1] HHS regions: http://www.hhs.gov/about/agencies/regional-offices/

Table 2 Domains for the Experimental Evaluations

Domain	Training period	Test period	Spatial resolution	Data sets
Civil Unrest	2013-01-01∼2013-12-31	2014-01-01∼2014-12-31	country, state, city	Brazil, Colombia, Ecuador, El Salvador, Mexico, Paraguay, Uruguay, Venezuela
Influenza	2011-01-01∼2013-12-31	2014-01-01∼2014-12-31	country, HHS-region, state	the United States

2) Parameter Settings and Metrics: There is one tunable parameter in our MREF-I and MREF-II models, namely the regularization parameter γ. This parameter was set for all nine data sets based on a tenfold cross validation on the training set.

In the experiment, the event forecasting task is to predict whether there will be an event in the next time step for a specific location at several different spatial resolutions. For civil unrest data sets, a time step is one day and the spatial resolutions are country level, state level, and city level. For disease outbreaks, a time step is one week and spatial resolutions are country level, HHS-region level, and state level. For each spatial resolution, a predicted event is matched to a GSR event if the location for the current spatial resolution is matched and the date is within two time steps before the actual event occurrence; otherwise, it is considered a false forecast. To validate the prediction performance, different metrics are adopted. Precision designates the fraction of all the predictions that match actual events that occur. Recall denotes the percentage of all the actual events that have been successfully predicted. In addition, another metric, F-measure, is defined as the harmonic mean of precision and recall F-measure $= 2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precsion} + \text{Recall})$.

3) Comparison Methods: The following methods are included in the performance comparison.

a) LASSO [2]. Different LASSO models were built for each of the corresponding spatial resolutions. The feature set for each is represented by the set of keyword counts. To tune the regularization parameter, different values from the set $\mathscr{R}_p = \{0.01, 0.02, \ldots, 0.1, 0.2, \ldots, 1, 2 \ldots, 10\}$ were tested based on a tenfold cross validation on the training set. Specifically, for each data set, we partitioned the training set into ten equal segments along the time line. We then used nine segments for training the model and the remaining segment for validating the results, giving a total of ten rounds by iterating the segment used for the validation. For each round a validation performance is obtained and our focus is on the average performance across all ten rounds. The regularization parameter was set at 0.15 because this value achieved the best average performance for all ten rounds.

b) Multitask learning (MTL) [1]. In the multitask model, each task consists of the forecasting for each location and spatial resolution. Keyword counts are the features. As in LASSO, the values in $\mathcal{R}_p$ were tested to select the regularization parameters. Finally, the parameters were set as $\lambda_1 = 0.015$ and $\lambda = 0.001$, which were selected because they achieved the best overall performance in the tenfold cross validation.

c) Tree-guided group LASSO for multitask learning (TMTL) [43]. Here the relationships among the tasks follow the geohierarchy defined in Fig. 2. Specifically, each subtree consists of a parent task as root and all of its children as leaves, as defined in Definition 2. Keyword counts are the features. As with LASSO, the values in $\mathcal{R}_p$ were tested to select the regularization parameter, which was set as $\lambda = 0.3$ because this yielded the best overall performance in the tenfold cross validation.

d) Autoregressive exogenous (ARX) [9]. For each separate location, the count of future events is predicted based on the counts of both historical events and tweets indexed by the keywords. When forecasting, an output not less than "1" indicates event occurrence; otherwise, no event is deemed to have occurred.

e) Logistic regression (LR) [62]. For each spatial resolution, LR utilizes a logit function to map the tweet observations onto future event occurrences ("0" denotes no occurrence, "1" denotes occurrence). The input features here are the counts of keywords.

f) Latent Direchlet allocation-based logistic regression (LDA-LR) [6]. After extracting the latent topics by LDA from the tweets, the LDA-LR model was built on features that represent the proportions of the latent topics. Individual models were built for each spatial resolution. To set the value of the number of topics, several values $\{10, 20, \ldots, 100\}$ were tested based on a tenfold cross validation. The number of topics was set at 30 because this achieved the best performance in the cross validation.

g) Kernel-density-estimation-based logistic regression (KDE-LR) [16]. This approach utilizes KDE-smoothed historical-event counts and the proportions of latent topics as features, and builds a model for each spatial resolution. Similar to LDA-LR, the values $\{10, 20, \ldots, 100\}$ were tested based on a tenfold cross validation. Finally, the number of topics was set at 30 because it achieved the best performance in the cross validation.

B. Performance

In this section, the performances of all the methods are evaluated and compared. First, the specific spatial event forecasting performance for different spatial resolutions is

Table 3 Event Forecasting Performance on Multiple Civil Unrest Data Sets

City Level (precision, recall, F-measure)								
Method	Brazil	Colombia	Ecuador	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
ARX	0.63,0.47,0.54	0.30,0.40,0.35	0.33,0.47,0.39	0.44,0.42,0.43	**0.43**,0.20,0.27	0.52,0.27,0.36	0.53,0.60,0.56	0.51,0.23,0.32
LR	0.43,0.41,0.42	0.33,0.38,0.36	0.37,0.39,0.38	0.50,0.34,0.41	0.30,0.11,0.16	0.52,0.23,0.32	0.48,0.47,0.48	0.40,0.33,0.36
KDE-LR	0.99,0.01,0.02	**0.68**,0.01,0.01	0.16,0.13,0.15	0.28,0.09,0.14	0.02,0.15,0.04	0.04,0.35,0.07	0.13,**0.93**,0.22	0.69,0.03,0.06
LDA-LR	**1.00**,0.01,0.02	0.01,**0.63**,0.02	0.16,0.13,0.15	0.26,0.09,0.13	0.01,0.19,0.02	0.04,0.36,0.07	0.14,**0.93**,0.24	**0.99**,0.04,0.07
LASSO	0.74,0.45,0.56	0.40,0.41,0.40	0.34,0.42,0.38	0.62,0.36,0.46	0.18,0.42,0.25	0.72,0.25,0.37	0.61,0.46,0.52	0.19,**0.80**,0.31
MTL	0.68,0.48,0.56	0.37,0.44,**0.41**	0.24,**0.55**,0.34	0.42,**0.45**,0.43	0.42,0.24,0.31	0.57,0.29,0.38	0.60,0.54,0.56	0.37,0.45,0.41
TMTL	0.46,0.42,0.44	0.36,0.34,0.35	0.37,0.43,0.40	0.57,0.43,0.49	0.29,0.25,0.27	0.25,**0.42**,0.31	0.60,0.64,0.62	0.41,0.58,**0.48**
MREF-I	0.79,0.47,**0.59**	0.37,0.39,0.38	**0.38**,0.43,0.40	0.58,0.43,0.50	0.29,0.30,0.29	**0.75**,0.26,0.39	**0.66**,0.60,0.63	0.24,0.49,0.33
MREF-II	0.60,**0.50**,0.55	0.43,0.34,0.38	0.37 0.46 **0.42**	**0.66**,0.43, **0.52**	0.21,**0.74**,**0.33**	0.65 0.31 **0.42**	0.56,0.78,**0.65**	0.39,0.47,0.43
State Level (precision, recall, F-measure)								
Method	Brazil	Colombia	Ecuador	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
ARX	0.73,0.63,0.67	0.35,0.41,0.38	0.34,0.51,0.41	0.53,0.55,0.54	0.55,0.39,0.46	0.48,0.42,0.45	0.33,0.57,0.42	0.63,0.41,0.50
LR	0.53,0.56,0.55	0.34,**0.54**,0.41	0.21,0.69,0.32	0.51,0.53,0.52	0.30,0.89,0.45	0.58,0.37,0.45	0.49,0.45,0.47	0.55,0.48,0.51
KDE-LR	**1.00**,0.08,0.16	0.02,0.18,0.04	0.10,0.38,0.16	0.10,0.29,0.14	0.93,0.23,0.37	**1.00**,0.12,0.21	0.23,0.20,0.21	0.37,0.37,0.37
LDA-LR	**1.00**,0.08,0.16	**0.99**,0.05,0.09	0.08,**0.79**,0.15	0.08,0.32,0.12	**0.94**,0.23,0.37	**1.00**,0.12,0.21	0.19,0.21,0.20	0.41,0.40,0.41
LASSO	0.70,0.67,0.68	0.43,0.43,0.43	0.34,0.50,0.40	**0.64**,0.44,0.52	0.41,0.69,0.52	0.31,**0.77**,0.44	0.52,0.49,0.50	**0.64**,0.40,0.49
MTL	0.60,**0.72**,0.66	0.40,0.50,**0.45**	**0.39**,0.51,**0.44**	0.55,0.51,0.53	0.70,0.30,0.42	0.65,0.37,0.47	0.58,0.55,0.56	0.57,0.54,0.55
TMTL	0.61,0.36,0.45	0.37,0.38,0.37	0.36,0.49,0.41	0.61,0.51,**0.56**	0.42,0.34,0.38	0.43,0.50,0.46	0.52,0.52,0.52	0.54,0.37,0.44
MREF	0.75,0.64,**0.69**	0.36,0.51,0.43	0.37,0.49,0.42	0.27,**0.59**,0.37	0.35,0.77,0.49	0.58,0.41,0.48	**0.63**,**0.58**,**0.61**	0.53,0.42,0.47
MREF-II	0.65,0.71,0.68	0.41,0.45,0.43	0.37 0.49 0.42	0.46,0.58,0.52	0.44,**0.98 0.61**	0.70,0.39,**0.50**	**0.63**,**0.58**,**0.61**	0.57,**0.61**,**0.59**
Country Level (precision, recall, F-measure)								
Method	Brazil	Colombia	Ecuador	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
ARX	0.93,**1.00**,0.96	0.73,**0.97**,0.83	0.53,0.87,0.65	0.66,0.97,0.78	0.99,**1.00**,**1.00**	0.90,0.87,0.88	0.60,0.90,0.72	0.90,0.98,0.94
LR	0.95,**1.00**,0.97	0.79,**0.97**,0.87	0.56,**0.95**,0.70	0.78,0.82,0.80	**1.00**,0.98,0.99	0.89,0.97,0.93	0.63,0.93,0.75	0.92,0.96,0.94
KDE-LR	0.97,0.96,0.97	0.93,0.80,0.86	0.88,0.59,0.70	**0.85**,0.76,0.80	**1.00**,0.99,**1.00**	**1.00**,0.85,0.92	**0.97**,0.69,0.80	**1.00**,0.91,0.95
LDA-LR	0.96,0.96,0.96	**0.95**,0.82,0.88	**0.95**,0.57,0.71	0.82,0.78,0.80	0.93,**1.00**,0.96	0.91,0.92,0.91	0.94,0.70,0.80	**1.00**,0.91,0.95
LASSO	0.95,0.99,0.97	0.81,0.95,0.87	0.59,0.93,0.72	0.75,0.86,0.80	0.99,0.99,0.99	0.90,0.99,0.94	0.54,0.99,0.70	0.93,0.99,0.96
MTL	**0.98**,0.97,0.97	0.83,0.94,0.88	0.58,0.88,0.70	0.79,0.87,0.83	0.99,0.99,0.99	0.92,0.94,0.93	0.68,0.75,0.71	0.95,0.95,0.95
TMTL	0.82,0.98,0.89	0.88,0.92,**0.90**	0.67,0.87,0.76	0.70,0.87,0.78	**1.00**,**1.00**,**1.00**	0.94,0.98,**0.96**	0.67,0.72,0.70	0.88,**1.00**,0.94
MREF	0.97,**1.00**,**0.98**	0.86,0.94,**0.90**	0.66,0.91,0.76	0.76,**0.98**,0.86	**1.00**,**1.00**,**1.00**	0.93,0.99,**0.96**	0.69,0.97,0.81	0.96,**1.00**,**0.98**
MREF-II	0.93,**1.00**,0.97	0.82,0.96,0.89	0.68,0.91,**0.78**	0.78,**0.98**,**0.87**	0.99,**1.00**,**1.00**	0.90,**1.00**,**0.96**	0.70,**1.00**,**0.82**	0.95 0.99 0.97
Overall (precision, recall, F-measure)								
Method	Brazil	Colombia	Ecuador	El Salvador	Mexico	Paraguay	Uruguay	Venezuela
ARX	0.76,0.70,0.73	0.46,0.59,0.52	0.40,0.62,0.49	0.54,0.65,0.59	0.66,0.53,0.59	0.63,0.52,0.57	0.49,0.70,0.58	0.68,0.54,0.60
LR	0.64,0.66,0.65	0.49,**0.63**,0.55	0.38,**0.68**,0.49	0.60,0.56,0.58	0.53,0.66,0.59	0.66,0.52,0.58	0.53,0.62,0.57	0.62,0.59,0.60
KDE-LR	**0.99**,0.35,0.52	0.54,0.33,0.41	0.38,0.37,0.37	0.41,0.38,0.39	0.65,0.46,0.54	0.68,0.44,0.53	0.44,0.61,0.51	0.69,0.44,0.54
LDA-LR	**0.99**,0.35,0.52	**0.65**,0.50,**0.57**	0.40,0.50,0.44	0.39,0.40,0.39	0.63,0.47,0.54	0.65,0.47,0.55	0.42,0.61,0.50	**0.80**,0.45,0.58
LASSO	0.80,0.70,0.75	0.55,0.60,**0.57**	0.42,0.62,0.50	**0.67**,0.55,0.60	0.53,0.69,0.60	0.64,**0.67**,**0.65**	0.56,0.65,0.60	0.59,**0.73**,0.65
MTL	0.75,0.72,0.73	0.53,**0.63**,**0.57**	0.40,0.65,0.50	0.59,0.61,0.60	**0.70**,0.51,0.59	0.71,0.53,0.61	0.62,0.61,0.61	0.63,0.65,0.64
TMTL	0.63,0.59,0.59	0.54,0.55,0.54	**0.47**,0.60,0.53	0.62,0.60,0.61	0.57,0.53,0.55	0.54,0.63,0.58	0.60,0.63,0.61	0.61,0.64,0.62
MREF	0.84,0.70,**0.76**	0.53,0.61,**0.57**	**0.47**,0.61,0.53	0.53,**0.66**,0.59	0.55,0.70,0.61	**0.75**,0.55,0.63	**0.66**,0.72,0.67	0.58,0.65,0.61
MREF-II	0.73,**0.74**,0.73	0.55,0.58,**0.57**	**0.47**,0.62,**0.54**	0.63,**0.66**,**0.63**	0.55,**0.91**,**0.65**	**0.75**,0.57,0.63	0.63,**0.79**,**0.69**	0.64,0.69,**0.66**

discussed for civil unrest and influenza outbreaks, after which the precision–recall curves for the overall forecasting performance are examined.

1) Civil Unrest Event Forecasting Performance at Multiple Spatial Resolutions: In Table 3, the performances of our MREF-I and MREF-II and competing methods are compared for civil unrest event forecasting. Three metrics, namely precision, recall, and F-measure, were adopted to quantify the performance. The model performances for each of the spatial resolution levels and for the overall performance are shown in the table. The overall performance is the averaged performance across the different spatial resolutions.

Table 3 shows that the forecasting performance generally improves as the spatial resolution becomes coarser. For example, at the city level, the F-measure is typically 0.2–0.5, while for the state level, it is typically 0.3–0.6, and for the country level, it increases to about 0.8–1.0. Of these data sets, all the methods generally achieved better performances for Brazil, which is a large country with a large number of civil unrest events. In all, the proposed MREF-II model outperformed all the other methods in six data sets in overall performance, five data sets in city-level performance, three data sets in state-level performance, and five data sets in country-level performance. This is because MREF-II leverages the tasks' relationships in terms of geohierarchy, geo-resolution, and sufficiently considers the geo-parent–child constraints in Corollary 1. MREF-II also achieved good performance at the finest granularity, namely city level, outperforming the other methods by around 6% in five data sets and placing second in two more. This is because MREF-II can provide better predictions at the finest resolution by borrowing information from coarser resolutions, which effectively handles the shortage of finest level data in social media data sets. MREF-I achieved the second best performance by partially considering the parent–child constraints between locations. In general, the performance of the methods considering the feature sparsity is better than those achieved by existing methods. Specifically, LASSO, MTL, TMTL, and our MREF-I and MREF-II models all achieved better performances for each spatial resolution level than any of the others. LASSO, MTL, TMTL, MREF-I, and MREF-II obtained the best overall performance in seven of the eight data sets shown, demonstrating the effectiveness of utilizing regularization terms for filtering out unrelated features and ensuring the model's generalizability. Among these, the multitask-learning-based methods such as MTL, TMTL, MREF-I, and MREF-II also take into account the relatedness of different

Table 4 Forecasting Performance on Influenza Outbreak Data Set (Precision, Recall, and F-Measure)

Method	State-level	Region-level	Country-level	Overall	Runtime
ARX	0.09,0.52,0.15	0.12,0.86,0.21	0.58,0.98,0.73	0.26,**0.79**,0.39	**21 sec**
LR	0.08,0.20,0.12	0.26,0.48,0.33	0.85,0.95,**0.90**	0.40,0.54,0.46	37 sec
KDE-LR	0.74,0.07,0.12	0.24,0.21,0.22	**1.00**,0.53,0.69	0.66,0.27,0.38	2026 sec
LDA-LR	0.56,0.03,0.05	0.73,0.14,0.24	0.65,0.53,0.59	0.65,0.23,0.34	296 sec
LASSO	0.12,**0.84**,0.20	0.18,**1.00**,0.30	0.77,**1.00**,0.87	0.36,0.95,0.52	118 sec
MTL	**0.94**,0.12,0.21	**0.93**,0.18,0.21	**1.00**,0.68,0.81	**0.96**,0.33,0.49	45 sec
TMTL	0.15,0.54,0.24	0.49,0.35,0.41	0.70,**1.00**,0.82	0.45,0.63,0.49	656 sec
MREF	0.17,0.57,0.27	0.59,0.35,0.44	0.75,**1.00**,0.86	0.50,0.64,0.56	923 sec
MREF-II	0.24,0.45,**0.31**	0.84,0.38,**0.52**	0.79,0.94,0.86	0.62,0.59,**0.60**	518 sec

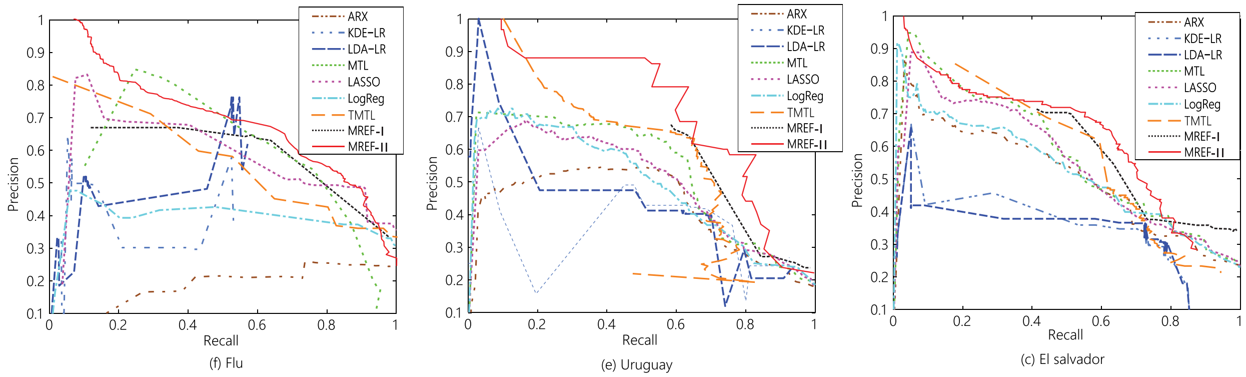
geographical locations, enabling them to handle the data scarcity inherent in small locations. Among the other methods LR, KDE-LR, and LDA-LR, all of which utilize the logistic regression framework, obtained similar performances, with KDE-LR and LDA-LR being especially close because they both consider the latent topics. The performance of ARX is not as good as regularization-based methods, which consistently outperformed ARX by 1%–18%.

2) Influenza Outbreak Event Forecasting Performance in Multiple Spatial Resolutions: In Table 4, the performances of MREF-I, MREF-II, and the competing methods are illustrated for influenza outbreak event forecasting. Their performances for all the different spatial resolutions and their overall performance were investigated.

As in Table 3, Table 4 shows that the forecasting performance generally becomes better when the spatial resolution becomes coarser. For example, at the state level, the F-measure is typically 0.1–0.2, at the region level, the F-measure is typically 0.2–0.4, and at the country level, the F-measure increases to about 0.6–0.9. In general, the performance of the methods utilizing regularization terms is better than other methods. In particular, LASSO, MTL, TMTL, MREF-I, and MREF-II achieve better performance at each spatial resolution level than the others. LASSO, MTL, TMTL, MREF-I, and MREF-II obtained the best overall performances, with F-measures of around 0.5, while the other methods were slightly lower, at around 0.3–0.4. This demonstrates the effectiveness of utilizing regularization terms for filtering out unrelated features and ensuring the model's generalizability. KDE-LR and LDA-LR again achieved similar performances because they both consider the latent topics as features. The performance of ARX was once more not as good as those of regularization-based methods, which outperformed it by over 20%. MREF-II outperformed all the other methods for overall performance, by 13% at the state level, 18% at the HHS-region level, and 7% overall. This again demonstrates the advantage enjoyed by MREF-II due to characterizing the location relatedness and sufficiently leveraging the parent–child constraint of the hierarchical locations.

3) Efficiency on Running Time: The rightmost column of Table 4 shows the training time efficiency comparison for forecasting influenza outbreaks. The running times on the test sets for all the comparison methods are effectively instantaneous (i.e., less than 0.01 second for one prediction) so are not provided here. According to Table 4, the running time of ARX was 21 s, outperforming the other methods. The running times achieved by all these methods were below at most 40 min for a huge three-year-long training set for week-wise event forecasting tasks, making this eminently practical for real-world applications. The efficiency evaluation results on civil unrest data sets followed a very similar pattern and thus these data are not provided here.

4) Event Forecasting Performance on Precision–Recall Curves: Fig. 4 illustrates the overall event forecasting performance on precision–recall curves for three data sets in two domains, namely civil unrest and influenza outbreaks. These curves were drawn by varying the boundary between values for positive and negative predictions. The other civil unrest data sets followed a similar pattern to the "El Salvador" and "Uruguay" data sets and thus are not provided here due to limited space. The overall performance shown is the averaged performance for different spatial resolutions. For the three data sets shown in Fig. 4, MREF-II generally outperformed the other methods because it is in most cases the closest to the (1,1) points in the plots. Moreover, the ROC curves for MREF-II were consistently above the other methods in these data sets,

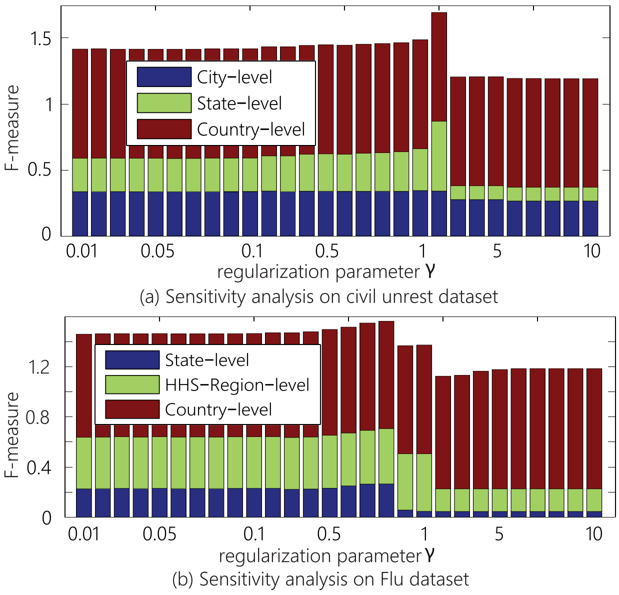


(f) Flu

(e) Uruguay

(c) El salvador

Fig. 4. ***Precision-recall curves for the performances on different data sets.***



(a) Sensitivity analysis on civil unrest dataset

(b) Sensitivity analysis on Flu dataset

Fig. 5. ***Sensitivity analyses on regularization parameter γ.***

when precision and recall vary. MREF-I consistently achieved the second best performance consistently. Other than MREF-I and MREF-II, the models MTL, TMTL, and LASSO achieve the most competitive results. The performance of KDE-LR and LDA-LR exhibit similar patterns because they utilize latent topics as features, unlike the other methods. Once again, ARX obtained a particularly poor performance for the flu data set, although it achieved an average performance in the other data sets.

C. Sensitivity Analyses

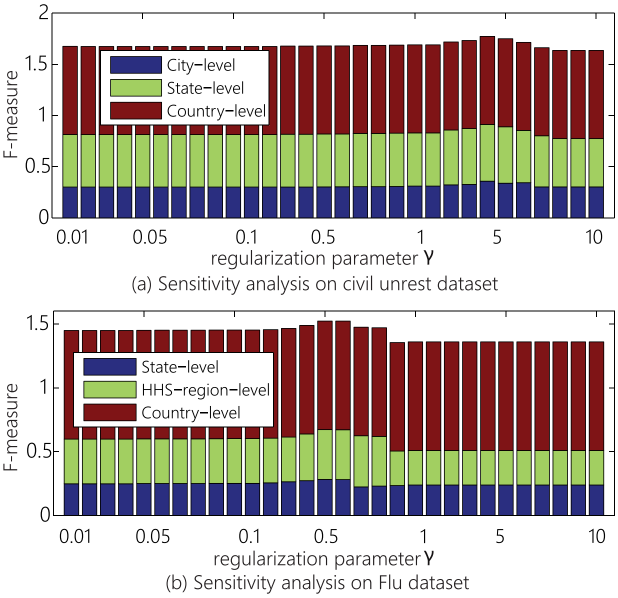
The proposed models MREF-I and MREF-II both include a tunable parameter, namely the regularization parameter γ, for the regularization terms. For civil unrest data sets, only the sensitivity analysis results for the Colombia data set are illustrated here as a typical example. The other civil unrest data sets follow very similar patterns.

1) Parameter Sensitivity of MREF-I: Fig. 6(a) illustrates the performance for civil unrest event forecasting of MREF-I versus γ. By varying it over a wide range from 0.01 to 10, the performance in terms of F-measures for civil unrest event forecasting can be illustrated for different spatial resolutions, namely city level, state level, and country level. For all spatial resolutions, the F-measures at $\gamma = 0.01{\sim}2$ are basically higher than that at $\gamma = 2{\sim}10$. The country level is more consistent across different values of γ. Starting from $\gamma = 0.01$, the F-measures in different spatial resolutions increase slightly as γincreases, reaching their highest values in the range of $\gamma = 0.5{\sim}2$, then decreasing at the city level and state level in the range of $\gamma > 2$. These results show that the value of the regularization parameter $\gamma$and the regularization terms influence the overall performance, and it should be neither too small nor too large.

Fig. 6(b) shows the performance of MREF-I versus γ for the influenza outbreaks data set. For all spatial resolutions, the F-measures at $\gamma = 0.01{\sim}0.8$ are essentially higher than those at $\gamma > 0.8$, especially for the city level and the state level; the country level is more consistent across different values of γ. Starting from $\gamma = 0.01$, the F-measures in different spatial resolutions increase slightly as γ increases until $\gamma = 0.8$, where the highest value is achieved, after which they decrease to smaller values for the F-measures at the city level and state level, where they are in the range of $\gamma > 1$. These results show that an appropriate value of regularization parameter γ that is neither too small nor too large can again optimize the model performance.

2) Parameter Sensitivity of MREF-II: Fig. 6(a) illustrates the performance for civil unrest event forecasting of MREF-II versus γ. Similar to the experiment on MREF-I, the performance in terms of F-measures for civil unrest event forecasting are illustrated for different spatial resolutions when γ varies from 0.01 to 10. Starting from $\gamma = 0.01$, the F-measures at different spatial resolutions increase slightly as γ increases across the range of $\gamma = 0.5{\sim}2$, decreasing at the city level and state level across the range of $\gamma > 6$. These results show that the value of regularization parameter $\gamma$and the regularization terms influence the overall performance, and it should therefore be neither too small or too large, which is similar to the situation for MREF-I.

Fig. 6(b) shows the performance of MREF-II versus γ for the influenza outbreaks data set. For all spatial resolutions, the F-measures at $\gamma = 0.01{\sim}0.8$ are essentially higher than those at $\gamma > 0.8$, particularly for the city level and the state level; the country level is more consistent across different values of γ. The F-measures is stable in the range of $\gamma = 0.01{\sim}0.1$ across different spatial resolutions but slightly increases when



(a) Sensitivity analysis on civil unrest dataset

(b) Sensitivity analysis on Flu dataset

Fig. 6. ***Sensitivity analyses on regularization parameter γ.***

$\gamma = 0.1{\sim}0.8$, where the highest value is achieved, after which it drops to smaller values at the city level and state level. These results again show that an appropriate value of γ that is neither too small nor too large yields the best model performance, especially at the city and state levels.

VII. CONCLUSION

For spatial event forecasting, the accuracy and discernibility of the predictive model are two the key concerns. Their joint consideration and optimization present several challenges, however. In this paper, we propose a new multiresolution spatial event forecasting framework to address all the challenges simultaneously. To achieve this, we propose two novel multitask learning models that leverage the heterogeneous relationships among the prediction tasks, and develop effective parameter optimization algorithms based on ADMM and dynamic programming. Experiments on several data sets in two different domains were conducted to evaluate the performance and parameter sensitivity of the proposed models. The results demonstrate that because of the effective utilization of the shared information across different spatial resolutions and neighborhoods, the proposed model outperforms existing methods used for comparison. ■

REFERENCES

[1] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan, “Multi-task learning for spatio-temporal event forecasting,” in *Proc. KDD*, 2015, pp. 1503–1512.

[2] N. Ramakrishnan *et al.*, “‘Beating the news’ with EMBERS: Forecasting civil unrest using open source indicators,” in *Proc. KDD*, 2014, pp. 1799–1808.

[3] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, “Predicting elections with twitter: What 140 characters reveal about political sentiment,” in *Proc. ICWSM*, vol. 10. 2010, pp. 178–185.

[4] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, Mar. 2011.

[5] M. Arias, A. Arratia, and R. Xuriguera, “Forecasting with twitter data,” *Trans. Intell. Syst. Technol.*, vol. 5, no. 1, p. 8, 2013.

[6] 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*. New York, NY, USA: Springer-Verlag, 2012, pp. 231–238.

[7] S. Aghababaei and M. Makrehchi, “Mining social media content for crime prediction,” in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell. (WI)*, Oct. 2016, pp. 526–531.

[8] H.-W. Kang and H.-B. Kang, “Prediction of crime occurrence from multi-modal data using deep learning,” *PloS One*, vol. 12, no. 4, p. e0176244, 2017.

[9] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, “Predicting flu trends using twitter data,” in *Proc. IEEE Conf. Comput. Commun. Workshops*, Apr. 2011, pp. 702–707.

[10] Y. Ning, S. Muthiah, H. Rangwala, and N. Ramakrishnan, “Modeling precursors for event forecasting via nested multi-instance learning,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1095–1104.

[11] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, and N. Ramakrishnan, “Combining heterogeneous data sources for civil unrest forecasting,” in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Jul. 2015, pp. 258–265.

[12] M. J. Fard, P. Wang, S. Chawla, and C. K. Reddy, “A Bayesian perspective on early stage event prediction in longitudinal data,” *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 12, pp. 3126–3139, Dec. 2016.

[13] K. Radinsky and E. Horvitz, “Mining the Web to predict future events,” in *Proc. WSDM*, 2013, pp. 255–264.

[14] L. Hu, J. Li, L. Nie, X.-L. Li, and C. Shao, “What happens next? Future subevent prediction using contextual hierarchical LSTM,” in *Proc. AAAI*, 2017, pp. 3450–3456.

[15] F. Qiao, P. Li, X. Zhang, Z. Ding, J. Cheng, and H. Wang, “Predicting social unrest events with hidden Markov models using GDELT,” *Discrete Dyn. Nature Soc.*, vol. 2017, May 2017, Art. no. 8180272.

[16] M. S. Gerber, “Predicting crime using Twitter and kernel density estimation,” *Decision Support Syst.*, vol. 61, pp. 115–125, May 2014.

[17] T. Rekatsinas *et al.*, “SourceSeer: Forecasting rare disease outbreaks using multiple data sources,” in *Proc. SIAM Int. Conf. Data Mining*, 2015, pp. 379–387.

[18] L. Zhao, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan, “Hierarchical incomplete multi-source feature learning for spatiotemporal event forecasting,” in *Proc. KDD*, 2016, pp. 2085–2094.

[19] L. Zhao, J. Chen, F. Chen, W. Wang, C.-T. Lu, and N. Ramakrishnan, “Simnest: Social media nested epidemic simulation via online semi-supervised deep learning,” in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2015, pp. 639–648.

[20] Q. Zhang, N. Perra, D. Perrotta, M. Tizzoni, D. Paolotti, and A. Vespignani, “Forecasting seasonal influenza fusing digital indicators and a mechanistic disease model,” in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 311–319.

[21] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan, “Spatiotemporal event forecasting in social media,” in *Proc. SDM*, 2015, pp. 963–971.

[22] D. Wang and W. Ding, “A hierarchical pattern learning framework for forecasting extreme weather events,” in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Jun. 2015, pp. 1021–1026.

[23] N. Alsaedi, P. Burnap, and O. Rana, “Can we predict a riot? Disruptive event detection using twitter,” *ACM Trans. Internet Technol.*, vol. 17, no. 2, p. 18, 2017.

[24] M. Hasan, M. A. Orgun, and R. Schwitter, “A survey on real-time event detection from the twitter data stream,” *J. Inf. Sci.*, p. 0165551517698564, Mar. 2017.

[25] E. Schubert, M. Weiler, and H.-P. Kriegel, “Signitrend: Scalable detection of emerging topics in textual streams by hashed significance thresholds,” in *Proc. KDD*, 2014, pp. 871–880.

[26] M. Adedoyin-Olowe, M. M. Gaber, C. M. Dancausa, F. Stahl, and J. B. Gomes, “A rule dynamics approach to event detection in twitter with its application to sports and politics,” *Expert Syst. Appl.*, vol. 55, pp. 351–360, Aug. 2016.

[27] X. Zhang, X. Chen, Y. Chen, S. Wang, Z. Li, and J. Xia, “Event detection and popularity prediction in microblogging,” *Neurocomputing*, vol. 149, pp. 1469–1480, Feb. 2015.

[28] H. Cai, Z. Huang, D. Srivastava, and Q. Zhang, “Indexing evolving events from tweet streams,” *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 11, pp. 3001–3015, Nov. 2015.

[29] W. Yu, J. Li, M. Z. A. Bhuiyan, R. Zhang, and J. Huai, “Ring: Real-time emerging anomaly monitoring system over text streams,” *IEEE Trans. Big Data*, 2017, DOI: 10.1109/TBDATA.2017.2672672.

[30] M. Shao, J. Li, F. Chen, H. Huang, S. Zhang, and X. Chen, “An efficient approach to event detection and forecasting in dynamic multivariate social media networks,” in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 1631–1639.

[31] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake shakes twitter users: Real-time event detection by social sensors,” in *Proc. WWW*, 2010, pp. 851–860.

[32] S. Unankard, X. Li, and M. A. Sharaf, “Emerging event detection in social networks with location sensitivity,” *World Wide Web*, vol. 18, no. 5, pp. 1393–1417, 2015.

[33] J. Foley, M. Bendersky, and V. Josifovski, “Learning to extract local events from the Web,” in *Proc. 38th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2015, pp. 423–432.

[34] X. Dong, D. Mavroeidis, F. Calabrese, and P. Frossard, “Multiscale event detection in social media,” *Data Mining Knowl. Discovery*, vol. 29, no. 5, pp. 1374–1405, 2015.

[35] J. Krumm and E. Horvitz, “Eyewitness: Identifying local events via space-time signals in twitter feeds,” in *Proc. 23rd SIGSPATIAL Int. Conf. Adv. Geograph. Inf. Syst.*, 2015, p. 20.

[36] C. Zhang *et al.*, “Geoburst: Real-time local event detection in geo-tagged tweet streams,” in *Proc. 39th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2016, pp. 513–522.

[37] Q. Li, A. Nourbakhsh, S. Shah, and X. Liu, “Real-time novel event detection from social media,” in *Proc. IEEE 33rd Int. Conf. Data Eng. (ICDE)*, Apr. 2017, pp. 1129–1139.

[38] W. Feng *et al.*, “STREAMCUBE: Hierarchical spatio-temporal hashtag clustering for event exploration over the

twitter stream," in *Proc. IEEE 31st Int. Conf. Data Eng. (ICDE)*, Apr. 2015, pp. 1561–1572.

[39] H. Cai, Y. Yang, X. Li, and Z. Huang, "What are popular: Exploring twitter features for event detection, tracking and visualization," in *Proc. 23rd ACM Int. Conf. Multimedia*, 2015, pp. 89–98.

[40] M. Granroth-Wilding and S. Clark, "What happens next? Event prediction using a compositional neural network model," in *Proc. 30th AAAI Conf. Artif. Intell.*, 2016.

[41] S. Muthiah *et al.*, "Planned protest modeling in news and social media," in *Proc. IAAI*, 2015, pp. 3920–3927.

[42] B. Ahmed *et al.*, "Multi-task learning with weak class labels: Leveraging iEEG to detect cortical lesions in cryptogenic epilepsy," in *Proc. Mach. Learn. Healthcare Conf.*, 2016, pp. 115–133.

[43] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," in *Proc. ICML*, 2010, pp. 543–550.

[44] N. J. Yadwadkar, B. Hariharan, J. E. Gonzalez, and R. Katz, "Multi-task learning for straggler avoiding predictive job scheduling," *J. Mach. Lear. Res.*, vol. 17, no. 106, pp. 1–37, 2016.

[45] A. Acharya, R. J. Mooney, and J. Ghosh, "Active multitask learning using supervised and shared latent topics," in *Proc. Pattern Recognit. Big Data*, 2016, p. 75.

[46] G. Cheng, P. Zhou, and J. Han, "Learning rotation-invariant convolutional neural networks for object detection in VHR optical remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7405–7415, Dec. 2016.

[47] K. Lin, J. Xu, I. M. Baytas, S. Ji, and J. Zhou, "Multi-task feature interaction learning," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1735–1744.

[48] Y. Jiang, C.-S. Perng, and T. Li, "Meta: Multi-resolution framework for event summarization," in *Proc. SDM*, 2014, pp. 605–613.

[49] D. Agarwal, R. Agrawal, R. Khanna, and N. Kota, "Estimating rates of rare events with multiple hierarchies through scalable log-linear models," in *Proc. KDD*, 2010, pp. 213–222.

[50] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.

[51] G. Cheng, J. Han, L. Guo, Z. Liu, S. Bu, and J. Ren, "Effective and efficient midlevel visual elements-oriented land-use classification using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4238–4249, Aug. 2015.

[52] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.

[53] V. Sekar, Y. Xie, M. K. Reiter, and H. Zhang, "A multi-resolution approach for worm detection and containment," in *Proc. Int. Conf. Dependable Syst. Netw. (DSN)*, 2006, pp. 189–198.

[54] H. Moon, S. Yi, G. S. Choi, Y.-S. Jeon, and J. Kim, "A multi-resolution port scan detection technique for high-speed networks," *J. Inf. Sci. Eng.*, vol. 31, no. 5, pp. 1613–1632, 2015.

[55] B. Doucoure, K. Agbossou, and A. Cardenas, "Time series prediction using artificial wavelet neural network and multi-resolution analysis: Application to wind speed data," *Renew. Energy*, vol. 92, pp. 202–211, Jul. 2016.

[56] M. Cooper, J. Foote, A. Girgensohn, and L. Wilcox, "Temporal event clustering for digital photo collections," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 1, no. 3, pp. 269–288, 2005.

[57] D. Jiang, C. Yao, Z. Xu, and W. Qin, "Multi-scale anomaly detection for high-speed network traffic," *Trans. Emerg. Telecommun. Technol.*, vol. 26, no. 3, pp. 308–317, 2015.

[58] J. Zhou, J. Chen, and J. Ye, *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Tempe, AZ, USA: Arizona State Univ., 2011.

[59] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Found. Trends Mach. Learn.*, vol. 3, no. 1, pp. 1–122, Jan. 2011.

[60] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Optimization with sparsity-inducing penalties," *Found. Trends Mach. Learn.*, vol. 4, no. 1, pp. 1–106, Jan. 2012.

[61] M. J. Paul and M. Dredze, "A model for mining public health topics from twitter," *Health*, vol. 11, pp. 6–16, 2012.

[62] R. Compton *et al.*, "Using publicly visible social media to build detailed forecasts of civil unrest," *Secur. Inform.*, vol. 3, no. 1, pp. 1–10, 2014.

ABOUT THE AUTHORS

Liang Zhao received the Ph.D. degree from Virginia Tech, Blacksburg, VA, USA.

He is an Assistant Professor at the Information Science and Technology Department, George Mason University, Fairfax, VA, USA. His research interests include natural language processing, text mining, machine learning, and robotics. In recent years, he has worked primarily on applications to social media, civil unrests, and public health informatics.

Junxiang Wang received the B.S. degree from East China Normal University, Shanghai, China. in 2012. Currently, he is working toward the Ph.D. degree at the Information Science and Technology Department, George Mason University, Fairfax, VA, USA, supervised by Prof. L. Zhao. His research focuses on data mining in social media and convex optimization.

Feng Chen received the B.S. degree from Hunan University, Changsha, China, in 2001, the M.S. degree from Beihang University, Beijing, China, in 2004, and the Ph.D. degree from Virginia Tech, Blacksburg, VA, USA, in 2012, all in computer science.

He is an Assistant Professor with the University at Albany, SUNY, Albany, NY, USA. His research focuses on the detection of emerging events and other relevant patterns in the mobile context and/or data mining of spatial temporal, textual, or social media data.

Chang-Tien Lu received the M.S. degree in computer science from the Georgia Institute of Technology, Atlanta, GA, USA, in 1996 and the Ph.D. degree in computer science from the University of Minnesota, Minneapolis, MN, USA, in 2001.

He is an Associate Professor with the Department of Computer Science, Virginia Tech, Blacksburg, VA, USA. His research interests include spatial databases, data mining, geographic information systems, and intelligent transportation systems.

Prof. Lu served as the Vice Chair of ACM SIGSPATIAL from 2011 to 2014. He is an ACM Distinguished Scientist.

Naren Ramakrishnan received the Ph.D. degree in computer sciences from Purdue University, West Lafayette, IN, USA, in 1997.

He is currently the Thomas L. Phillips Professor of Engineering in the Department of Computer Science, Virginia Tech, Arlington, VA, USA. His research has been supported by NSF, DHS, NIH, NEH, DARPA, IARPA, ONR, General Motors, HP Labs, NEC Labs, and Advance Auto Parts.

Prof. Ramakrishnan has served as both Program Chair and General Chair of the IEEE International Conference on Data Mining.