Multi-resolution Spatial Event Forecasting in Social Media

Liang Zhao
George Mason University
lzhao9@gmu.edu

Feng Chen
University of Albany, SUNY
fchen5@albany.edu

Chang-Tien Lu, Naren Ramakrishnan
Virginia Tech
ctlu@vt.edu, naren@cs.vt.edu

***Abstract*—Social media has become a significant surrogate for spatial event forecasting. The accuracy and discernibility of a spatial event forecasting model are two key concerns, which respectively determine how accurate and how detailed the model's predictions could be. Existing work pays most attention on the accuracy alone, seldom considering the accuracy and discernibility simultaneously, because this would requires a considerably more sophisticated model while still suffering from several challenges: 1) the precise formulation of the trade-off 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 model 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 model based on multitask learning, whose parameters are optimized by our proposed algorithm based on the Alternative Direction Method of Multipliers (ADMM). Extensive experimental evaluations on 11 datasets from different domains demonstrated the effectiveness of our proposed approach.**

1. Introduction

Social media like Twitter and Weibo have become popular platforms, serving as real-time "sensors" for social trends and incidents [26]. 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 a potential precursor 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" [18]. A great deal of recent research has widely explored and demonstrated the power of social media for spatial event forecasting for topics such as crimes [23], civil unrest [22], and disease outbreaks [1].

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 trade-off 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 country, say Mexico, which has 31 states and over 2000 cities, and we want 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. Moreover, 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 possess country-level observations or train a city-level prediction model effectively if we only have state-level labels. Social media is composed of such noisy data that it provides social sensors with different geographical discernibilities. For example, geo-tagged tweets provide pinpoint geographical coordinates if their users enable this function on their mobile device, but this is not a common situation; other users may provide their city information while some only provide information on their state, country, or nothing at all, leaving their postings with different spatial resolutions of city, state, country, or the planet earth, respectively.

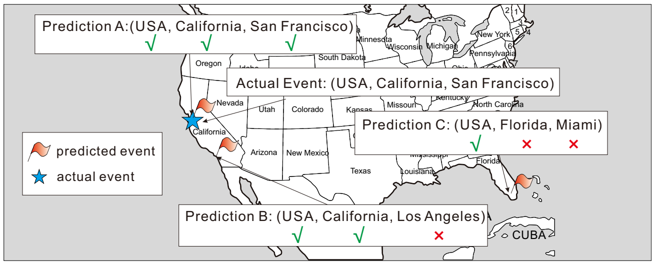


Figure 1: Spatial event forecasting performance. The qualities of these 3 predictions are different because they achieve correct prediction in different discernibilities

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 [12], [18], [24]. For those seeking to address this issue, the major challenges can be summarized as follows. **1) Trade-off between accuracy and discernibility**. Traditionally, we focus on evaluating whether or not a prediction is correct rather than "how correct" it actually needs to be for practical purposes. For example, Figure 1 shows three event predictions, A, B, and C, for a future date "August 13, 2016". 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 contains geo-information 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, only around 3% of all the data possesses spatial coordinates that include street-level geo-information but 30%-50% contains city-level or state-level information. In this case, only 3% of all the data would be used, hence the model performance is limited by insufficient data. **3) Interaction and heterogeneity of geographical locations.** Nearby locations could have regional correlation, such as being influenced by the same regional epidemics, natural disasters, and social events. In the meanwhile, 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 be a strong signal of an upcoming civil unrest event in a city with a population of a few million users but could be a strong indicator 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 above-mentioned challenges, we propose a novel model, **Multi-Resolution Spatial Event Forecasting (MREF)**, based on multi-task learning that jointly reinforces the accuracy and discernibility of event forecasting. In our MREF, each task is 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 are 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 contains merely coarse-grained spatial information. In addition, to characterize the geographical neighborhood relationship among tasks, a tree-structure geographical hierarchy is developed. The major contributions of this paper are as follows:

1) **Formulating a framework for multi-resolution spatial event forecasting.** Here, multi-resolution spatial event forecasting is formulated as a multi-task 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 multi-task model with heterogeneous task relationships.** In the proposed multi-task 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 multi-task model is a non-smooth inequality-constrained overlapping group lasso problem which is challenging to solve. By introducing auxiliary variables, we develop an effective ADMM-based algorithm to ensure the global optimal solution for this problem.

The rest of this paper is organized as follows. Section 2 reviews existing work. Section 3 introduces the problem setup. Section 4 elaborates our MREF model and its parameter optimization algorithm. In Section 5, extensive experiments to evaluate the performance of MREF are conducted and analyzed; the work is summarized and conclusions drawn in Section 6.

2. Related Work

The related work of this paper is summarized by categories in the following.

Event detection: There is a large amount of work on the identification of ongoing events, including disease outbreaks [21], earthquakes [19] and various other types of events [14]. Generally, for event detection, either classification or clustering is utilized to extract tweets of interest and then the spatial [19], temporal [20], or spatiotemporal burstiness [10] of the extracted tweets is examined to identify the potential occurrence of ongoing events. Utilizing retrospective analysis on tweets, Dong et al. proposed a wavelet-based clustering method to extract the historical events with different time durations and spatial sizes [10]. However, instead of forecasting events in the future, these approaches typically uncover them only after they have occurred.

Event forecasting: Currently, most research in this area focuses solely on temporal events, although some of the models developed are also able to handle spatial information. The research on temporal events includes the forecasting of elections [22], stock market movements [7], disease outbreaks [25], box office ticket sales [5], and crimes [23]. These studies can be classified into three categorizes: 1) Linear regression models [7]; 2) Nonlinear models [23]; and 3) Time series-based methods [1]. However, few existing approaches are able to characterize the information in a spatial dimension in order to forecast spatial events. Gerber utilized a logistic regression model for spatiotemporal events forecasting [12]. Zhao et al. [27] developed a multi-level model to characterize the hierarchical features from multiple data sources and predict spatio-temporal social events. Ramakrishnan et al. [18] built separate LASSO models for different locations to predict their occurrence. Zhao et al. [24] proposed a multi-task learning framework for event forecasting that jointly learns multiple related spatial locations. However, existing methods typically only consider events using a single geographical granularity and do not jointly optimize the discernibility and accuracy.

Multi-task learning: In multi-task learning (MTL), multiple related tasks are learned simultaneously to improve generalization performance [26]. Many MTL approaches have been proposed in the past [28]. Evgeniou et al. proposed a regularized MTL that constrained the models of all the tasks to be close to each other [11]. This task relatedness can also be characterized by constraining multiple tasks to share a common underlying structure, such as a common set of features [4], a common subspace [3], or using a tree-structured model [15]. For example, Kim et al. [15] proposed a multi-task learning model which leverges a tree-structured relationship among the tasks. MTL approaches have been applied in many domains, including computer vision and biomedical informatics. To the best of our knowledge, however, ours is the first work that applies MTL for multi-resolution spatial civil unrest forecasting.

Multi-resolution sensing: Multi-resolution sensing approaches have been typically applied in domains such as computer vision and satellite remote sensing [13]. To analyze the rates of advertisements' responses in websites, Agarwal et al. [2] developed a method that can estimate predictions for fine-grained geo-locations. Aiming at a retrospective analysis of historical events, Jiang et al. [13] designed a framework to extract and summarize events from different views with different resolutions. Currently, few researchers are utilizing multi-resolution in spatial event forecasting. To our knowledge, we are the first to apply multiple geographical resolutions for civil unrest forecasting.

3. 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 sub-collection of tweets at tth time interval and T is the set of time intervals. According to the granularity of geo-information, 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 whole spatial area of location q_i is included within location s_j, we say q_i is a **spatial subregion** of s_j, denoted as $q_i \sqsubseteq s_j$ or equally $s_j \sqsupseteq q_i$ $(i > j)$.*

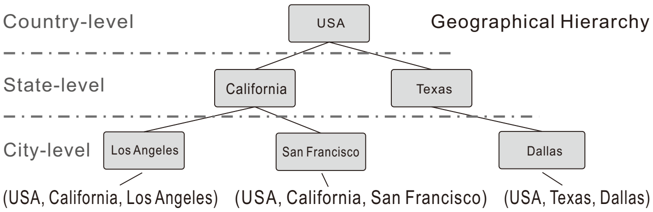


Figure 2: The location tuples based on geographical hierarchy

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

A tweet sub-collection 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, $n = 1, \cdots, N$. $X_{t,s_n} \in \mathbb{N}^{K \times 1}$ 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 $S = \{S_n\}_n^N$ as the set of all the locations. Because not all of the tweets possess location information under the finest spatial resolution, we know that $\{X_{t,s_n}\}_{s_n}^{S_n} \subseteq \{X_{t,s_{n-1}}\}_{s_{n-1}}^{S_{n-1}}$, $n = 2, \cdots, N$. 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 $Y_{\tau,s} = (Y_{\tau,s_1}, \cdots, Y_{\tau,s_N})$. The problem of this paper can thus be formulated as follows:

Problem Formulation: Given the tweets data X_t in N different spatial resolutions, the goal is to predict the occurrence of a future event for location $s = (s_1, \cdots, s_N)$ at time interval τ, where s_n $(n = 1, \cdots, N)$ 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 a mapping from tweets data to future event predictions $f : X_{t,s} \to \{Y_{\tau,s_1}, \cdots, Y_{\tau,s_N}\}$ for locations s at N spatial resolutions.

Definition 3 (Multi-resolution Event Forecasting Error) *The multi-resolution 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:*

$$\mathcal{L}(W) = \sum\nolimits_n^N \sum\nolimits_{s_n}^{S_n} \mathcal{L}(W_{s_n})$$

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 $f(X_{t,s_n}, W_{s_n})$ against the labels Y_{τ,s_n} for all the time intervals T. $\mathcal{L}(W_{s_n})$ can be a logistic loss [28] where $f(X_{t,s_n}, W_{s_n}) = 1/(1 + e^{-W_{s_n} \cdot X_{t,s_n}})$.

Due to the different characteristics of different locations and in different spatial resolutions, it is unfeasible to build

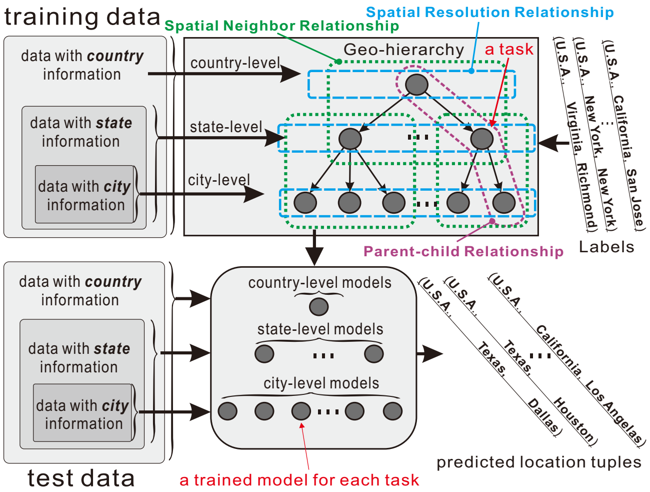

training data
Spatial Neighbor Relationship
Spatial Resolution Relationship
data with country information
data with state information
data with city information
country-level
state-level
city-level
Geo-hierarchy
a task
(U.S.A., Virginia, Richmond)
(U.S.A., New York, New York)
(U.S.A., California, San Jose)
Parent-child Relationship
Labels
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

Figure 3: A schematic view of the proposed model

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*. 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) *Multi-resolution event forecasting paradox*. Contradictory predictions at different spatial resolutions could also happen. For example, a model that predicts there will be an event in "Los Angelas" could also predict that there will be no event in "California". To address these three challenges, in the next section, we propose a novel multi-task learning model named MREF based on mixed-structured task relatedness and a non-smooth constraint.

4. Multi-Resolution Spatial Event Forecasting

In this section we propose a new model, named Multi-Resolution Spatial Event Forecasting (MREF), based on multi-task feature learning. In Section 4.1, the multiple types of task relatedness are characterized mathematically, and these are then integrated into the new multi-task feature learning framework in Section 4.2. In Section 4.3, an effective algorithm based on ADMM is proposed that ensures the global optima.

4.1. 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. Figure 3 illustrates the proposed multi-task learning framework that characterizes all three major aspects of relatedness among all the locations (tasks) for the problem of multi-resolution spatial event forecasting: 1) Spatial neighborhood relationships, 2) Spatial resolution relationships, and 3) Parent-child relationships, which are elaborated in turn below.

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 top right of Figure 3; the location (task) nodes in a sub-tree are within a spatial neighborhood.

As illustrated in top right of Figure 3, a geographical hierarchy is a tree whose nodes consist of all the spatial locations and the links are the parent-child relationships among them. In this tree of geographical hierarchy, denote $\mathcal{T} = \{\mathcal{T}_i\}_i$ as the set of sub-trees that are defined as $\mathcal{T}_i = \{s_n\} \cup \{s'_{n+1} | s'_{n+1} \sqsubseteq s_n, n < N\}$, which means a sub-tree 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. Denote $Q^{(n)}_{\cdot,k}$ as the model parameters for feature k for all the locations in 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 in the following lemma and theorem.

Lemma 1 *If there is no event in a location, then there is no event in all of its subregions. Formally, without loss of gen-*

erality, 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 1 *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.

4.2. Objective Function

The above consideration of the heterogeneous relatedness of forecasting tasks leads to a new multi-task feature learning framework by applying a general paradigm of multi-task learning, namely to minimizing the penalized empirical loss:

$$\min_{W} \mathcal{L}(W) + \Omega(W) \tag{3}$$

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, a constraint is also added based on Theorem 1. In all, the objective function for our multi-task feature learning model is as follows:

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

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_1 and γ_2 are regularization parameters such that $\gamma_{\text{P}} = \gamma / \sum_{\mathcal{T}_i}^{\mathcal{T}} \sqrt{|\mathcal{T}_i|}$ and $\gamma_{\text{Q}} = \gamma / \sum_n^N \sqrt{|S_n|}$ where γ is the regularization parameter that balances the trade off between the loss function $\mathcal{L}(W)$ and the regularization terms.

The objective function in Equation (4) encompasses the joint consideration of the heterogeneous task relationships. However, to solve this objective function two challenges must first be overcome: 1) *non-smooth inequality constraint*, and 2) *overlapping among the coupled sub-trees*, which are discussed and addressed in the following.

1. **Non-smooth inequality constraint.**

The non-smooth function $max(\cdot)$ in the inequality constraint in Equation 4 makes the objective function difficult to solve. To address this challenge, we propose to replace this term with an alternative constraint that applies a sufficiency condition to the original constraint:

$$f(X_{s_n,t}, W_{s_n}) \geq f(X_{s'_{n+1},t}, W_{s'_{n+1}}),\ s'_{n+1} \sqsubseteq s_n \tag{5}$$

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

2. **Overlapping among the coupled sub-trees.**

It can be seen from Figure 3 that a node in the geographical hierarchy tree could belong to two sub-trees. For example, state-level nodes belong to a sub-tree whose root is a country-level node, but they can also be the root of another sub-tree 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 an efficient optimization solution by introducing two auxiliary variables, U and V. U is the model parameter set for the set of sub-trees $\mathcal{T}_{\mathcal{O}}$ with roots in odd-number (i.e., $n = 1, 3, 5, \cdots$) spatial resolutions, while V represents the set of sub-trees $\mathcal{T}_{\mathcal{E}}$ with roots in even-number (i.e., $n = 2, 4, 6, \cdots$) levels. Thus, neither U nor V contain overlapping sub-trees. 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:

$$\begin{aligned} \min_{W} \mathcal{L}(W) + \gamma_0 \sum\nolimits_k^K \sum\nolimits_{\mathcal{T}_i}^{\mathcal{T}_{\mathcal{O}}} \|U_{\mathcal{T}_i,k}\|_F + \gamma_1 \sum\nolimits_k^K \sum\nolimits_{\mathcal{T}_i}^{\mathcal{T}_{\mathcal{E}}} \|V_{\mathcal{T}_i,k}\|_F \\ + \gamma_2 \sum\nolimits_k^K \sum\nolimits_n^N \|Q^{(n)}_{\cdot,k}\|_F \qquad\qquad \\ s.t. \qquad W = P + Q,\ P = U,\ P = V, \qquad \\ g(X,W) + \beta = 0, \beta = \beta_+,\ \beta_+ \geq 0 \qquad \end{aligned} \tag{6}$$

where $g(X,W) = \{g(X_{s_n,t}, W_{s_n})\}_{s_n,t}^{S',T}$ is a matrix of which each element $g(X_{s_n,t}, W_{s_n}) = f(X_{s'_{n+1},t}, W_{s'_{n+1}}) - f(X_{s_n,t}, W_{s_n})$ 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 $\gamma_2 = \gamma / \sum_n^N \sqrt{|S_n|}$ where γ is the regularization parameter that balances the trade off between the loss function $\mathcal{L}(W)$ and the regularization terms.

4.3. Parameter Optimization of MREF

The objective function in Equation (6) is convex because the loss function, regularization terms, and constraints are all convex. To solve the convex optimization problem with constraints, the alternating direction method of multipliers (ADMM) has begun to be widely utilized as an efficient algorithm that first breaks the original large problem into smaller subproblems that can be solved easily and fast. Here we propose a ADMM-based framework that solves Equation

(6) by first obtaining its augmented Lagrangian format as follows:

$$\begin{aligned}\min_{\Theta} \mathcal{L}(W) &+ \gamma_0 \sum\nolimits_{k,\mathcal{T}_i}^{K,\mathcal{T}_\mathcal{O}} \|U_{\mathcal{T}_i,k}\|_F + \gamma_1 \sum\nolimits_{k,\mathcal{T}_i}^{K,\mathcal{T}_\mathcal{E}} \|V_{\mathcal{T}_i,k}\|_F \\ &+ \gamma_2 \sum\nolimits_k \sum\nolimits_n^N \|Q_{\cdot,k}^{(n)}\|_F + \langle \alpha_1, W-P-Q \rangle \\ &+ \frac{\rho}{2}\|W-P-Q\|_F^2 + \langle \alpha_2, P-U \rangle + \frac{\rho}{2}\|P-U\|_F^2 \\ &+ \langle \alpha_3, P-V \rangle + \frac{\rho}{2}\|P-V\|_F^2 + \langle \alpha_4, g(W)+\beta \rangle \\ &+ \frac{\rho}{2}\|g(W)+\beta\|_F^2 + \langle \alpha_5, \beta-\beta_+ \rangle + \frac{\rho}{2}\|\beta-\beta_+\|_F^2 \end{aligned} \tag{7}$$

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 mulipliers 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, called mixed-structured multi-task learning, as shown in Algorithm 1. It alternately optimizes each of the parameters in Θ until convergence is achieved. Lines 4-5 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. [8]. The detailed optimization steps are described in more detail below.

Algorithm 1 Mixed-structured Multi-task Learning

```
    Input: $X$, $Y$, $\gamma$
    Output: solution $W$
 1: Initialize $\rho = 1$, $W, U, V, P, Q, \alpha_i, \beta, \beta^+ = \mathbf{0}, i = 1, \cdots, 5$.
 2: Choose $\varepsilon_p > 0, \varepsilon_d > 0$.
 3: repeat
 4:     Update $W, U, V, P, Q$ by Equations (8), (9), and (10).
 5:     Update $\{\alpha_i\}_{i=1}^5, \beta, \beta^+$ by Equations (11)and (13).
 6:     Update primal and dual residuals $p$ and $d$.
 7:     if $r > 10s$ then
 8:         $\rho \leftarrow 2\rho$                    # Update penalty parameter
 9:     else if $10r < s$ then
10:         $\rho \leftarrow \rho/2$                   # Update penalty parameter
11:     else
12:         $\rho \leftarrow \rho$                     # Update penalty parameter
13:     end if
14: until $p < \varepsilon^p$ and $d < \varepsilon^d$     # Convergence criterion
```

1. Update W, fix others.

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

$$\begin{aligned} W \leftarrow \operatorname*{argmin}_W \ & \mathcal{L}(W) + \langle \alpha_2, g(W)+\beta \rangle + \frac{\rho}{2}\|g(W)+\beta\|_F^2 \\ &+ \langle \alpha_1, W-P-Q \rangle + \frac{\rho}{2}\|W-P-Q\|_F^2 \end{aligned} \tag{8}$$

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:

$$\begin{aligned} P \leftarrow \operatorname*{argmin}_P \ & \langle \alpha_1, W-P-Q \rangle + \langle \alpha_2, P-U \rangle + \frac{\rho}{2}\|P-U\|_F^2 \\ &+ \frac{\rho}{2}\|W-P-Q\|_F^2 + \langle \alpha_3, P-V \rangle + \frac{\rho}{2}\|P-V\|_F^2 \end{aligned} \tag{9}$$

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

3. Update U, V, Q, fix others.

The optimization of U, V, and Q are all problems of least squares loss functions with $\ell_{2,1}$ norms:

$$\begin{aligned} U &\leftarrow \operatorname*{argmin}_U \gamma_0 \sum\nolimits_{k,\mathcal{T}_i}^{K,\mathcal{T}_\mathcal{O}} \|U_{\mathcal{T}_i,k}\|_F + \langle \alpha_2, P-U \rangle + \frac{\rho}{2}\|P-U\|_F^2 \\ V &\leftarrow \operatorname*{argmin}_V \gamma_0 \sum\nolimits_{k,\mathcal{T}_i}^{K,\mathcal{T}_\mathcal{E}} \|V_{\mathcal{T}_i,k}\|_F + \langle \alpha_2, P-V \rangle + \frac{\rho}{2}\|P-V\|_F^2 \\ Q &\leftarrow \operatorname*{argmin}_Q \gamma_2 \sum\nolimits_k \sum\nolimits_n^N \|Q_{\cdot,k}^{(n)}\|_F + \langle \alpha_1, W-P-Q \rangle \\ &\qquad + \frac{\rho}{2}\|W-P-Q\|_F^2 \end{aligned} \tag{10}$$

where all 3 problems can be efficiently solved by using proximal operators [6].

4. Update β, fix others.

The optimization of β can be formulated as the following least squares problem:

$$\begin{aligned} \beta \leftarrow \operatorname*{argmin}_\beta \ & \langle \alpha_4, g(W)+\beta \rangle + \frac{\rho}{2}\|g(W)+\beta\|_F^2 \\ &+ \langle \alpha_5, \beta-\beta_+ \rangle + \frac{\rho}{2}\|\beta-\beta_+\|_F^2 \end{aligned} \tag{11}$$

where the solution is: $\beta = \frac{1}{2}(\beta_+ - g(W)) - \frac{1}{2\rho}(\alpha_4 + \alpha_5)$.

5. Update β_+, fix others.

The optimization of β_+ can be formulated as a least squares problem with linear inequality constraint:

$$\beta_+ \leftarrow \operatorname*{argmin}_{\beta_+ \geq 0} \langle \alpha_5, \beta-\beta_+ \rangle + \frac{\rho}{2}\|\beta-\beta_+\|_F^2 \tag{12}$$

To eliminate inequality constraint, first let $c^2 = \beta_+, c \in \mathbb{R}$ and we get 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 $\alpha_i (i = 1, \cdots, 5)$

The updating of the dual variables α_i is as follows:

$$\begin{aligned} &\alpha_1 \leftarrow \alpha_1 + \rho \cdot (W-P-Q) \\ &\alpha_2 \leftarrow \alpha_2 + \rho \cdot (P-U), \ \alpha_3 \leftarrow \alpha_3 + \rho \cdot (P-V) \\ &\alpha_4 \leftarrow \alpha_4 + \rho \cdot (F+\beta), \ \alpha_5 \leftarrow \alpha_5 + \rho \cdot (\beta-\beta_+) \end{aligned} \tag{13}$$

5. Experiments

In this section, the proposed model MREF is evaluated on 11 real-world datasets from two different domains. After the experiment setup has been introduced in Section 5.1, the effectiveness of the methods is evaluated against several existing methods on different spatial resolutions, along with an analysis of the performances on precision-recall curves for all the comparison methods, in Section 5.2.

Table 1: Datasets and Labels

Dataset	#Tweets	Label sources [1]	#Events
Argentina	160,564,890	Clarín; La Nación; Infobae	1427
Brazil	185,286,958	O Globo; O Estado de São Paulo; Jornal do Brasil	3417
Chile	97,781,414	La Tercera; Las Últimas Notícias; El Mercurio	776
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

5.1. Experimental Setup

5.1.1. Datasets and Labels. The experimental evaluations in this study are based on 11 datasets on different domains. Of these, 10 datasets 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 datasets, Table 2 shows the specific country from which the Twitter data was gathered for each dataset. The raw Twitter data is collected from 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), exclusively provided by MITRE [16]. GSR is a collection of civil unrest news reports from the most influential newspaper outlets in Latin America [18], as shown in Table 1. For civil unrest forecasting, 3 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").

For the dataset 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 [17]; the time period of this dataset is also shown in Table 2. The data collection for the influenza dataset is 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). CDC typically organizes the influenza surveillance data by HHS regions[2], which groups the US's states into 10 regions. CDC publishes weekly influenza-like illness (ILI) activity level within each state in the United States using the proportion of outpatient visits to healthcare providers for ILI. There are 4 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, 3 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").

5.1.2. Parameter Settings and Metrics. There is one tunable parameter in our MREF model, namely the regularization parameter γ. This parameter was set for all 10 datasets based on 10-fold cross validation on the training set.

In the experiment, the event forecasting task is to predict whether or not there will be an event in the next time-step for a specific location at several different spatial resolutions. For civil unrest datasets, 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 2 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 $= 2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precsion+Recall})$.

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

1. *LASSO* [18]. Different LASSO models are built for corresponding spatial resolutions. The feature set is the set of keyword counts. The regularization parameter is set as 0.15 based on a 10-fold cross validation on the training set.

2. *Multitask Learning (MTL)* [26]. In multi-task model, each task is the forecasting for each location and spatial resolution. Keyword counts are the features. The regularization parameters $\lambda_1 = 0.015$ and $\lambda = 0.001$ are set based on a 10-fold cross-validation.

3. *Tree-guided Group Lasso for Multi-task Learning (TMTL)* [15]. Here the relationships among the tasks follow the geo-hierarchy defined in Figure 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. The regularization parameter $\lambda = 0.3$ are set based on a 10-fold cross-validation.

4. *Autoregressive exogenous (ARX)* [1]. For each separate location, the count of future events is predicted and dependent on both the counts of 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.

5. *Logistic regression (LR)* [9]. For each spatial resolution, LR utilizes a logit function to map the tweets observations into future event occurrences ("0" denotes no

[1] In addition to the top 3 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.

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

Table 2: Domains for the Experimental Evaluations

Domain	Training period	Test period	Spatial resolution	Datasets
Civil Unrest	2013-01-01∼2013-12-31	2014-01-01∼2014-12-31	country, state, city	Argentina, Brazil, Chile, 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

Table 3: Event forecasting performance on multiple civil unrest datasets

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	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

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

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**

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

occurrence, “1” denotes occurrence). The input features here are the counts of keywords.

6. *Latent Direchlet allocation based Logistic regression (LDA-LR)* [23]. After extracting the latent topics by LDA from the tweets, the LDA-LR model is built on features that are the proportions of the latent topics. Individual models are built for each spatial resolution. The number of topics for each dataset is set based on 10-fold cross-validation.

7. *Kernel density estimation-based logistic regression (KDE-LR)* [12]. This approach utilizes KDE-smoothed historical-event counts and the proportions of latent topics as features, and builds a model for each spatial resolution. The number of topics for each dataset is set based on 10-fold cross-validation.

Table 4: Forecasting performance on influenza outbreak dataset (Precision, Recall, and F-measure)

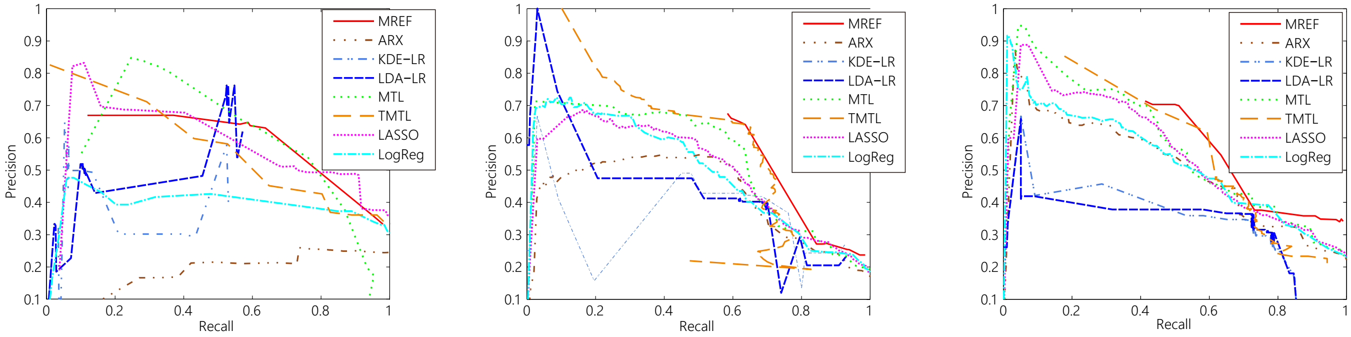
Method	State-level	Region-level	Country-level	Overall	Runtime
ARX	0.09,0.52,0.15	0.12,0.86,0.21	0.58,0.98,0.73	0.26,**0.79**,0.39	**21 sec**
LR	0.08,0.20,0.12	0.26,0.48,0.33	0.85,0.95,**0.90**	0.40,0.54,0.46	37 sec
KDE-LR	0.74,0.07,0.12	0.24,0.21,0.22	**1.00**,0.53,0.69	0.66,0.27,0.38	2026 sec
LDA-LR	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

5.2. Performance

In this section, the performances of all the methods are evaluated and compared. First, the specific spatial event forecasting performance in different spatial resolutions is discussed for civil unrest and influenza outbreaks, after which the Precision-Recall curves for the overall forecasting performance are examined.

5.2.1. Civil unrest event forecasting performance at multiple spatial resolutions. In Table 3, the performance of our MREF and competing methods are compared for civil unrest event forecasting. Three metrics, namely precision, recall, and F-measure, are adopted to quantify the performance. Due to space limitation, only 8 out of the 10 datasets are illustrated; the results for the other 2 datasets, Argentina and Chile, are similar to the 8 shown. Model performance for each of the spatial resolution levels and for the overall performance are shown. The overall performance is calculated based on Definition 3.

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 to



(f) Flu

(e) Uruguay

(c) El salvador

Figure 4: Precision-Recall curves for the performances on different datasets

0.5, while for the state-level, it is typically 0.3 to 0.6, and for country-level, it increases to about 0.8 to 1.0. In general, the performance of the methods utilizing regularization terms is better than for other methods. Specifically, LASSO, MTL, TMTL, and our MREF model achieve better performances for each spatial resolution level than the others. LASSO, MTL, TMTL, and MREF obtain the best overall performance in 7 of the 8 datasets shown, demonstrating the effectiveness of utilizing regularization terms for filtering out unrelated features and ensuring the model's generalizability. Among these, the MTL and TMTL methods also take into account the relatedness of different geographical locations, enabling it to handle the data scarcity inherent in small locations. Apart from location relatedness, MREF also considers the parent-child relationships between locations and supports information sharing among different spatial resolutions. Among the other methods LR, KDE-LR, and LDA-LR, all of which utilize the logistic regression framework, KDE-LR and LDA-LR obtain similar performances because they both consider the latent topics. The performance of ARX is not as good as regularization-based methods, which consistently outperform ARX by 1%∼18%. Of these datasets, all the methods generally achieve better performances for Brazil, which is a large country with a large number of civil unrest events. In all, our MREF model outperforms all the other methods in 6 datasets in overall performance, 5 datasets in city-level performance, 4 datasets in state-level performance, and all 8 datasets in country-level performance. This is because MREF leverages the tasks' relationships in terms of geo-hierarchy, geo-resolution, and geo-parent-child constraints in Theorem 1. Moreover, MREF achieves good performance at the finest granularity, namely city-level, outperforming the other methods by around 9% in 5 datasets and placing second in other 2 more. This is because MREF 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 datasets.

5.2.2. Influenza outbreak event forecasting performance in multiple spatial resolutions. In Table 4, the performance of MREF and the competing methods are illustrated for influenza outbreak event forecasting. Their performances for all the different spatial resolutions and their overall performance have been 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 to 0.2, at the region-level, the F-measure is typically 0.2 to 0.4, and at the country-level, the F-measure increases to about 0.6 to 0.9. In general, the performance of the methods utilizing regularization terms is better than other methods. In particular, LASSO, MTL, TMTL and MREF achieve better performance at each spatial resolution level than the others. For example, LASSO, MTL, TMTL, and MREF obtain the best overall performances, with F-measures of around 0.5, while the other methods are lower, at around 0.3 to 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 achieve similar performances because they both consider the latent topics as features. The performance of ARX is not as good as those of regularization-based methods, which outperforms ARX by over 20%. MREF outperforms all the other methods for overall performance, by 11% at the state-level, 7% at the region-level, and 8% overall. This again demonstrates the advantage enjoyed by MREF due to characterizing the location relatedness and heterogeneity of locations.

5.2.3. Efficiency on running time. The rightmost column of Table 4 shows the training time efficiency comparison for forecasting influenza outbreaks. The running times on test set for all the comparison methods are instant (i.e., less than 0.01 second for one prediction) so that are not provided here. According to Table 4, the running time of ARX was 21 seconds, outperforming the other methods. The running times achieved by all these methods were only a maximum of 40 minutes for 3-year-long huge training set for week-wise event forecasting tasks, making this eminently practical for real-world applications. The efficiency evaluation results on civil unrest datasets follow a similar pattern and are not provided due to the space limitation.

5.2.4. Event forecasting performance on Precision-Recall curves. Figure 4 illustrates the event forecasting overall performance on Precision-Recall curves for 3 datasets in two domains, namely civil unrest and influenza outbreaks.

These curves are drawn by varying the boundary between values for positive and negative predictions. The other civil unrest datasets follow a similar pattern of the "El Salvador" and "Uruguay" datasets and are not provided here due to the space limitation. The calculation of the overall performance again follows that provided in Definition 3. For the 3 datasets shown in Figure 4, MREF generally outperforms the other methods because it is in most cases the closest to (1,1) points in the plots. Moreover, the ROC curves of MREF are consistently above the other methods in these datasets, when precision and recall vary. Other than MREF, 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 obtains a particularly poor performance for the flu dataset, although it achieves an average performance in the other datasets.

6. Conclusion

The accuracy and discernibility of a spatial event forecasting model are two key concerns. But the joint consideration and optimization of them suffer from several challenges. In this paper, we propose a new multi-resolution spatial event forecasting framework to address all the challenges simultaneously. To achieve this, we propose a novel multi-task learning model that leverages the heterogeneous relationships among the prediction tasks, and develop an effective parameter optimization algorithm based on ADMM. Experiments on 11 datasets in two different domains were conducted to evaluate the performance and parameter sensitivity of the proposed model. The results demonstrated that because of the effective utilization of the shared information across different spatial resolutions and neighborhoods, the proposed model outperforms the other comparison methods.

Acknowledgment

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

References

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

[2] D. Agarwal, R. Agrawal, R. Khanna, and N. Kota. Estimating rates of rare events with multiple hierarchies through scalable log-linear models. In *KDD 2010*, pages 213–222. ACM, 2010.

[3] R. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[4] A. Argyriou, T. Evgeniou, and M. Pontil. Multi-task feature learning. *NIPS 2007*, 19:41, 2007.

[5] M. Arias, A. Arratia, and R. Xuriguera. Forecasting with Twitter data. *TIST*, 5(1):8, 2013.

[6] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

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

[8] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[9] R. Compton, C. Lee, J. Xu, L. Artieda-Moncada, T.-C. Lu, L. De Silva, and M. Macy. Using publicly visible social media to build detailed forecasts of civil unrest. *Security Informatics*, 3(1):1–10, 2014.

[10] X. Dong, D. Mavroeidis, F. Calabrese, and P. Frossard. Multiscale event detection in social media. *Data Mining and Knowledge Discovery*, 29(5):1374–1405, 2015.

[11] T. Evgeniou and M. Pontil. Regularized multi-task learning. In *KDD 2004*, pages 109–117. ACM, 2004.

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

[13] Y. Jiang, C.-S. Perng, and T. Li. Meta: Multi-resolution framework for event summarization. In *SDM*, pages 605–613. SIAM, 2014.

[14] F. Jin, W. Wang, L. Zhao, E. Dougherty, et al. Misinformation propagation in the age of twitter. *Computer*, (12):90–94, 2014.

[15] S. Kim and E. P. Xing. Tree-guided group lasso for multi-task regression with structured sparsity. In *ICML 2010*, pages 543–550, 2010.

[16] MITRE. http://www.mitre.org/.

[17] M. J. Paul and M. Dredze. A model for mining public health topics from Twitter. *Health*, 11:16–6, 2012.

[18] N. Ramakrishnan, P. Butler, S. Muthiah, N. Self, et al. 'Beating the news' with EMBERS: Forecasting civil unrest using open source indicators. In *KDD 2014*, pages 1799–1808. ACM, 2014.

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

[20] E. Schubert, M. Weiler, and H.-P. Kriegel. Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds. In *KDD 2014*, pages 871–880. ACM, 2014.

[21] A. Signorini, A. M. Segre, and P. M. Polgreen. The use of Twitter to track levels of disease activity and public concern in the us during the influenza an H1N1 pandemic. *PloS one*, 6(5):e19467, 2011.

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

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

[24] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan. Spatiotemporal event forecasting in social media. In *SDM 15*, pages 963–971. SIAM, 2015.

[25] 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 *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 639–648. IEEE, 2015.

[26] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *KDD 2015*, pages 1503–1512. ACM, 2015.

[27] L. Zhao, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan. Hierarchical incomplete multi-source feature learning for spatiotemporal event forecasting. pages 2085–2094, 2016.

[28] J. Zhou, J. Chen, and J. Ye. *MALSAR: Multi-tAsk Learning via StructurAl Regularization*. Arizona State University, 2011.