Efficient Learning with Exponentially-Many Conjunctive Precursors to Forecast Spatial Events

Liang Zhao, Feng Chen, and Yanfang Ye

Abstract—Forecasting spatial societal events in social media is significant and challenging. Most existing methods consider the frequencies of keywords or n-grams to be features, but have not explored the exponentially large space of the conjunctions of those features, such as keyword co-occurrence in messages, which can serve as crucial precursor rules. Due to the inherent exponential complexity of ensemble rule learning, existing work typically adopts greedy/heuristic strategies. This means that they cannot guarantee the solution's optimality, which would require a considerably more sophisticated model for spatial event forecasting, while still suffering from major challenges: 1) Exponentially-dimensional feature learning with distant supervision, 2) Numerical values of conjunctive features, and 3) Spatially heterogeneous conjunction patterns. To concurrently address all these challenges with a theoretical guarantee, we propose a novel spatial event forecasting model which learns numerical conjunctive features efficiently. Specifically, to consider their magnitude, traditional Boolean rules are innovatively generalized to deal with numerical conjunctive features with amenable computational properties. To handle the geographical similarity and heterogeneity in numerical conjunctive feature learning, we propose a new model that implements through a new bi-space hierarchical sparsity regularization for locations and features. Moreover, we propose a new algorithm to optimize the model parameters and prove that it enjoys theoretical guarantees for both the error bounds and time efficiency. Extensive experiments on multiple datasets demonstrate the effectiveness and efficiency of the proposed method.

Index Terms—conjunctive feature learning, spatial event forecasting, multi-task learning, hierarchical kernel learning

1 Introduction

Currently, user-generated contents such as microblogs have become ubiquitous, which serve as real-time "sensors" for social trends and incidents [26]. People use social media to plan, advertise, and organize future social events such as the planned protests in the "Arab Spring" and "Occupy Wall Street" [28]. The predictive power of microblogs for social event forecasting has been widely explored by a great deal of recent research on topics such as crimes [17], civil unrest [40], and disease outbreaks [2]. These research works share essentially similar workflows. First, the model features are typically defined as the counts of terms (e.g., keywords and hashtags) under the domain of interest. The feature values in the aggregated collections of massive microblogs are considered to jointly reflect the social tendencies. The predictive model is then trained to map the social indicators to the model response, in this case the occurrence of future events.

However, the count for a single keyword may not be sufficiently informative to serve as a precursor for forecasting social events. For example, Figure 1(a) shows that instead of either "teacher" or "reform", the count of their conjunction in the same tweets reflects the public concern regarding educational policy. Similarly, as shown in Figure 1(b), the count for anyone of "election", "president", and "fraud" individually is a very noisy signal, while the massive co-occurrence of them all in the same tweets is a very informative precursor for the subsequent three waves of protests against the results of the presidential election in Mexico. Therefore, unlike the incidence of single keywords, the co-occurrence of keywords, such as "president+election+fraud" typically conveys much more definite meaning and is thus a significantly more powerful precursor for future protest events. In this paper, we call such new features *conjunctive features*, which in this case refer to two or more co-occurring keywords in Figure 1, and the standalone atomic features as *primitive features*, which here mean single keywords.

Conjunctive features are highly informative and thus more interpretable by human observers, which is crucial if decision makers are to understand and utilize the predictive models. However, it is impossible for domain experts to manually provide an extensive set of all the keyword conjunctions that are precursors. Better methods for automatically learning an extensive set of significant keyword conjunctions from the data are clearly required, but due to the exponential complexity in storage and time of computation, this problem is conventionally unfeasible even for a moderate sized of keyword set. For example, among as few as 100 keywords, there are $2^{100} \approx 10^{30}$ possible combinations of conjunctive features to store and compute, which are far beyond the existing memory capacity and computational power.

Existing methods on supervised rule mining typically utilize greedy or heuristic-based methods [15], where the

• *Liang Zhao was with the Department of Information Science and Technology, George Mason University.*
E-mail: lzhao9@gmu.edu

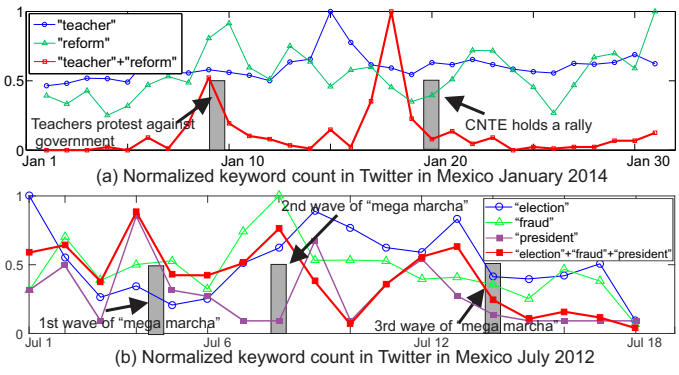


Figure 1: The burstiness of the counts of some keyword co-occurrences preceded the major events. The keyword co-occurrence features could be much more determinative and interpretable than single keywords as precursors for future events.

optimality of the solution cannot be guaranteed. Despite the significance of this problem, to the best of our knowledge, there has been very little work reported on spatial event forecasting that extensively considers conjunctive features. Even utilizing the simplest settings, several substantial theoretical and practical challenges make this problem unfeasible to solve: **1) Numerical values of conjunctive features.** Existing methods for conjunctive feature learning typically require the conjunctive features to be Boolean-valued if efficient computation is to be feasible. However, in spatial event forecasting, instead of binary values, the numerical frequencies of the keyword conjunctions occurring in the same location at a specific time serves as the indicator, which does not satisfy the Boolean assumption that is universally applied in existing work. **2) Exponentially-dimensional learning with distant supervision.** Due to the immense volume of microblog messages, it is prohibitively labor-intensive to label each individual message. When forecasting spatial events, typically only labels at the aggregate level (e.g., city-level) are available, which can thus only provide distant supervision when learning important conjunctive features. **3) Spatially heterogeneous conjunction patterns.** Different geo-locations may share similar conjunctive features but also have their own exclusive ones within a particular geo-neighborhood. For example, "occupy+street" can be a good indicator for general civil unrest across many different geo-locations, but "occupy+wall+street" typically appears in New York State while "occupy+Texas+state" typically happens in Texas.

In order to simultaneously overcome all the above-mentioned challenges, we propose a novel model named Hierarchical-Task Numerical conjunctive feature Learning (HTNL) for spatial event forecasting. Specifically, a novel kernel is formulated to represent every possible numerical conjunctive feature (NCF) and then this exponentially large set of kernels is correlated via a Directed Acyclic Graph (DAG). The complexity in NCF selection is reduced from exponential to polynomial by utilizing the sparsity structure from the DAG and the favorable computational properties of the proposed kernel. Finally, the similarities between the selected NCFs within geo-hierarchical neighborhoods are enforced to boost model generalizability using the newly proposed bi-space regularization strategy in both feature and location spaces. The major contributions of this paper are as follows:

- **Develop a generic framework for conjunctive precursor learning for spatial event forecasting.** A generic framework is proposed for spatial event forecasting that optimally learns the NCFs, taking into account both geographical similarity and heterogeneity. A number of related classic approaches are shown to be special cases of our model.
- **Propose a novel hierarchical multitask model for NCF learning.** First, every possible NCF is formulated as a novel kernel with structured sparsity on a DAG. Then the similarity of sparsity patterns is enforced using a newly proposed bi-space regularization strategy that utilizes geo-hierarchical knowledge to boost up model generalizability.
- **Design an efficient optimization method with a theoretical guarantee of optimality.** The proposed model requires the optimization of an NCF set that is exponentially large and geographically correlated. The new algorithm leverages both the topological sparsity among NCFs and the computational efficiency of the proposed kernel, and provides theoretical guarantees for both error bounds and time complexity.
- **Conduct extensive experiments for performance evaluations.** The proposed method is evaluated on multiple datasets in different domains and found to significantly outperform the existing methods in prediction performance. Moreover, the conjunctive features discovered by the model clearly demonstrate its effectiveness and interpretability.

The rest of this paper is organized as follows. Section 2 reviews the background and related work and Section 3 introduces the problem setup. Sections 4 and 5 presents our proposed model and an efficient model parameter optimization algorithm, respectively. The experiments on synthetic and real-world datasets are presented in Section 6, and the paper concludes with a summary of the research in Section 7.

2 Related Work

Event Detection and Forecasting in Social Media. A considerable amount of work has been done on detecting ongoing events, including disease outbreaks [31], earthquakes [30] and various other types of events [38], [26]. Generally, for event detection, either classification or clustering is utilized to extract tweets of interest and then the spatial [30], temporal [31], or spatiotemporal burstiness [14] of the extracted tweets is examined to identify the potential occurrence of an ongoing event. However, these approaches typically uncover events only after they have commenced. To forecast future events, several event forecasting methods have been proposed, most of which focus on temporal events and ignore the underlying geographical information, such as the forecasting of elections [35], stock market movements [7], disease outbreaks [2], box office ticket sales [4], crimes [36], and others [39], [10]. These works typically utilize linear/nonlinear regression models [4], [7] or time series-based methods [2]. Few existing approaches provide true spatiotemporal resolution for predicted events. In [17], Gerber used logistic regression for spatiotemporal event

forecasting using topic-related tweet volumes as features, while Ramakrishnan et al. [28] built separate LASSO models for different locations to predict the occurrence of civil unrest events. Zhao et al. [39] proposed a multi-task learning framework that jointly learns multiple related spatial locations. But it requires extra knowledge on dynamic features. Innovatively formulating the mobility prediction in transportation system as a video prediction task, StepDeep is proposed by Shen et al. [32] based on novel spatial-temporal convolution layers. The method StepDeep requires spatial grid data as input, which cannot be adapted into the setting of this paper.

Rule Ensemble Learning (REL). Given a set of basic propositional features describing the data, the goal of REL is to supervisedly learn a set of feature conjunctions with good predictability. To handle the inherent exponential complexity of this problem, many REL methods have been proposed majorly in three categories: 1) *filter-based methods*, which assume that important conjunctive features must be frequent and thus only retain frequent instances for classification [9], [11], [33]. This category is correlated to *discriminative frequent pattern mining* [9]. However, frequency and the predictability of features are not equivalent because predictability is dependent on the specific prediction task while frequency is not; 2) *heuristic/boosting-based methods*, where researchers address the challenge in Category 1, by learning the feature conjunctions and the predictive model concurrently [29], [25], [15]. To ensure computational efficiency, heuristic strategies based on greedy or boosting methods are generally utilized. And only sub-optimum or local optimum can be found. 3) *optimization-based methods*: To address the problem in Category 2 and ensure efficiency, recently few methods have been proposed for conjunctive feature selection with theoretical guarantee on the error bound to the global optima [5], [20]. This is achieved by utilizing an active set algorithm to scale down the solution space. To check the optimality of current active set efficiently, the "product-of-sum" property [5] of Boolean rules must be exploited. However, existing optimization-based methods do not apply in more general situations where the rules are numerical because they violate the "product-of-sum" property. Classic approaches such as discretizating the numerical values into multiple binary features arbitrarily scale up the number of basic propositional features and thus exponentially enlarge solution space. In order to address this problem, our paper proposes a new method that can directly handle numerical rules efficiently without discretization.

Multi-task learning: Multi-task learning (MTL) learns multiple related tasks simultaneously to improve generalization performance [3], [24]. Many MTL approaches have been proposed over the last decade [34]. In [21], Kim et al. proposed a regularized MTL which constrained the models of all tasks to be close to each other. The task relatedness can also be modeled by constraining multiple tasks to share a common underlying structure, e.g., a common set of features [37], or a common subspace [1]. MTL approaches have been applied in many domains including computer vision and bioinformatics.

3 Problem Formulation

In this section, the problem in this paper is formulated. Section 3.1 poses the problem of "precursor rule learning for event forecasting". Denote $X = \{X_{s,t}\}_{s,t}^{S,T}$ as a collection of input data (e.g., microblog data), where T is the set for time intervals and S is the set of the spatial locations. $X_{s,t}$ denotes the data for tth time interval (e.g., tth date) at location s such that $X_{s,t} \in \mathbb{Z}^{n_{s,t} \times |V|}$, where $n_{s,t}$ denotes the number of microblog messages sent during time interval t at location s, and $|V|$ denotes the size of the vocabulary V, which is a set of *primitive features* that can include occurrences of specific keywords, hashtags, and hyperlinks. $X_{s,t}$ is defined as a matrix whose element $[X_{s,t}]_{i,v} \in \{0, 1\}$ denotes the occurrence (with value 1) or not (with value 0) of the primitive feature v in the ith message in location s during time interval t. The important notations in this paper are listed in Table 1.

As explained earlier, a *conjunctive feature* (or *feature conjunction*) is defined as the conjunction of a set of distinct primitive features such as keywords that co-occur in the same message. Hence, the set of all the possible conjunctive features is denoted as $\mathcal{V} = \{v | v \subseteq V\}$, whose size is $|\mathcal{V}| = 2^{|V|}$. $\phi_v(X_{s,t}) \in \mathbb{R}^+ \bigcup \{0\}$ denotes the frequency of the conjunctive feature v in location $s \in S$ at time $t \in T$. Therefore, instead of assigning this a Boolean value, in our problem the conjunctive feature is generalized to a numerical value, referred to as the ***numerical conjunctive feature (NCF)***.

NCFs have topological relationships with each other. We denote these relationships using a *directed acyclic graph (DAG)* known as a *feature conjunction lattice*: $\mathcal{G}(\mathcal{V}, \mathcal{E})$, as illustrated in Figure 2(b). In a feature conjunction lattice, the top node (i.e., Level 0) is an empty conjunctive feature while the nodes in Level 1 are the primitive features V. A node $v_1 \in \mathcal{V}$ is called the *parent* of another node $v_2 \in \mathcal{V}$ if $v_2 \subset v_1$ and $|v_2| + 1 = |v_1|$; hence v_2 is a *child* of v_1. Let $D(v)$ and $A(v)$ denote the set of descendants and ancestors of $v \in \mathcal{V}$, respectively. We assume that both $D(v)$ and $A(v)$ include the node v. For a subset of nodes $\mathcal{U} \subset \mathcal{V}$, we define the *hull* and *sources* of $\mathcal{U}$ as $H(\mathcal{U}) = \bigcup_{v \in \mathcal{U}} A(v)$ and $S(\mathcal{U}) = \{v | A(v) \cap \mathcal{U} = \{v\}\}$, respectively. $|\mathcal{U}|$ denotes the number of NCFs in set $\mathcal{U}$ while $\bar{\mathcal{U}}$ denotes the complementary set of $\mathcal{U}$, namely all the NCFs that are in $\mathcal{V}$ but not in $\mathcal{U}$.

Define $Y = \{Y_{s,t}\}_{s,t}^{S,T}$ as the event occurrences, where $Y_{s,t} \in \{1, -1\}$ such that $Y_{s,t} = 1$ means there is an event in location s at time t, otherwise $Y_{s,t} = -1$. The following is our problem definition of spatial forecasting task: Given the input data $X_s = \{X_{s,t}\}_t^T$ for location $s \in S$, and the primitive feature set V, our goal is to predict the output, namely the future event occurrence $Y_{s,\tau}$. And in the meanwhile, we also discover the set of NCFs $\mathcal{U}_s \subseteq \mathcal{V}$ that are crucial precursors for a future event in each location s, and thus learn a mapping function for event forecasting:

$$f : \{\phi_v(X_{s,t})\}_{v \in \mathcal{U}_s} \rightarrow Y_{s,\tau} \tag{1}$$

where $\tau = t + q$, and q is the lead time for forecasting. Among all of the $|\mathcal{V}| = 2^{|V|}$ candidate NCFs, typically only a few are useful precursors for forecasting.

Table 1: Important Notations

Notations	Explanations
$X_{s,t}$ and $Y_{s,t}$	Input data and event occurrence in location s at time t
$\mathcal{V}$	The set of all the conjunctive features
$\phi_v(X_{s,t})$	The frequency of $v \in \mathcal{V}$ in location s at time t
$D(v)$ and $A(v)$	The sets of descendants and ancestors of $v \in \mathcal{V}$
W_s	Weight vector for the conjunctive features in location s
G	The set of geographical neighborhoods
$\Theta_{d,r}$	Boundary and interior of a generalized d-simplex
$\alpha_{s,t}$	dual variable for location s at time t
p, $\hat{p}$, and $\bar{p}$	p, $\hat{p}$, and $\bar{p}$-norms, $\bar{p} = \hat{p}/(\hat{p}-1)$, $\hat{p} = p/(2-p)$

There are three technical challenges involved in solving this problem. **First, exponential solution space.** This problem is extremely difficult to solve even for a modest size of V because of the exponentially large size of $|\mathcal{V}|$, which causes intractability in both memory and computation. **Second, numerical values of conjunctive features.** To ensure an efficient solution, the state-of-the-art methods require a conjunctive feature to have a Boolean value. However, in our problem the conjunctive feature value $\phi_v(X_{s,t})$ must be numerical and therefore the Boolean assumption is not satisfied, creating a serious challenge for the model efficiency. **Third, geographical influences in NCF learning**. For spatial event forecasting, both the geographical relationship and heterogeneity in the conjunctive feature learning are crucial and must be considered, a combination that has never been addressed by existing methods. To address all three of these challenges, we have developed the new model presented in the following section.

4 MODEL

We propose a new model, HTNL, to address the challenges described above. First, NCF is mathematically defined and analyzed in Section 4.1. Second, geographical relationships and heterogeneity are considered for NCFs in Section 4.2. The final objective function and its relation to existing models are proposed in Sections 4.3 and 4.4, respectively.

4.1 Computational Properties of NCFs

The state of the art requires a Boolean value for conjunctive features to ensure efficiency, because this is the only way the conjunctive features can be efficiently computed by multiplying the primitive features that they consist of. However, for our problem of spatial event forecasting, the Boolean assumption is not satisfied and a new and generic version, namely NCF, is required. To ensure the computational efficiency is retained in such a generalized setting, the unique formulation and properties of NCF are explored in the following.

4.1.1 Calculation of the NCF

As noted in Section 3, the value of NCF v in location s at time t, namely $\phi_v(X_{s,t})$, is defined as the spatiotemporally accumulated occurrence of NCF v. Given that $[X_{s,t}]_{v,i} \in \{0,1\}$, $\phi_v(X_{s,t})$ is computed as:

$$\phi_v(X_{s,t}) = \sum_i \Big(\bigwedge\nolimits_{j \in v} [X_{s,t}]_{i,j} \Big) = \sum_i \prod\nolimits_{j \in v} [X_{s,t}]_{i,j} \quad (2)$$

where $\bigwedge_{j \in v}[X_{s,t}]_{i,j}$ is the logical "and" among the values of the primitive features. Equation (2) builds a logical mapping between a spatial location and all the messages it contains. This mapping is important because it enables us to leverage the distant supervision on spatial-location level (i.e., the event occurrence label $Y_{s,t}$ for each location which contains multiple messages inside it) to learn the feature occurrence patterns, which is finer-grained on message-level.

4.1.2 The kernel that induces the feature mapping $\phi_v(X_{s,t})$

The computation of NCF $\phi_v(X_{s,t})$ is a nonlinear mapping from the input. In the following, we prove that $\phi_v(X_{s,t})$ is induced by a kernel and thus can benefit from efficient computation through kernel methods and kernel hierarchy.

Lemma 1. *$k_v(X_{s,t}, X_{s',t'}) = \phi_v(X_{s,t}) \cdot \phi_v(X_{s',t'})$ is a kernel.*

The proof of Lemma 1 is in Appendix A. The predictive mapping f in Equation (1) can be instantiated as the linear combination of a subset of NCFs: $f(W, \{\phi_v(X_{s,t})\}_v^{\mathcal{U}}) = \sum_v^{\mathcal{U}} W_{s,v}\phi_v(X_{s,t}) + b$, where $W_{s,v}$ represents the weight of the NCF v for location s. Thus, learning such a mapping function is equivalent to optimizing a subset of $\mathcal{U}$ and their corresponding weights $W = \{W_{s,v}\}_{s,v}^{S,\mathcal{U}}$. Mathematically, this can be achieved by jointly optimizing the empirical risk term and regularization term:

$$\min_{W_{s,v}} C \sum_{s,t}^{S,T} \mathcal{L}(Y_{s,t}, f(W_{s,\cdot}, \{\phi_v(X_{s,t})\}_v^{\mathcal{V}})) + \sum_s^S \frac{1}{2}\Omega^2(W_s) \quad (3)$$

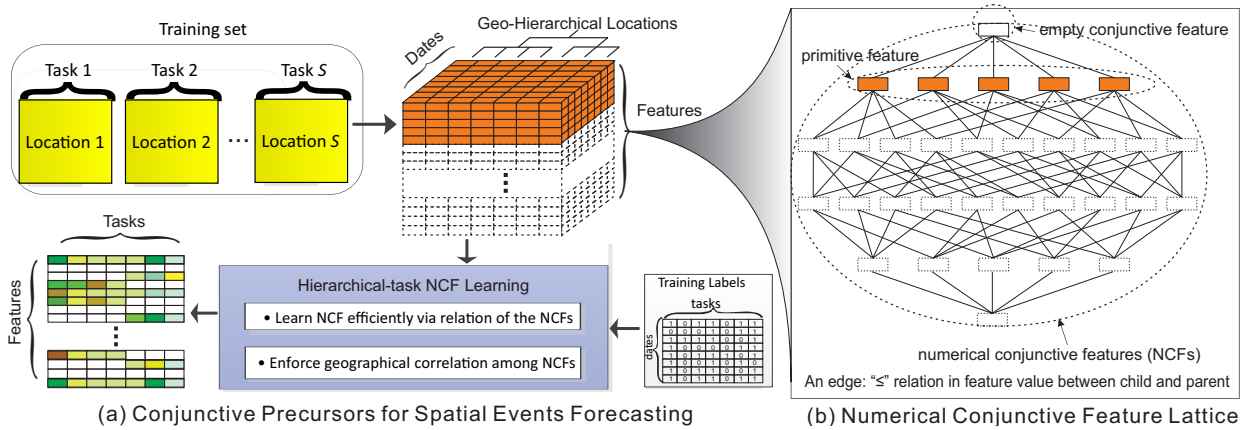
where $\mathcal{L}(\cdot)$ is the loss function, which is convex and proper. To address the classification problem, this could be a hinge loss. $\Omega(\cdot)$ is the regularization term that enforces sparsity so that only a few $W_{s,v}$ will retain nonzero values to form the subset $\mathcal{U}$. Due to the property in Lemma 1, the efficient *representer theorem* [16] can be utilized to formulate the predictive function $f(W, \{\phi_v(X_{s,t})\}_v^{\mathcal{U}})$ as a linear combination of hierarchical kernels. We denote $W_{s,\cdot} = \{W_{s,v}\}_v^{\mathcal{V}}$. The major computational challenge in solving Equation (3) comes from the large size of $D(v)$, which is exponential to $|V|-|v|$, Thanks to the favorable properties of our proposed kernel in Lemma 1, this computation can be reduced to be linear with $|V|$, which will be proved in Theorem 2 in Section 5.

4.2 Geographical relationships of NCFs

The geographical relationships of NCFs include both geographical similarity and geographical heterogeneity.

4.2.1 Geographical similarity: general conjunctive precursors.

For a domain of interest, the sparsity among NCFs for different locations can be learned jointly because they follow



(a) Conjunctive Precursors for Spatial Events Forecasting (b) Numerical Conjunctive Feature Lattice

Figure 2: The proposed hierarchical-task numerical conjunctive feature learning (HTNL). (a) The flowchart of proposed HTNL. (b) The NCF lattice where the edge denotes the "$\leq$" relation among NCFs

the same DAG relation shown in Figure 2. Specifically, the majority of the important conjunctive features tend to be the smaller ones (i.e., those consisting of fewer primitive features), while most long conjunctive features can normally be enforced to zeros. To achieve this, we enforce $\ell_{p,1}$-norm ($p \in (1, 2)$) on the norms of the descendants of each NCF so that longer conjunctive features will be subject to greater penalties.

$$\Omega(W) = \sum\nolimits_{v\in\mathcal{V}} d_v \left(\sum\nolimits_{u\in D(v)} \left(r_{D(v)}(W_{\cdot,v}) \right)^p \right)^{\frac{1}{p}} \quad (4)$$

where $d_v = a^{|v|}$, $a > 0$ is the regularization parameter corresponding to NCF with a specific length, and $p = (1, 2]$ controls the sparsity. An NCF can be selected only when all of its ancestors are selected, in which case $p = 2$; otherwise, an NCF could be selected even if its ancestors are zeros. $r_{D(v)}(W_{\cdot,v})$ is the norm for each NCF which will be detailed in next subsection in Equation (5).

4.2.2 Geographical heterogeneity: regional conjunctive precursors.

Although different locations may share similar general textual expressions, the strength of this similarity typically varies. The textual expressions within the same spatial neighborhood tends to be more similar than those far away. For example, events that occur at neighboring locations at around the same time could well involve similar topics, so the texts from neighboring locations may share a number of common keyword conjunctions that are related to the events. As shown in Figure 3, the top popular civil-unrest-related conjunctive features for different major cities in major Latin American countries in January 2013 are shown. Here conjunctive features with similar meanings are marked by similar colors. It can be seen that the cities in the same spatial regions or the same countries are more likely to have similar popular conjunctive features. And the locations in different geographical neighborhoods (e.g., regions, provinces, countries, and other administrative or geographical divisions) may probably have distinct conjunctive features due to their different public concerns, natural features, and societal issues.

Here we treat the event prediction for each geographical location $s \in S$ as a task. This means $X_{s,t}$ and $Y_{s,t}$ are the input and output for time t in the s-task, respectively. To

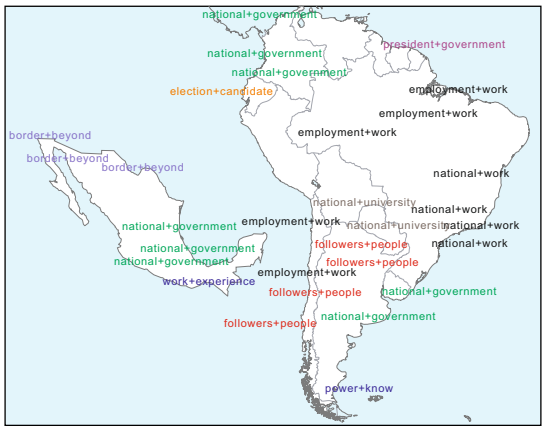


Figure 3: Under civil unrest domain, the top popular conjunctive features for different major cities in major Latin American countries in Jan 2013. It shows some geographical regions share similar conjunctive feature patterns.

take into account such geographical heterogeneity, we propose a new hierarchical multi-task learning strategy for each NCF's norm $r_{D(v)}(W_{\cdot,v})$ by enforcing an ℓ_2-norm on each geographical neighborhood $g \subseteq S$. Using such scenario, the above-mentioned geo-heterogeneity among different spatial locations S will be jointly considered, with those locations j in the same geo-neighborhood g more similar, which is enforced by the ℓ_2-norm which is also named group Lasso:

$$r_{D(v)}(W_{\cdot,v}) = \sum\nolimits_g^G \|\{W_{j,u}\}_{j\in g}\|_2, \; u \in D(v) \quad (5)$$

where G is the set of all the geographical neighborhoods. Combining Equations (4) and (5), we propose the following novel bi-space regularization term:

$$\Omega(W) = \sum_{v\in\mathcal{V}} d_v \left(\sum\nolimits_{u\in D(v)} \left(\sum\nolimits_g^G \|\{W_{j,u}\}_{j\in g}\|_2 \right)^p \right)^{\frac{1}{p}} \quad (6)$$

where the sparsity of NCFs are enforced by considering the hierarchical structures in two spaces. Specifically, the conjunction relation in DAG is modeled by the outer $\ell_{p,1}$-norm and geographical hierarchy is modeled by the inner $\ell_{2,1}$-norm.

4.3 Objective function of HTNL

The regularization term in Equation (6) is non-smooth and multilevel. To simplify its form, the following elegant equiv-

alence is proposed in Lemma 2 which is proved in Appendix B:

Lemma 2. *Define* $\Theta_{d,r} = \{x \in \mathbb{R}^d | x \geq 0, \sum_i x_i^r \leq 1\}$. *The regularization term* $\Omega(W)$ *defined in Equation 6 can be transferred to an equivalent problem as follows:*

$$\Omega(W) = \min_{\gamma,\lambda,\mu} \sum_{v\in\mathcal{V}} \frac{d_v^2}{\gamma_v} \sum_{u\in D(v)} \frac{1}{\lambda_{u,v}} \sum_g^G \frac{1}{\mu_{u,v,g}} \sum_{j\in g} \|W_{j,u}\|^2 \quad (7)$$

where $\gamma \in \Theta_{|\mathcal{V}|,1}$, $\lambda_v = \Theta_{|D(v)|,\hat{p}}$, *and* $\mu_{u,v,g} = \Theta_{|G|,1}$. $\hat{p} = p/(2-p)$.

The proof of Lemma 2 is in Appendix B. The equivalent form is an elegant quadratic form of the weights $W_{j,u}$, making it possible to utilize the representer theorem to solve the empirical risk minimization problem. By introducing this equivalent form into Equation (3), we obtain:

$$\min_{\gamma,\lambda,\mu,W} C \sum_{s,t}^{S,T} \mathcal{L}(Y_{s,t}, f(W_s, \{\phi_v(X_{s,t})\}_v^{\mathcal{V}})) + \sum_{u\in\mathcal{V}} \sum_g^G \Psi_{u,g}^{-1}(\gamma,\lambda,\mu) \sum_{j\in g} \|W_{j,u}\|^2 \quad (8)$$

where $\Psi_{u,g}(\gamma,\lambda,\mu) = \left(\sum_{v\in A(u)} \frac{d_v^2}{\gamma_v \lambda_{v,u} \mu_{v,u,g}}\right)^{-1}$. The above problem is convex in γ, λ, μ, and W. $\mathcal{L}(\cdot)$ is the hinge loss.

4.4 Relationship to previous models

In this section, we show that our proposed model HTNL is the general form of several state-of-the-art models:

4.4.1 Generalization of multiple kernel learning

Let $p = 2$, $\Omega(\cdot) = \|W\|_2$, $|G| = 1$, and $n_{s,t} \equiv 1, \forall s,t \in S,T$. Our model in Equation (8) is reduced to multiple kernel learning [18]:

$$\min_w C \sum_i^n \mathcal{L}(y_i, f(w, x_i)) + \|w\|_2^2$$

where w is the set of feature weights, n is the number of samples, x_i and y_i are the ith input and output of the model.

4.4.2 Generalization of hierarchical kernel learning

Let $p = 2$, $|G| = 1$, and only allow Boolean conjunctive features, i.e., $n_{s,t} \equiv 1, \forall s,t \in S,T$. Our model in Equation (8) is thus reduced to hierarchical kernel learning [5]:

$$\min_{\gamma,w} C \sum_i^n \mathcal{L}(y_i, f(w, \{\Pi_{j\in v} x_{i,j}\}_v^{\mathcal{V}})) + \sum_{u\in\mathcal{V}} \Psi_{u,1}^{-1}(\gamma,1,1) \|w_u\|_2^2$$

here $x_{i,j}$ is a binary value of jth primitive feature in ith input.

4.4.3 Generalization of generalized hierarchical kernel learning

Let $|G| = 1$ and only allow Boolean conjunctive features, i.e.,$n_{s,t} \equiv 1, \forall s,t \in S,T$. Our model in Equation (8) is thus reduced to generalized hierarchical kernel learning [19]:

$$\min_{\gamma,\lambda,w} C \sum_k^m \sum_i^{n_k} \mathcal{L}(y_{k,i}, f(w_k, \{\Pi_{j\in v} x_{k,i,j}\}_v^{\mathcal{V}})) + \sum_{u\in\mathcal{V}} \Psi_{u,1}^{-1}(\gamma,\lambda,1) \sum_k^m \|w_{k,u}\|^2$$

where $x_{k,i,j}$ is the binary value of the jth primitive feature in the ith input of the kth task and $w_{k,u}$ is the weight value of the uth feature of the kth task.

Algorithm 1 Hierarchical-multitask NCF Learning

Require: X, Y, C, G, and $\mathcal{V}$.
Ensure: solution W and b.
Initialize $\mathcal{U} = S(\mathcal{V})$, $\mathcal{W}_m, \Gamma, \Phi = \mathbf{0}$.
repeat
  repeat
    Normalize $\eta \leftarrow \eta / \sum_v^{\mathcal{U}} \eta_v$
    $\eta_u(\beta) \leftarrow \left(\sum_{v\in A(u)} d_v^p \beta_v^{(1-p)}\right)^{1/(1-p)}$
    Initialize $\xi_{g,u} = 1/|G|,\ u \in \mathcal{U}$
    repeat
      $\alpha \leftarrow$ solve Equation (13) given ξ
      $\xi \leftarrow$ solve Equation (13) given α
    until convergence
    step size $d \leftarrow \sqrt{\log(\mathcal{U})/k} / \|\nabla H(\eta)\|_\infty$
    $\eta \leftarrow \exp \mathbf{1} + \log \eta - s \cdot \nabla H(\eta)$
  until Convergence
  if Equation (14) is satisfied **then**
    break
  else
    Add the nodes violating Equation (14) to $\mathcal{U}$
  end if
until Forever

5 Optimization Algorithm

In this section, we propose a new efficient algorithm to solve the objective function of HTNL model in Equation (8) by leveraging its dual solutions. Specially, first, its dual form is proposed and simplified in Section 5.1 and then solved by the proposed algorithm described in Sections 5.2 and 5.3. Finally, theoretical analyses of the convergence and time complexity are presented in Section 5.4.

5.1 Duality form

The primal form in Equation (8) of the objective function can be reformulated into the following duality form:

$$\min_{\gamma,\lambda,\mu} \max_\alpha \sum_{t\in T, s\in S} \alpha_{s,t} - \frac{1}{2} \sum_u^{\mathcal{V}} \sum_g^G \Psi_{v,u,g}(\gamma,\lambda,\mu) \cdot h(g,u)$$
$$s.t. \quad \sum_{t\in T} \alpha_{s,t} \cdot y_{s,t} = 0, \forall s \in S$$
$$0 \leq \alpha_{s,t} \leq C, \forall s \in S, \forall t \in T \quad (9)$$

where $\alpha = \{\alpha_{s,t}\}_{s,t}^{S,T}$ is the dual variable. $h(g,u) = \sum_j^g \sum_{i,k} \alpha_{j,i} y_{j,i}\ \phi_u(X_{j,i}) \phi_u(X_{j,k}) y_{j,k} \alpha_{j,k}$. The above function is convex in γ, μ, and λ and concave in α. However, the problem as stated involves too many variables and is thus difficult to solve efficiently. To address this problem, Theorem 1 proposes a simplified equivalent formation.

Theorem 1. *The objective function in Equation* (9) *can be simplified into the following equivalent form.*

$$\min_\beta \max_\alpha \sum_{t\in T, s\in S} \alpha_{s,t} - \frac{1}{2} \left(\sum_{u\in\mathcal{V}} \eta_u(\beta) \cdot \hat{h}(u)^{\bar{p}} \right)^{1/\bar{p}} \quad (10)$$

where $\eta_u(\beta) = (\sum_{v\in A(u)} d_v^p \beta_v^{(1-p)})^{1/(1-p)}$. $\hat{h}(u) = \max_{g\in G} h(g,u)$.

The proof of Theorem 1 is in Appendix C.

5.2 Active Set Algorithm

Because the size of all the possible NCFs $\mathcal{V}$ is exponential to the size of the primitive features V, Equation (10) can easily be computationally unfeasible to solve even with a moderate size of V. To handle this problem, the sparsity of $\mathcal{V}$ is taken into account. This means that for the optimal

solution to Equation (10), most members of $v \in \mathcal{V}$ should be 0. Thus, solving the original problem in Equation (10) is equivalent to solving the following subproblem in Equation (11) where only the small subset of non-zero variables at the optimal solution of Equation (10) need to be involved. The computational effort required in the latter case will be significantly lower.

$$\min_{\beta} \max_{\alpha} \sum_{t \in T, s \in S} \alpha_{s,t} - \frac{1}{2} \left(\sum_{u \in \mathcal{U}} \eta_u(\beta) \cdot \hat{h}(u)^{\bar{p}} \right)^{1/\bar{p}} \quad (11)$$

where $\mathcal{U} = \{u | W_u \neq 0, u \in \mathcal{V}\}$ is the set of nonzero-weighted NCFs. This function is convex in β but concave in α.

However, the non-zero variables at the optimum are unknown beforehand. This leads us to leverage the active set algorithm [5], which efficiently updates and optimizes the set $\mathcal{U}$ until the optimality condition is satisfied. The specific procedures are shown in Algorithm 1. The algorithm initializes $\mathcal{U}$ as the top node in the DAG; the subproblem in Equation (11) is solved in Lines 3-13, which is elaborated in Section 5.3. The solution to the subproblem is then validated against the optimality condition of the original problem via Theorem 2. If the optimality condition is satisifed, then the algorithm is terminated; Otherwise, the NCFs that violate the optimality condition will be added to the current NCF subset $\mathcal{U}$ for the next iteration.

5.3 Solution to the Subproblem in Equation (11)

Equation (10) is simplified to a subproblem where only the NCFs with nonzero weights are retained. This enables the active set algorithm to solve the subproblem by varying the set of nonzero NCFs until the optimality condition is met.

To efficiently solve Equation (11), which is convex and Lipschitz continuous in β, we employ the mirror descent algorithm [8], which achieves a near-optimal convergence rate when the feasibility set is a simplex such as in our problem. In general, mirror descent iterations require β and α to be solved alternately. When fixing α, the gradient of the objective function in Equation (11) with respect to β is calculated and then β is updated by a descent step s. β is then fixed, and the updated α is used to solve the following problem:

$$H(\eta) = \max_{\alpha} \sum_{t,s}^{T,S} \alpha_{s,t} - \frac{1}{2} \left(\sum_{u \in \mathcal{U}} \eta_u(\beta) (\max_{g \in G} h(g,u))^{\bar{p}} \right)^{1/\bar{p}} \quad (12)$$

which is difficult to solve due to the "max" function inside the $\ell_{\hat{p}}$-norm. To remove the "max" term, an auxiliary variable ξ is introduced to transform Equation (12) to the following equivalent problem:

$$\max_{\alpha} \min_{\xi} \sum_{t,s}^{T,S} \alpha_{s,t} - \frac{1}{2} \left(\sum_{u \in \mathcal{U}} \eta_u(\beta) \sum\nolimits_g \xi_{g,u} h(g,u)^{\bar{p}} \right)^{1/\bar{p}} \quad (13)$$

where $\xi_u \in \Theta_{|G|,1}$. Thus α and ξ can be solved alternately until convergence is achieved. Specifically, when fixing ξ, solving α is similar to the $\ell_{\hat{p}}$-norm MKL problem [22] with a different feasibility set for the optimization variables. When α is updated and fixed, ξ is easily optimized by straightforward linear programming.

5.4 Theoretical Analysis

5.4.1 Optimality analysis for convergence criteria

Algorithm 1 will converge when the current candidate set of NCFs $\mathcal{U} \subseteq \mathcal{V}$ satisfies the optimality condition. To verify this, derivation of the sufficient condition of the optimality is proposed in Theorem 2 and proved in Appendix D.

Theorem 2. *Denote $(\beta_{\mathcal{U}}, \alpha_{\mathcal{U}})$ as an $\epsilon_{\mathcal{U}}$-approximate optimal solution of Equation (11) based on the current active set $\mathcal{U}$. It is then an optimal solution for Equation (10) with a duality gap less than ϵ if the following condition holds:*

$$\max_{g} \max_{u \in S(\mathcal{U})} \sum\nolimits_{v \in D(u)} \frac{h(g,v)}{(\sum_{x \in A(v) \cap D(v)} d_u)^2} \leq \left(\sum\nolimits_{u \in \mathcal{U}} \eta(\beta_{\mathcal{U}}) (\hat{h}(u))^{\bar{p}} \right)^{\frac{1}{\bar{p}}} + 2(\epsilon - \epsilon_{\mathcal{U}}) \quad (14)$$

In the proposed algorithm, the most time-consuming part is the verification of a sufficient condition of convergence because it involves the search of an exponential variable space. Due to the use of the NCF lattice in Figure 2(b) and our proposed kernel, this can be reduced to a polynomial complexity, as stated by Theorem 2 and proved in Appendix E:

Theorem 3. *The sufficient condition can be examined efficiently in polynomial time: $(\sum_s^S n_s^2) \cdot |\mathcal{U}^*| \cdot e + (\sum_s^S n_s^2) \cdot |\mathcal{U}^*|^2 \cdot e)$*

The remaining computation in Algorithm 1 primarily involves the solution of the subproblem in Equation (11). Denote $|\mathcal{U}^*|$ as the size of the final active set $\mathcal{U}^*$, Then Equation 11 is solved $O(|\mathcal{U}^*|)$ times in the worst case, which requires $log(|\mathcal{U}^*|)$ iterations. The dominant computation in each iteration is solving Equation 12, whose conservative complexity estimate is $O((\sum_s^S n_s^3) \cdot |\mathcal{U}^*|^2)$, where n_s denotes the size of the data for location s. This amounts to $O(n_{s,t}^3 \cdot S \cdot |\mathcal{U}^*|^3 log(|\mathcal{U}^*|)$. After combining this with the time complexity proved by Theorem 2, the overall computational complexity of the proposed algorithm is obained: $O(n_{s,t}^3 \cdot S \cdot |\mathcal{U}^*|^3 log(|\mathcal{U}^*| + (\sum_s^S n_s^2) \cdot |\mathcal{U}^*| \cdot e + (\sum_s^S n_s^2) \cdot |\mathcal{U}^*|^2 \cdot e)$.

6 Experiments

In this paper, the performance of the proposed model HTNL is evaluated using several synthetic datasets and real-world datasets. First, the datasets and experimental settings are introduced. Then, the effectiveness and efficiency of HTNL are evaluated against several existing methods that are the state-of-the-arts. In addition, qualitative evaluations on the selection of NCFs demonstrates the interpretability of HTNL. All the experiments were conducted on a 64-bit machine quad-core processor (i7CPU@ 3.10GHz) and 16.0GB memory.

6.1 Experiment Setup

6.1.1 Synthetic datasets

Several synthetic datasets were generated randomly. The generation procedures were as follows.

1) Generate NCFs. First, define a vocabulary V consisting of 1000 primitive features (i.e., $|V| = 1000$), which are nominal symbols denoted by distinct IDs: '# 1','# 2',.... Based on V, the "ground truth" NCFs set $\mathcal{U}^*$ was randomly

formed as follows: i) 10% elements in V were randomly selected on into $\mathcal{U}^*$ following uniform distribution; ii) then 10% of them were further selected to randomly combine with another primitive features to form NCFs with length of 2; iii) similar to the previous step, 10% of "length-2" NCFs were selected to form those with length of 3. **2) Generate NCFs' weights.** As mentioned above, each location is treated as a task. Here, $|S| = 12$ tasks were generated, where each four of them randomly formed as a group. For each sth task in group g, we randomly generated the "ground truth" NCFs weights $W_{s,\cdot} \in \mathbb{R}^{|\mathcal{U}^*|}$ as follows. We first randomly generated the "group-average" weights $\overline{W}_g \sim Gaussian(\mathbf{0}, 0.05I)$, where I is identity matrix. Then for each task in each group, we generated its NCFs weights $W_{s,\cdot} \sim Gaussian(\overline{W}_g, 0.01I)$. **2) Generate the input and output variables.** Next, for each task, we generated 200 samples, and each sample is a matrix $X_{s,t} \in \mathbb{Z}^{2\times|V|}$ whose each row is formed by randomly selecting $k \sim Poisson(8)$ elements from all the "ground truth" NCFs to be valued "1" following a Poisson distribution. Additional $k \sim Poisson(2)$ primitive features could also be selected to be assigned "1" from all the primitive features following a uniform distribution. All the other unselected features were set to 0. Furthermore, the response variable for tth sample of sth task is determined by logistic function: $Y_{s,t} = \text{sign}(\sum_i^2 [X_{s,t}]_{i,v} \cdot W_{s,v} + \varepsilon)$, where $\varepsilon \sim$ Gaussian$(0, 0.01)$. Based on the above strategy, 10 synthetic datasets were generated randomly.

6.1.2 Real-world civil unrest datasets

Table 2: Real-world datasets

Dataset	#Tweets	Label sources [1]	#Events
Argentina	160,564,890	Clarín; La Nación; Infobae	1427
Colombia	158,332,002	El Espectador; El Tiempo; El Colombiano	1287
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.	6,487,623,208	CDC Flu Activity Map	533

For the datasets on Latin America, the raw data was obtained by randomly sampling 10% (by volume) of the Twitter data from Jan 2013 to Dec 2014 in five countries as shown in Table 2. The Twitter data for the period from Jan 1, 2013 to Dec 31, 2013 was used for training, while the data for the second half of the period, from Jan 1, 2014 to Dec 31, 2014, was used for the performance evaluation. For the civil unrest domain, the feature set included $2^{100} \approx 10^{30}$ conjunctive features which were all the possible conjunctions of 100 civil unrest related words (such as "protest" and "riot") and hashtags (such as "#Megamarch") based on the keyword list in [28]. The event forecasting results were validated against a labeled events set, known as the gold standard report (GSR) publicly available[1], as shown in Table 2. An example of a labeled GSR event was given by the tuple: (CITY="Curitiba", STATE = "Paraná", COUNTRY = "Brazil", DATE = "2013-01-20").

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.

1. Open Source Indicators. https://doi.org/10.7910/DVN/EN8FUW

6.1.3 Real-world influenza dataset

For the datasets in the United States, the raw data was crawled from Jan 2013 to Dec 2014, as shown in Table 2. As in the first dataset, the Twitter data for the period from Jan 1, 2013 to Dec 31, 2013 was used for training while the second half of the period, from Jan 1, 2014 to Dec 31, 2014, was used for the performance evaluation. For the influenza outbreaks, the feature set consisted of over $2^{181} \approx 10^{54}$ features generated from the combinations of 181 influenza-related words extracted based on the keywords list used in [23]. The forecasting results for the flu outbreaks were validated against the corresponding influenza statistics reported by the Centers for Disease Control and Prevention (CDC)[2]. 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 4 ILI activity levels: minimal, low, moderate, and high, where the level "high" corresponds to a salient flu outbreak and was considered for forecasting. An example of a CDC flu outbreak event is: (STATE = "Virginia", COUNTRY = "United States", WEEK = "01-06-2013 to 01-12-2013").

6.1.4 Parameter Settings and Metrics

The event forecasting task is to predict whether or not there will be an event in the next time-step for a specific location. For civil unrest datasets, a time step is one day while for disease outbreaks, a time step is one week. There are several parameters for our proposed model HTNL. First, p (with three optional values: $\{1.1, 1.5, 1.9\}$ since $1 < p \leq 2$ and p=2 will be tested in gHKL model introduced in the following) and C (with candidate values:$\{0.01, 0.1, 1, 10, 100\}$) were determined with a 3-fold cross validation. The parameters were set as $d_v = 2^{|v|}$ suggested by Jawanpuria et al. [20]. The geographical hierarchy was "state-city" administrative relation for civil unrest datasets while "HHSregion[3]-state" relation for influenza dataset.

6.1.5 Comparison Methods

The proposed HTNL were compared with 7 state-of-the-art methods on spatial event forecasting and predictive rule learning described as follows.

1. *Least absolute shrinkage and selection operator (LASSO)* [28]. LASSO utilizes a simple ℓ_1-norm to jointly achieve curve fitting and select the primitive features. In addition to merely using the primitive as the features, we also tried frequent patterns as the features based on frequent pattern mining techniques which extract the frequent conjunctive features which appear at least in 1% among all the tweets in training set as features. The LASSO model using such features is named "LASSO-Freq" here. The feature set is the set of the primitive features, namely keyword counts. The regularization parameter is set based on a 3-fold cross validation on the training set.

2. *Tree-guided Group Lasso for Multi-task Learning (TMTL)* [21]. The relationships among tasks follow the geo-hierarchy

2. CDC FluView. http://www.cdc.gov/flu/weekly/fluviewinteractive.htm

3. HHSregions:http://www.hhs.gov/about/agencies/regional-offices/

Table 3: Evaluation results of all methods in effectiveness and efficiency on 9 datasets

	Prediction Performance Area Under the Curve (AUC) of ROC						Train. Time	Test Time
Method	Argentina	Colombia	Paraguay	Uruguay	Venezuela	Influenza	(second)	(second)
LASSO	0.5738	0.6441	0.6013	0.6526	0.5722	0.6738	40	10^{-3}
LogReg	0.7268	**0.7384**	0.7044	0.7274	0.6792	0.4851	18	10^{-3}
LASSO-Freq	0.5857	0.5560	0.5138	0.5801	0.5560	0.6854	120	10^{-3}
LogReg-Freq	0.5158	0.4905	0.5012	0.5114	0.5239	0.6554	52	10^{-3}
KDE-LDA	0.7665	0.6919	0.6654	0.7279	0.7214	0.2827	656	10^{-2}
MREF	0.7264	0.5296	0.6171	0.6812	0.5887	0.4969	444	10^{-3}
TMTL	0.7069	0.5633	0.6129	0.6931	0.6586	0.4989	203	10^{-3}
RuleFit	0.7246	0.5101	0.5008	0.5698	0.7080	0.6100	**3**	10^{-3}
gHKL	0.6850	0.5198	0.6067	0.6878	0.6970	0.5000	76	10^{-3}
HTNL	**0.8264**	**0.7384**	**0.7374**	**0.7538**	**0.7508**	**0.6951**	132	10^{-3}

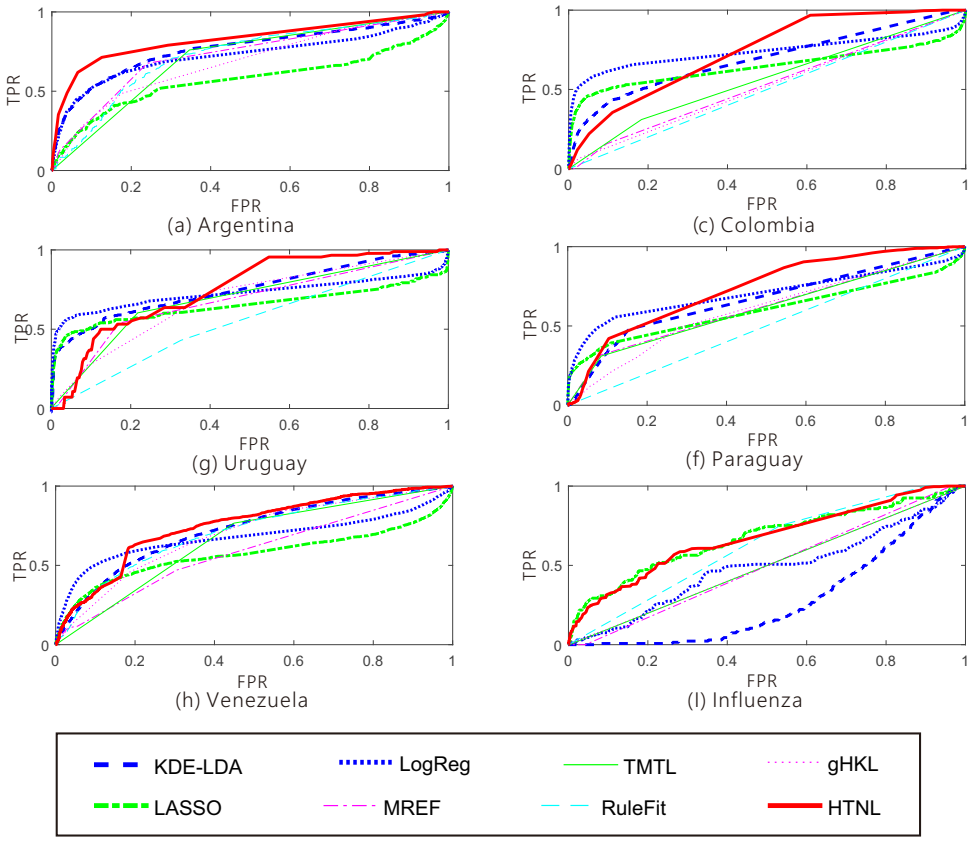


Figure 4: The AUC curves of our HTNL and all the comparison methods. HTNL performed consistently best in general.

Table 4: The NCF precursors (translated in English) discovered for different datasets. (# NCF: The number of selected NCFs; avg. len.: The average length of all the NCFs; The symbol "+" denotes the logical "and" within an NCF)

	Argentina	Colombia	Paraguay	Uruguay	Venezuela	Influenza
Top 10 NCFs (length$\geq$ 2)	government+congress	government+water	national+university	know+hope	government+high	bed+flu+home
	government+deputies	water+problem	to+avoid	security+rights	violence+order	bed+home
	to+end+hate	violence+protest	war+avoid	project+president	national+support	cold+sick
	let's+fight	national+mayor	know+rights	power+death	candidate+march	you+flu
	followers+free	national+control	work+rights	power+matches	national+students	is+epidemic
	to+know+results	mayor+control	find+food	president+hope	patria+control	flu+bed
	protest+against	national+government	people+commitment	death+matches	national+government	sick+stomach
	national+triumph	national+water	national+freedom	project+hope	fight+policy	yousick
	hate+hunger	national+freedom	national+marches	national+university	students+protest	have+today
	hate+class	power+death	national+central	national+fight	government+violence	not+well
# NCFs	131	106	108	71	325	2017
avg. len.	1.2290	1.2837	1.1062	1.0000	1.6892	1.9499

defined by the administrative relation introduced in Section 6.1.4. Keyword counts are the features. The regularization parameter $\lambda = 0.3$ are set based on a 3-fold cross-validation.

3. *Logistic regression (LR)* [10]. LR utilizes a logit function to map the tweets observations into future event occurrences ("-1" denotes no occurrence, "1" denotes occurrence). The input features here are the counts of keywords. In addition to merely using the primitive as the features, we also tried frequent patterns as the features based on frequent pattern mining techniques which extract the frequent conjunctive features which appear at least in 1% among all the tweets in training set as features. The LR model using such

features is named "LR-Freq" here.

4. *Kernel density estimation-based logistic regression (KDE-LR)* [17]. 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 3-fold cross-validation.

5. *Multi-resolution Event Forecasting (MREF)* [40]. This method jointly models the prediction tasks in multiple geographical levels by utilizing their geo-hierarchical relation. The features are the primitive features, namely the counts of keywords. The major parameter is the regularization parameter that is set by 3-fold cross-validation.

6. *RuleFit* [15]. RuleFit is a well-recognized rule ensemble learning algorithm. All the parameters were set to the default values mentioned by the authors and recommended by several publications [19], [12], [13]. To be specific, the model was set in the mixed linear-rule mode, average size of tree was set 4, and the maximum number of trees were set as 500. This method is used to learn the predictive rules for future events. The model inputs are binary occurrence of keywords combinations, as it can only handle Boolean rules.

7. *Generalized hierarchical kernel learning (gHKL)* [19]. gHKL is a state-of-the-art rule ensemble learning algorithm. There are two sensitive parameters, the type of norm and the value of regularization parameter. We have tried 1.1-, 1.5-, and 1.9-norms, while tuned the with regularization parameter with values in $\{10^{-3}, 10^{-2}, \cdots, 10\}$ using 3-fold cross validation.

6.2 Performance

In this section, the proposed HTNL is evaluated quantitatively in effectiveness, efficiency, and scalability on synthetic and real-world datasets. In addition, the illustration of the selected NCFs by the proposed HTNL is also presented.

6.2.1 Quality of NCFs selection

The quality of the selected NCFs by the methods for conjunctive feature learning, namely RuleFit, gHKL, and our HTNL were evaluated on synthetic datasets against the "ground-truth" NCFs. Precision and recall were utilized to evaluate the percentage of correct NCFs among all the selected NCFs, and the percentage ground-truth NCFs successfully selected, respectively. F-measure, which is the harmonic mean of precision and recall, is also shown in Table 5. In general, the proposed HTNL outperforms the comparison methods by 15% in F-measure, and also achieved the best recall and second-best precision. Moreover, HTNL always achieved much smaller standard deviations than the comparison methods, showing a robust performance across different datasets. Both precision and recall are near 90% of HTNL, demonstrating that it can effectively discover the ground-truth conjunctive features just based on the training data. Moreover, both HTNL and gHKL can effectively consider both the difference and similarity among different tasks, this explains why they outperform RuleFit. Furthermore, HTNL can further consider the grouping relationship among different tasks, which helps it further outperform gHKL.

Table 5: Conjunctive feature selection performance in the precision, recall, and F-measure on 10 synthetic datasets.

Method	Precision	Recall	F-measure
RuleFit	**0.908**±0.064	0.568±0.116	0.688±0.083
gHKL	0.730±0.095	0.936±0.018	0.817±0.047
HTNL	0.869±**0.051**	**0.937±0.015**	**0.901±0.027**

6.2.2 Performance on real-world datasets

In the experiment based on real-world datasets, Twitter data collection was partitioned into a sequence of date-interval subcollections. The event forecasting task was to utilize one day tweet data to predict whether or not there would be an event in the next day for a specific city (for the civil unrest domain), or a specific state (for the influenza outbreaks domain), which means the lead time $q = 1$. To perform this task, we created a training set and a test set for each city (or state), where each data sample was the daily tweet observation with the above-mentioned features. The predicted events were structured as tuples of (date, city/state). A predicted event was matched to a real event if both the date and location attributes were matched. To validate the prediction performance, the Area Under the Curve (AUC) of Receiver operating characteristic (ROC) curve were adopted. ROC curve illustrates the performance of a binary classifier as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. The AUC measures the area below this curve, which is a well-recognized metric for the comprehensive performance of a classifier.

Table 3 summarizes the effectiveness and efficiency of the proposed HTNL on different datasets. The AUC measure was adopted to quantify the performance. First, the results shown in Table 3 demonstrates that the methods that take into account the spatial information, especially the geographical hierarchy, performed better. Specifically, KDE-LDA, MREF,TMTL, and the proposed HTNL typically performed the best in most situations. KDE-LDA performed much better on civil unrest datasets than the influenza dataset. This might be because this method was specially designed to forecast crimes, which are small-scale social events unlike influenza epidemics in "state" level. LogReg also achieved a very competitive performance with AUC larger than 0.73 on three datasets. Second, HTNL outperformed all the other methods in all the six datasets. This is because HTNL not only considers the geo-hierarchy, but more importantly, is to consider the NCFs like the frequencies of keyword co-occurrences as new features that capture crucial precursors for future events. In contrast, the Boolean rule learning methods including RuleFit and gHKL only achieved the AUCs around 0.6 on the datasets, generally worse than the other methods. This is because they can only consider the binary occurrence of keywords on each date instead of the frequencies of keywords. Thus they lose much information of the magnitude of the social indicators. Among all the datasets, the overall performance for Argentina was generally the best while the Influenza outbreaks forecasting was a relatively difficult prediction task with lower AUCs for most of the methods. Finally, the method directly utilizing frequent pattern mining strategy to treat the most frequent conjunctive features as the features cannot

achieve a competitive performance. Specifically, LogReg-Freq achieved a worse performance than that of LogReg because there are much more features (i.e., frequent patterns as conjunctive feautres) involved purely based on the input of training set, which are likely to enforce the over-fitting somehow due to the largely-increased number of features against the number of training samples. LASSO-Freq performs clearly better than LogReg-Freq because it can regularize the large number of features involved in LogReg, and hence achieved better generalizability. Both of LASSO-Freq and LogReg-Freq cannot perform as good as our method because the frequent patterns selected based on their method can only consider the frequency of the features. But more frequent does not necessarily mean more important and beyond merely inputs, our method can jointly consider both inputs and outputs and hence is able to learn which conjunctive feature inputs are important for prediction outputs.

6.2.3 Efficiency and scalability

The rightmost column of Table 3 shows the training and test time efficiency comparison among HTNL and the competing methods for forecasting influenza outbreaks. The efficiency evaluation results on civil unrest datasets followed a similar pattern and are not provided due to space limitations. The test runtime for all the methods are extremely small (no more than 0.01 sec) for each prediction, though KDE-LDA is relatively slower (i.e., 10^{-2}) due to extra computation required for computing for the latent topics. For the training runtime, Table 3 shows that RuleFit required smallest amount of time of only 3 seconds, because of two reasons 1) it binarizes the numerical frequencies into Boolean values as inputs; and 2) it utilizes an efficient heuristic procedure to obtain a suboptimal solution. Simpler methods like LASSO and LogReg also achieved high efficiency with less than 50 seconds. In addition, even though HTNL need optimize a problem with exponentially large set of candidate features, it still achieved highly efficient computation. This is because of the good property of the proposed kernel in Lemma 1, which is proved to reduce the exponential time complexity down to polynomial as proved in Theorem 3.

In addition, Figure 5 illustrates the scalability of the proposed HTNL in synthetic datasets in the runtime when the size of the datasets vary. Each setting of synthetic datasets was generated randomly for ten times and thus the standard deviation was calculated and shown by the error bars. Specifically, Figure 5(a) shows that when the number of features in active set does not change, the runtime basically will not increase because only the active set is essentially involved in subproblem computation in Equation (11). This demonstrated the theoretically-advantageous and practically-useful characteristics of HTNL which can handle large number of sparse features where only few of them are useful for the prediction tasks. Furthermore, Figure 5(c) shows that when the features in active set increases, the runtime generally increases super-linearly, which verifies the time complexity proved in Theorem 3. In addition, the runtime is linear in the number of tasks while super-linear in the number of total samples, which again match Theorem 3.

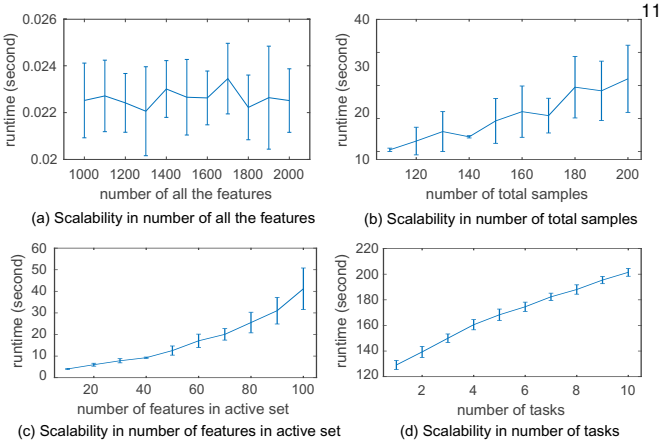


Figure 5: The scalability of the proposed method.

6.2.4 Event forecasting performance on ROC curves

In Figure 4, the event forecasting performance in ROC curves for six datasets is illustrated. For all these datasets, the proposed HTNL performed consistently among the best whose curves were farthest away from the point (1,0). Specifically, For the the civil unrest datasets like "Argentina", "Paraguay", "Uruguay", and "Venezuela", HTNL generally performed the best, with ROC curves covering the largest areas above the x-axis. For the dataset of "Colombia", HTNL, KDE-LDA, and LogReg were among the best, where HTNL typically performed the best when FPR was larger than 0.5. This merit is important because it indicates that HTNL tends to provide the most extensive true positive alarms among all the methods. Comprehensive detections of sensitive social events are important to many applications such as social emergency management. For the influenza dataset, according to Figure 4(i), HTNL consistently outperformed the other methods with different FPR and TPR values. LogReg and LASSO also achieved quite competitive performance.

6.2.5 Qualitative Evaluation

Another advantage of our HTNL is its strong interpretability compared to most of the spatial event forecasting models that merely use primitive features. Table 4 shows the results on the selection of NCFs by the proposed HTNL for the six real-world datasets. Specifically, the top 10 NCFs with length larger than 1, namely precursor rules, are listed. The amount and average length of the selected NCFs are also presented. The original Spanish words were translated into English by Google Translator [4]. The symbol "+" denotes the logical "and" within an NCF. An NCF will be triggered only if all its words connected by "+" co-occur in a tweet. According to Table 4, the proposed HTNL effectively selected high-quality NCFs robustly for all the datasets in two different domains, namely civil unrest and influenza outbreaks. For civil unrest datasets, the NCFs in Table 4 typically represent the motivations or propaganda of the protest events. For example, the high frequency of the NCF "water"+"problem" could probably be one important reason that causes social unrest in Colombia, while the NCFs like "government"+"congress" and "government"+"deputies" could be the triggers for those future events in Argentina. The NCFs can also be propaganda-related, such as "to"+"end"+"hate"

4. Google Translate: https://translate.google.com/

and "let's"+"fight". In contrast to civil unrest datasets, those top NCFs for influenza dataset were typically not about the motivation or advertisement of organized social events, but the symptoms or discussions about flu. For example, NCF like "bed"+"flu"+"home" appearing together in a tweet was likely to be a strong indicator for a person's disease status. "sick"+"stomach" could also be a symptom of stomach flu. Additionally, the numbers of total NCFs optimally selected for different datasets show that for larger countries, the numbers and average lengths of NCFs tend to be larger. This is because larger size of population typically leads to more various social issues and thus the social events could be indicated by more diverse precursors. Finally, the average length of the selected features for Uruguay dataset is 1, which indicates that the selected conjunctive features are all single keyword features. The reason is because Uruguay has relatively small number of tweets and fewer tweets per event on average, as shown in Table 2. This makes this dataset to have small number of existence of conjunctive features for each sample. This indicates that for this country, conjunctive features with size larger than one are not strong signals to indicate future events because of the scarcity of them..

7 Conclusion

Forecasting spatial societal events in social media is significant. It is also very challenging because the precursors of the future events are not straightforward and can be sophisticated ensemble of underlying rules. Most existing methods simplified this problem by considering frequencies of keywords or predefined phrases as features due to the challenges such as the inherent exponential complexity of ensemble rule learning, distant supervision, numerical values, geographical relations. To jointly handle all the challenges with theoretical guarantee, we propose a novel spatial event forecasting model named HTNL which learns the NCFs efficiently. an efficient algorithm is proposed to optimize the model parameters and prove its theoretical guarantees for error bound and time efficiency. Extensive experiments on multiple datasets demonstrate the effectiveness and efficiency of the proposed method. Moreover, qualitative analysis on the extracted NCFs explicitly shows the strong interpretability of HTNL.

Acknowledgement

This work was supported by the National Science Foundation grant: # 1755850.

References

[1] Ayan Acharya, Raymond J Mooney, and Joydeep Ghosh. Active multitask learning using supervised and shared latent topics. *Pattern Recognition and Big Data*, page 75, 2016.

[2] Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. Predicting flu trends using Twitter data. In *IEEE Conference on Computer Communications Workshops*, pages 702–707, 2011.

[3] Bilal Ahmed, Thomas Thesen, Karen Blackmon, Ruben Kuzniecky, Orrin Devinsky, Jennifer Dy, and Carla Brodley. Multi-task learning with weak class labels: Leveraging ieeg to detect cortical lesions in cryptogenic epilepsy. In *Machine Learning for Healthcare Conference*, pages 115–133, 2016.

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

[5] Francis Bach. High-dimensional non-linear variable selection through hierarchical kernel learning. *arXiv preprint arXiv:0909.0844*, 2009.

[6] Christopher M Bishop. Pattern recognition and machine learning (information science and statistics). 2006.

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

[8] Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8(3-4):231–357, 2015.

[9] Hong Cheng, Xifeng Yan, Jiawei Han, and Chih-Wei Hsu. Discriminative frequent pattern analysis for effective classification. In *ICDE 2007*, pages 716–725. IEEE, 2007.

[10] Ryan Compton, Craig Lee, Jiejun Xu, Luis Artieda-Moncada, Tsai-Ching Lu, Lalindra De Silva, and Michael Macy. Using publicly visible social media to build detailed forecasts of civil unrest. *Security informatics*, 3(1):4, 2014.

[11] Gao Cong, Kian-Lee Tan, Anthony KH Tung, and Xin Xu. Mining top-k covering rule groups for gene expression data. In *SIGMOD 2005*, pages 670–681. ACM, 2005.

[12] Krzysztof Dembczyński, Wojciech Kotłowski, and Roman Słowiński. Maximum likelihood rule ensembles. In *ICML 2008*, pages 224–231. ACM, 2008.

[13] Krzysztof Dembczyński, Wojciech Kotłowski, and Roman Słowiński. Ender: a statistical framework for boosting decision rules. *Data Mining and Knowledge Discovery*, 21(1):52–90, 2010.

[14] Xiaowen Dong, Dimitrios Mavroeidis, Francesco Calabrese, and Pascal Frossard. Multiscale event detection in social media. *Data Mining and Knowledge Discovery*, 29(5):1374–1405, 2015.

[15] Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, pages 916–954, 2008.

[16] Thomas Gärtner, Peter A Flach, Adam Kowalczyk, and Alexander J Smola. Multi-instance kernels. In *ICML 2002*, volume 2, pages 179–186, 2002.

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

[18] Mehmet Gönen and Ethem Alpaydın. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12(Jul):2211–2268, 2011.

[19] Pratik Jawanpuria, Saketha N Jagarlapudi, and Ganesh Ramakrishnan. Efficient rule ensemble learning using hierarchical kernels. In *ICML 2011*, pages 161–168, 2011.

[20] Pratik Jawanpuria, Jagarlapudi Saketha Nath, and Ganesh Ramakrishnan. Generalized hierarchical kernel learning. *The Journal of Machine Learning Research*, 16(1):617–652, 2015.

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

[22] Marius Kloft, Ulf Brefeld, Sören Sonnenburg, and Alexander Zien. Lp-norm multiple kernel learning. *Journal of Machine Learning Research*, 12(Mar):953–997, 2011.

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

[24] K. Lin and J. Zhou. Interactive multi-task relationship learning. In *ICDM 2016*, pages 241–250, 2016.

[25] Dmitry M Malioutov and Kush R Varshney. Exact rule learning via boolean compressed sensing. In *ICML 2013*, pages 765–773, 2013.

[26] Sara Melvin, Wenchao Yu, Peng Ju, Sean Young, and Wei Wang. Event detection and summarization using phrase network. In *ECML-PKDD 2017*, pages 89–101. Springer, 2017.

[27] Charles A Micchelli and Massimiliano Pontil. Learning the kernel function via regularization. *Journal of machine learning research*, 6(Jul):1099–1125, 2005.

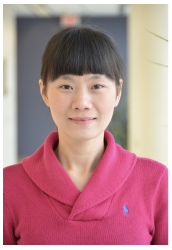
[28] Naren Ramakrishnan, Patrick Butler, et al. Beating the news with EMBERS: forecasting civil unrest using open source indicators. In *KDD 2014*, pages 1799–1808. ACM, 2014.

[29] Ulrich Rückert and Stefan Kramer. A statistical approach to rule learning. In *ICML 2006*, pages 785–792. ACM, 2006.

[30] Takeshi Sakaki, Makoto Okazaki, and Yutaka Matsuo. Earthquake shakes Twitter users: real-time event detection by social sensors. In *WWW 2010*, pages 851–860, 2010.
[31] Erich Schubert, Michael Weiler, and Hans-Peter Kriegel. Signitrend: scalable detection of emerging topics in textual streams by hashed significance thresholds. In *KDD 2014*, pages 871–880. ACM, 2014.
[32] Bilong Shen, Xiaodan Liang, Yufeng Ouyang, Miaofeng Liu, Weimin Zheng, and Kathleen M. Carley. Stepdeep: A novel spatial-temporal mobility event prediction framework based on deep neural network. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, pages 724–733, New York, NY, USA, 2018. ACM.
[33] Gyorgy J Simon, Vipin Kumar, and Peter W Li. A simple statistical model and association rule filtering for classification. In *KDD 2011*, pages 823–831. ACM, 2011.
[34] C. Su, F. Yang, S. Zhang, Q. Tian, L. S. Davis, and W. Gao. Multi-task learning with low rank attribute embedding for multi-camera person re-identification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(5):1167–1181, May 2018.
[35] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. *ICWSM 2010*, 10:178–185, 2010.
[36] Xiaofeng Wang, Matthew S Gerber, and Donald E Brown. Automatic crime prediction using events extracted from Twitter posts. In *SBP 2012*, pages 231–238. Springer, 2012.
[37] Neeraja J Yadwadkar, Bharath Hariharan, Joseph E Gonzalez, and Randy Katz. Multi-task learning for straggler avoiding predictive job scheduling. *Journal of Machine Learning Research*, 17(106):1–37, 2016.
[38] Chao Zhang, Liyuan Liu, Dongming Lei, Quan Yuan, Honglei Zhuang, Tim Hanratty, and Jiawei Han. Triovecevent: Embedding-based online local event detection in geo-tagged tweet streams. In *KDD 2017*, pages 595–604. ACM, 2017.
[39] Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Spatiotemporal event forecasting in social media. In *SDM 2015*, pages 963–971. SIAM, 2015.
[40] Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Multi-resolution spatial event forecasting in social media. In *ICDM 2016*, pages 689–698, 2016.

Liang Zhao is an Assistant Professor at Information Science and Technology Department of George Mason University. He received the Ph.D. degree from Virginia Tech, 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.

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 USA, in 2012, all in computer science. He is an Assistant Professor with the University at Albany, SUNY. 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.



Yanfang Ye Yanfang Ye received the B.Eng. and M.Sc. degrees in computer science and technology from Fuzhou University, Fuzhou, China, in 2003 and 2006, respectively, and the Ph.D. degree from Xiamen University, Xiamen, China, in 2010. She is currently an Assistant Professor with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, WV, USA. Her current research interests include data mining, machine learning, cyber security, and smart devices.

APPENDIX A
PROOF OF LEMMA 1

Proof. Define $k_{v,i,j}(X_{s,t}, X_{s',t'}) = \phi_{v,i}(X_{s,t}) \cdot \phi_{v,j}(X_{s,t})$, therefore we have $k_v(X_{s,t}, X_{s,t}) = \sum_{i,j} k_{v,i,j}(X_{s,t}, X_{s',t'})$. In the following, we will first prove $k_{v,i,j}(X_{s,t}, X_{s',t'})$ is a kernel.

$$\begin{aligned} k_{v,i,j}(X_{s,t}, X_{s',t'}) &= \phi_{v,i}(X_{s,t}) \cdot \phi_{v,j}(X_{s,t}) \\ &= \prod\nolimits_{k\in v} [X_{s,t}]_{i,k} \cdot \prod\nolimits_{k\in v} [X_{s,t}]_{j,k} \\ &= \prod\nolimits_{k\in v} \langle [X_{s,t}]_{i,k}, [X_{s',t'}]_{j,k} \rangle \end{aligned} \tag{15}$$

According to the properties of kernel, because each $\langle [X_{s,t}]_{i,k}, [X_{s',t'}]_{j,k} \rangle$ is a kernel on each feature dimension, the multiplication of kernels on all the dimensions $k_{v,i,j}(X_{s,t}, X_{s',t'})$ must be a kernel.

Finally, according to the theory of convention kernels [6], $k_v(X_{s,t}, X_{s,t})$ is a kernel if and only if each $k_{v,i,j}(X_{s,t}, X_{s',t'})$ is a kernel. The proof is completed. □

APPENDIX B
PROOF OF LEMMA 2

Proof. According to Micchelli and Pontil [27], we have the following equality holds based on the Holder inequality.

$$\min_{b\in\Theta_{d,r}} \sum\nolimits_{i=1}^{d} a_i/b_i = \left(\sum\nolimits_{i=1}^{d} a_i^{r/(r+1)}\right)^{(r+1)/r} \tag{16}$$

where $\Theta_{d,r} = \{x \in \mathbb{R}^d | x \geq 0, \sum_i x_i^r \leq 1\}$. By repeatedly applying Equation (16) on Equation (6), we have the following equation:

$$\begin{aligned} \Omega(W)^2 &= (\sum\nolimits_{v\in\mathcal{V}} d_v \left(\sum\nolimits_{u\in D(v)} \left(\sum\nolimits_{g}^{G} \|\{W_{j,u}\}_j^g\|_2\right)^p\right)^{\frac{1}{p}})^2 \\ &= \min_{\gamma} \sum\nolimits_{v\in\mathcal{V}} \frac{d_v^2}{\gamma_v} \left(\sum\nolimits_{u\in D(v)} \left(\sum\nolimits_{T}^{\mathcal{T}} \|W_u\|_2\right)^p\right)^{(\frac{1}{p})} \\ &= \min_{\gamma,\lambda} \sum\nolimits_{v\in\mathcal{V}} \frac{d_v^2}{\gamma_v} \sum\nolimits_{u\in D(v)} \frac{1}{\lambda_{v,u}} \left(\sum\nolimits_{T}^{\mathcal{T}} \|W_u\|_2\right)^2 \\ &= \min_{\gamma,\lambda,\mu} \sum\nolimits_{v\in\mathcal{V}} \frac{d_v^2}{\gamma_v} \sum_{u\in D(v)} \frac{1}{\lambda_{v,u}} \sum_{g}^{G} \frac{1}{\mu_{u,v,g}} \sum_{j}^{g} \|W_{j,u}\|^2 \end{aligned}$$

where $\gamma \in \Theta_{|\mathcal{V}|,1}$, $\lambda_v = \Theta_{D(v),\hat{p}}$, and $\mu_{u,v,g} = \Theta_{G,1}$. $\hat{p} = p/(2-p)$ □

APPENDIX C
PROOF OF THEOREM 1

In order to prove Theorem 1, we need to first prove the following lemma:

Lemma 3. *The following two problems are equivalent.*

$$\max_{\gamma,\lambda,\mu} \left(\sum_{u\in\mathcal{V}} \sum_{g}^{G} \Psi_{v,u,g}(\gamma,\lambda,\mu) \cdot h(g,u)\right)^{\bar{p}} \tag{17}$$

$$\min_{\delta\in\Delta} \min_{a} a, \quad s.t. a \geq d_v^{-2\bar{p}} \left(\sum_{u\in D(v)} \delta_{v,u}^2 \cdot \hat{h}(u)\right)^{\bar{p}} \tag{18}$$

Proof. According to the Proposition 11 in Jawanpuria et al. [19], Equation (17) is equivalent to the following:

$$\max_{\gamma,\mu} \min_{\delta\in\Delta} \max_{x\in\mathcal{V}} \left(\sum_{u\in D(x)} \sum_{g}^{G} \frac{\delta_{x,u}^2 \lambda_{x,u} \mu_{g,u} h(g,u)}{d_u^2}\right)^{\bar{p}} \tag{19}$$

By applying Chebyshev approximation, Equation (19) is equivalent to:

$$\max_{\lambda}\min_{\mu}\max_{\delta\in\Delta}\max_{a} a,\quad s.t.,\ a \geq \sum_{u}^{D(v)}\sum_{g}^{G}\left(\frac{\delta_{v,u}^2\lambda_{v,u}\mu_{g,u}h(g,u)}{d_v^2}\right)^{\bar{p}}$$

The above is equivalent to the following according to the Sion-Kakutani minimax theorem.

$$\max_{\delta\in\Delta}\max_{a}\max_{\lambda}\min_{\mu} a,\quad s.t.,\ a \geq \sum_{u}^{D(v)}\sum_{g}^{G}\left(\frac{\delta_{v,u}^2\lambda_{v,u}\mu_{g,u}h(g,u)}{d_v^2}\right)^{\bar{p}}$$

Using the Holder's inequality for λ and μ, we obtain the following:

$$\min_{\delta\in\Delta}\min_{a} a \quad , s.t.\ \ a \geq (d_v^{-2}(\sum\nolimits_{u\in D(v)}(\delta_{v,u}^2\hat{h}(u))^{\bar{p}})^{1/\bar{p}})^{\bar{p}}$$

where $\hat{h}(u) = \|\{h(g,u)\}_g^G\|_\infty$. The proof is completed. □

Now we utilize Lemma 3 to prove Theorem 1:

Proof. The proof of the Theorem 1 is equal to proving that Equation (17) is equivalent to $\max_{\beta\in\Theta_{|\mathcal{V}|,1}}\sum_{u\in\mathcal{V}}(\hat{h}(u))^{\bar{p}}\eta_u(\beta)$, which is now proved in the following. The Lagrangian of Equation (18) is:

$$L(\delta,a,\beta) = a + \sum_{v\in\mathcal{V}}\beta_v(d_v^{-2\bar{p}}\sum_{u\in D(v)}(\delta_{v,u}^2\hat{h}(u))^{\bar{p}} - a) \tag{20}$$

Let the derivative of L with respect to a to be 0, we obtain:

$$\max_{\beta\in\Theta_{|\mathcal{V}|,1}}\min_{\delta\in\Delta}\sum_{u\in\mathcal{V}}\sum_{u\in D(v)}\beta_v(d_v^{-2}\delta_{v,u}^2\hat{h}(u))^{\bar{p}} \tag{21}$$

which can be transformed to the following:

$$\max_{\beta\in\Theta_{|\mathcal{V}|,1}}\sum_{u\in\mathcal{V}}\hat{h}(u)^{\bar{p}}\min_{\delta\in\Delta}\sum_{u\in D(v)}\beta_v(d_v^{-2}\delta_{v,u}^2)^{\bar{p}} \tag{22}$$

whose equivalence is obtained using the Holder's inequality.

$$\max_{\beta\in\Theta_{|\mathcal{V}|,1}}\sum_{u\in\mathcal{V}}\hat{h}(u)^{\bar{p}}\left(\sum_{v\in A(u)}d_v^p\beta_v^{1-p}\right)^{1/(1-p)} \tag{23}$$

which is equivalent to $\max_{\beta\in\Theta_{|\mathcal{V}|,1}}\sum_{u\in\mathcal{V}}(\hat{h}(u))^{\bar{p}}\eta_u(\beta)$. The proof is completed. □

Appendix D
Proof of Theorem 2

Proof. Consider the following function of α and β.

$$F(\alpha,\beta) = \sum_{t\in T,s\in S}\alpha_i - \frac{1}{2}\sum_{u\in\mathcal{V}}\eta_u(\beta)\cdot\hat{h}(u) \tag{24}$$

which is concave in α and convex in β. According to Sion-Kakutani minimax theorem, there is no duality gap between the following min-max interchange problems:

$$\min_{\beta}\max_{\alpha}F(\alpha,\beta) = \max_{\alpha}\min_{\beta}F(\alpha,\beta) \tag{25}$$

Therefore, the error of the current solution (α,β) to the global optima is bounded by the duality gap below:

$$\max_{\alpha'}F(\alpha',\beta) - \min_{\beta'}F(\alpha,\beta') \leq P_\beta^* - \min_{\beta'}F(\alpha,\beta') \leq P_\beta - D_\beta + e(\beta)$$

where P_β is the value of the primal objective (in Equation 8) while P_β^* is its optimal value. D_β is the value of the dual objective. $e(\beta) = F(\alpha,\beta^*) - F(\alpha,\beta) \geq 0$ is the error based on the solution β.

Therefore to ensure the duality gap is less than ϵ, we need:

$$F(\alpha,\beta^*) - F(\alpha,\beta) + P_\beta - D_\beta \leq \epsilon \tag{26}$$

which is equivalent to

$$\max_{\beta}\left(\sum_{u\in\mathcal{V}}\eta_u(\beta)\cdot\hat{h}(u)\right)^{\frac{1}{\bar{p}}} - \left(\sum_{u\in\mathcal{V}}\eta_u(\beta)\cdot\hat{h}(u)\right)^{\frac{1}{\bar{p}}} + P_\beta - D_\beta \leq \epsilon \tag{27}$$

According to Proposition 5 in [5], we have,

$$\max_{\beta}\left(\sum_{u\in\mathcal{V}}\eta_u(\beta)\cdot\hat{h}(u)\right)^{\frac{1}{\bar{p}}} \leq \max\{\Omega(W_\mathcal{U})^2, \max_{u\in S(\mathcal{U})}\sum_{v\in D(u)}\frac{h(g,v)}{(\sum_{x\in A(v)\cap D(v)}d_u)^2}\} \tag{28}$$

Combine Equation (27) and (28), and consider the fact that $P_\beta - D_\beta \geq 0$, the proof is completed. □

Appendix E
Proof of Theorem 3

Before proving Theorem 2, we first present the following lemma:

Lemma 4.

$$\begin{aligned}\phi_x(I)\cdot\phi_x(J) &= \left(\sum\nolimits_{i\in I}\prod\nolimits_{c\in V}x_{I,i}\right)\cdot\left(\sum\nolimits_{j\in J}\prod\nolimits_{c\in V}x_{J,j}\right)\\ &= \sum\nolimits_{i\in I,j\in J}\left(\prod\nolimits_{c\in V}x_{I,i}\cdot\prod\nolimits_{c\in V}x_{J,j}\right)\end{aligned}$$

Now we give the proof of Theorem 3:

Proof. Then the time complexity of the computation of the kernel matrix can be reduced from exponential to polynomial, which will be proved in the following.

$$\begin{aligned}&\sum_{v\in D(u)}\frac{h(g,v)}{(\sum_{x\in A(v)\cap D(v)}d_u)^2}\\ =&\sum_{v\in D(u)}\frac{\sum_s^g\sum_{t_1,t_2\in T}\alpha_{s,t_1}y_{s,t_1}\phi_v(X_{s,t_1})\phi_v(X_{s,t_2})y_{s,t_2}\alpha_{s,t_2}}{(\sum_{x\in A(v)\cap D(u)}d_x)^2}\\ =&\sum\nolimits_s^g\sum\nolimits_{t_1,t_2\in T}m_{s,t_1,t_2}\sum_{v\in D(u)}\frac{\phi_v(X_{s,t_1})\phi_v(X_{s,t_2})}{(\sum_{x\in A(v)\cap D(u)}d_x)^2}\end{aligned}$$

where $m_{s,t_1,t_2} = \alpha_{s,t_1}y_{s,t_1}y_{s,t_2}\alpha_{s,t_2}$. By applying Lemma 4, the above equation can be transformed to:

$$\begin{aligned}&\sum_s^g\sum_{t_1,t_2\in T}m_{s,t_1,t_2}\sum_{v\in D(u)}\frac{\sum_{i\in n(t_1),j\in n(t_2)}(\prod_{c\in V}X_{t_1,i}\prod_{c\in V}X_{t_2,j})}{(\sum_{x\in A(v)\cap D(u)}d_x)^2}\\ =&\sum_s^g\sum_{t_1,t_2\in T}m_{s,t_1,t_2}\sum_{i\in n(t_1),j\in n(t_2)}\sum_{v\in D(u)}\frac{(\prod_{c\in V}X_{t_1,i}\prod_{c\in V}X_{t_2,j})}{(\sum_{x\in A(v)\cap D(u)}d_x)^2}\\ =&\sum_s^g\sum_{t_1,t_2\in T}m_{s,t_1,t_2}\sum_{i\in n(t_1),j\in n(t_2)}\frac{\prod_{c\in S_u}K_{c,i,j}}{a^{2|S_u|}}\sum_{v\in D(u)}\frac{\prod_{c\in(S_v-S_u)}k_{c,i,j}}{(a+1)^2|S_v-S_u|}\\ =&\sum_s^g\sum_{t_1,t_2\in T}m_{s,t_1,t_2}\sum_{i\in n(t_1),j\in n(t_2)}\frac{\prod_{c\in S_u}K_{c,i,j}}{a^{2|S_u|}}\prod_{c\in B-S_u}\left(\frac{K_{c,i,j}}{(a+1)^2}+1\right)\end{aligned}$$

Therefore, the computation of the kernel matrix requires time in $O(\sum_s^S n_s^2)\cdot|\mathcal{U}^*|\cdot e)$. Besides, the calculation of $m_{s,t_1,t_2}, s\in S, t_1,t_2\in T$ and right-hand side of Equation (14) requires $O(\sum_s^S n_s^2)\cdot|\mathcal{U}^*|^2\cdot e)$. The proof is completed. □