STAPLE: Spatio-Temporal Precursor Learning for Event Forecasting

Yue Ning* Rongrong Tao* Chandan K. Reddy*
Huzefa Rangwala† James C. Starz‡ Naren Ramakrishnan*

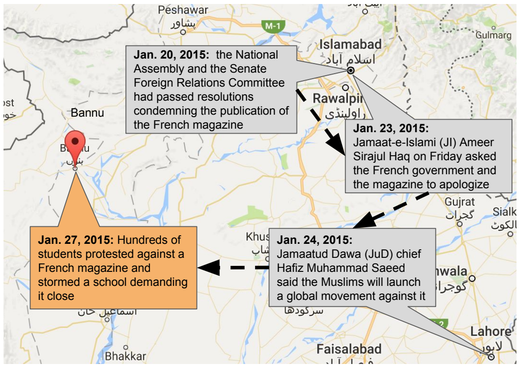
Abstract

Large-scale societal events such as civil unrest movements occur due to a variety of factors including economics, politics, and security. Societal event detection can be modeled as a system of inter-connected locations, where each location is recording a set of time-dependent observations. In order to detect event occurrence and automatically reconstruct the precursors and signals, it is essential to model relationships between the different locations w.r.t. how events evolve over time. However, existing methods for precursor discovery do not capture or exploit spatial and temporal correlations inherent in event occurrences. The absence of such modeling not only creates shortcomings in the quality of inference but also curtails interpretation by human analysts. Furthermore, forecasting is inhibited when training data is sparse. In this paper, we develop a novel multi-task model with dynamic graph constraints within a multi-instance learning framework. Our model tackles the problem of scarce data distribution and reinforces co-occurring location-specific precursors with augmented representations. Through studies on civil unrest movements in numerous countries, we demonstrate the effectiveness of the proposed method for precursor discovery and event forecasting.

Keywords: Multi-task learning; spatio-temporal precursor learning; event correlation.

1 Introduction

While studying large scale societal events, policy makers and professionals aim to reconstruct precursors to such events to help understand causative attributes. Given a document reporting an event of interest (e.g., a civil protest), precursors are other documents published earlier than reported incidents or happenings and can be viewed as influential in the lead up to the protest. Such analysis is typically done painstakingly with the aid of subject matter experts, but new algorithmic tools [13] have recently emerged that support such precursor discovery. A key challenge that persists is the incorporation of spatial and temporal correlations inherent in large-scale societal event occurrences. For example, civil unrest (protests or strikes) in a city is often influenced by happenings at nearby locations and while event counts might not be comparable, there is significant temporal and spatial correlation across event occurrences.



Jan. 20, 2015: the National Assembly and the Senate Foreign Relations Committee had passed resolutions condemning the publication of the French magazine

Jan. 23, 2015: Jamaat-e-Islami (JI) Ameer Sirajul Haq on Friday asked the French government and the magazine to apologize

Jan. 24, 2015: Jamaatud Dawa (JuD) chief Hafiz Muhammad Saeed said the Muslims will launch a global movement against it

Jan. 27, 2015: Hundreds of students protested against a French magazine and stormed a school demanding it close

Figure 1: Precursor event sequence discovered by our method for a protest event.

Figure 1 shows a precursor event sequence discovered by our proposed model. The target event is a student protest against a French satirical magazine in the city of Bannu in Pakistan. One week before this event, there were several highly related events that occurred in other cities and may have influenced this target event of interest. For instance, the government of Pakistan passed resolutions condemning the publication in this French magazine. Later, different groups expressed concern in small gatherings in Islamabad and Lahore followed by protests in multiple cities including Bannu.

In this paper, we propose **STAPLE**, a multi-task **S**patio-**T**empor**A**l **P**recursor **L**earning and **E**vent forecasting framework for multiple cities, specifically designed to discover precursors across geolocations with imbalanced class distributions and partial labels. The primary datasets of interest are open source news articles across the world encoded into events, where each event has vital information including a timestamp (in granularity of days), a geolocation (at the city level),

*Discovery Analytics Center, Computer Science department, Virginia Tech. {yning, rrtao, ckreddy, naren}@vt.edu

†Computer Science department, George Mason University. rangwala@cs.gmu.edu

‡Lockheed Martin ATL. james.c.starz@lmco.com

a description (plain text), and an event type (e.g., protests). Our objective is to build forecasting models for specific cities and to identify evidential precursors from multiple cities in the past news articles.

This problem is non-trivial and poses several unique challenges: (i) *Temporal ordering constraints on events.* Events are often carefully sequenced in terms of their precursors and ignoring temporal information that is inherent in event evolution leads to unsatisfactory results. (ii) *Lack of class labels for precursor documents.* While events of interest can be manually (or automatically) detected and classified, labels for associated precursor documents (which are larger in number) are not available and are expensive to obtain. (iii) *Data scarcity and imbalanced distribution in certain geolocations.* Although a few transfer learning algorithms [19, 16, 20] support inference of the type considered here, none of them can tackle the data insufficiency problem in the presence of spatio-temporal correlations. (iv) *Inadequacy of static features.* It has been demonstrated [22] that successful event forecasting requires moving beyond static features, e.g., combining keyword frequencies with dynamic graph features. Thus the proposed forecasting models must support learning of appropriate representations inherently within its model-building.

We observe that by taking advantage of spatio-temporal event correlations within a multi-task learning framework, about 86% of the cities in our dataset have improved F1 scores compared to the best state-of-the-art algorithm. 60% of cities have more than 20% improvements in F1 score, especially for cities with less training data. We summarize the key contributions of this paper as follows:

- **Dynamic graph constraints for precursor learning and event forecasting**: Our model exploits event count correlations across multiple locations under dynamic temporal constraints for jointly forecasting events and identifying precursors. A fusion penalty is proposed to coordinate the forecasting tasks in related cities.

- **Augmented representation learning for precursors**: By integrating document and entity embeddings within a multiple instance learning framework, it assists the prediction model to track significant entities that are evident based on their estimated probabilities.

- **Multi-task learning for precursor mining**: It alleviates the data insufficiency problem by simultaneously learning multiple related tasks and restricting all cities to share a common set of features with a consensus model.

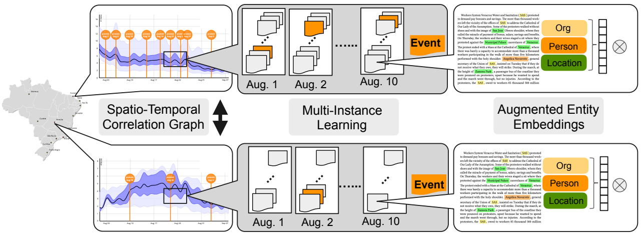


Figure 2: Overall System Framework.

- **Comprehensive set of experiments in real-world data**: We evaluate the proposed model on real-world datasets collected from six countries and more than one hundred cities. We conduct quantitative and qualitative analyses on the precursors inferred by the proposed model.

2 Problem Statement

Given multiple cities (or geolocations) within a country, we focus on the problem of predicting the occurrence of a future protest within a target city using captured open source news feeds. Figure 2 provides an overview of our proposed approach. Specifically, we hypothesize that the correlations between events occurring across "space" and "time" lead us to effectively forecast future events of interest. Besides forecasting the protest, we also aim to identify specific news articles as *precursors* across different locations for manual inspection. Our proposed **STAPLE** model seeks to capture these correlations by jointly training the models across different cities within a nested multiple instance learning framework [13].

Formally, given a set of multiple cities K, each city k, has a set of associated news articles and event indicators $(i = 1, ..., N_k)$ which are denoted by (X^k, Y^k).

$$(2.1) \qquad Y_i^k = \begin{cases} 1 & \text{if an event occurs after } X_i^k \\ 0 & \text{otherwise} \end{cases}$$

The news articles published H days before an indicator are given by X_i^k (bag). The proposed prediction model seeks to estimate the probability of occurrence of a future event in city k, P_i^k given X_i^k. Since instance-level labels are not provided, simple multi-instance learning(MIL) structures or complicated layers of MIL can both be applied in this problem. Given the reported performance [13], we use the nested multiple instance learning approach that allows for the transfer of class labels (target events) to individual news articles. It provides probabilistic estimates per document, per day, and per event for each city. The key contribution of this work is to *formalize a multi-task precursor learning model to study dynamic temporal relationships of multiple geolocations.* Also, we derive an optimization method to solve this problem.

3 Methodology Design

3.1 The Proposed Model Given K cities, let us assume $\Theta = (\boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^k, ..., \boldsymbol{\theta}^K)$ be the model parameters to be learned for each of the cities, where $\boldsymbol{\theta}^k \in \mathcal{R}^m$, and m is the dimension of the feature space representing each document. We model the document-level probability estimates p_{hj} for a news article j published on day h in city k to associate with the event of interest using a logistic function given by $p_{hj} = \sigma(x_{hj}^T \theta^k)$ where $\sigma(a) = 1/(1+e^{-a})$. Document embeddings (x_{hj}) are described in detail in Section 3.3. A specific document is considered to be more related to the event of interest when the probability estimate is high. Using the learned θ^k for a city, the nested multiple instance learning framework provides probabilistic estimates by averaging at the day-level (intermediate level) and at the event-level which considers a set of consecutive days before the target event.

The **STAPLE** model seeks to jointly learn the different model vectors, Θ, across the K cities using the multi-task learning paradigm, given by:

$$\min_{\Theta} \sum_{k \in K} \frac{N_k}{N} \mathcal{L}(\boldsymbol{\theta}^k) + \lambda R(\Theta) \tag{3.2}$$

where N_k is the number of training examples available for city k and N is the total number of training examples across all the cities. $\mathcal{L}$ is the loss function and R is the regularization function. The two-level multiple instance loss function is designed to minimize the prediction error associated with forecasting events, to enforce consistency in probabilistic estimates obtained for consecutive days, and also to maximize the difference between positive and negative instances using an unsupervised hinge loss function. The regularization term explicitly captures the spatio-temporal correlations between the occurrence of events across different cities. A more specific form of the objective function corresponding to Eq. (3.2) is given below:

$$\begin{aligned} \min_{\Theta} \sum_{k \in K} \Big(\frac{N_k}{N} \mathcal{L}(\boldsymbol{\theta}^k) + \frac{\lambda_1}{2} \sum_{i}^{N_k} \sum_{l \in \mathcal{G}_t} \alpha_{k,l}^{t_i} (\boldsymbol{\theta}^k - \boldsymbol{\theta}^l)^2 \\ + \frac{\lambda_2}{2} ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^k||_2^2 + \frac{\lambda_3}{2} ||\boldsymbol{\theta}^k||_2^2 \Big) \end{aligned} \tag{3.3}$$

where k, l are the indices for cities, θ_k is the model parameter for city k, $\hat{\theta}$ represents the global model, t_i is the time index for the current event indicator, $\alpha_{k,l}^{t_i}$ is the weight between city k and city l at time t_i, $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters. Different types of penalty functions allow us to enforce different behaviors in the evolution of the event across multiple geolocations.

Within the MTL framework the regularization term is designed to enforce different penalties such that related tasks share similar features or model parameters. The **STAPLE** model explicitly enforces pairs of cities within a country that have seen similar events occur in the past learn similar model vectors. $\mathcal{G}_{t_i}$ is the correlation graph for time t_i that determines the tasks (cities) with similar event profiles to each other.

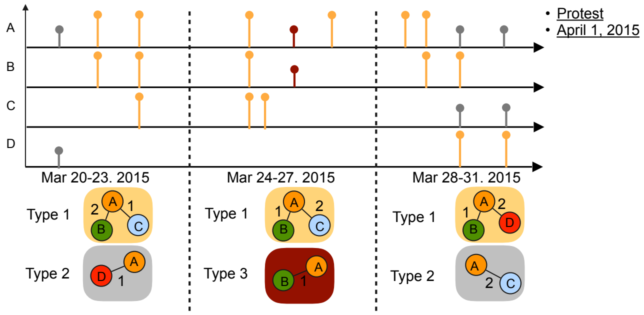


Figure 3: An example of the spatio-temporal correlation graph. Each node (A,B,C,D) represents a city. An edge between two nodes indicates that the same type of events occurred in the same time window for these two cities.

3.2 Event Correlation Graph We demonstrate the concept of the spatio-temporal correlation graph in Figure 3. Each node represents a city in a country. To predict the occurrence of a protest in city A on April 1, 2015, the model analyzes the past few days of data to discover if there is an event in city A and also in other cities (i.e., city B, C and D) from Mar. 20 to Mar. 31, 2015. The weight on the edge denotes the minimum number of common events between the two cities. From Mar. 20-23, the neighbor network for city A consists of three cities B, C, D with two types of events. $\alpha_{k,l}^{t_i}$ is the normalized weight on the edge between city k and city l, given by:

(3.4)

$$\alpha_{k,l}^{t_i} = \Big(\sum_{c} \sum_{t=t_i - H}^{t_i} \min(E_t^k(c), E_t^l(c)) \Big)' + \Big(\frac{1}{\mathrm{dist}(k,l)} \Big)'$$

Here c is the event type (such as a protest), and $E_t^k(c)$ is the number of events of type c in city k that occur within time window t. We scale the value of event count and $1/\mathrm{dist}()$ by feature scaling function $(x)' = \frac{(x - x_{\min})}{x_{\max} - x_{\min}}$ into a range of (0.0, 1.0). Given the spatio-temporal correlation graph $\mathcal{G}$ and the edge weights, it is reasonable to assume that two cities will share several common edges, as they tend to be influenced by the same set of covariates. When a city has large number of training examples, the empirical loss helps to reduce the prediction error. The spatio-temporal correlation constraint captures the task relatedness between multiple cities within a given time period. The *dist* function returns the distance

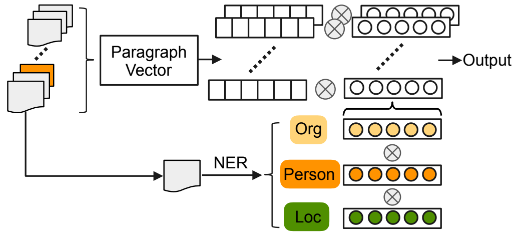


Figure 4: Learning representations. Top portion corresponds to the document embeddings and the bottom portion corresponds to the entity embeddings.

between city k and city l. We assume that two cities that are far away from each other have fewer correlations/similarities in their models.

Besides the spatio-temporal constraints we also enforce the learned individual model vectors to not deviate from the global average along with a $l2$-norm constraint on the weight vectors. The regularization parameters control the model complexity by enhancing robustness and the MTL constraints alleviate the data insufficiency problem for each individual task (if learned separately).

3.3 Learning Representations

One of the challenges in precursor mining for event forecasting is to select informative and related documents. A precursor event is not necessarily similar in semantics to the event of interest. In this work, we make use of *augmented distributed representations* of the documents to discover progression of precursors towards the target events. More specifically, we study the following aspects of representations for each historical news document $\mathbf{x}$:

Document Embeddings: Recent work has shown that the semantic relationships of words can be effectively captured using the geometry of a continuous embedding space [8, 12]. For the articles in each country, we learn distributed representations (vectors) [8] for text documents and utilize these embeddings in our experiments.

Entity Embeddings: In many societal events, the roles of significant entities such as government officers or large organizations are substantial and sometimes even influence the progression of events. We focus on location names, person names, and organization names as the primary entities of interest. Entity and relation embeddings have been studied in structured learning and knowledge graph modeling [10]. More specifically, we use the Stanford Named Entity Recognizer (NER) [5] to extract a set of entities for English and use a series of language enrichment steps (see [14] for details) for processing. We apply the Continuous Bag-of-Word model (CBOW) [11] to each dataset to obtain word level embeddings. Finally, we aggregate the embeddings of entities into the representation for text documents as shown in Figure 4.

Algorithm 3.1 STAPLE Model Learning

```
Input: $Dataset(X^k, Y^k), k \in [1, ...K], \tau$
Output: $\Theta = [\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^k]$
for $\tau$ iterations do
    randomize$(X^k, Y^k)$
    for city $k$ in $K$ do
        for $i \in [1, 2, ..., N_k]$ do
                        ▷ training examples for city $k$
            graph construct $\mathcal{G}_{t_i}$
            calculate $\alpha^{t_i}_{k,l}$                ▷ Eq. (3.4)
            calculate gradient $\nabla(\Theta)$   ▷ Eq. (3.5, 3.6)
            update $\Theta$ using $\nabla(\Theta)$ based on SGD
return $\Theta$
```

3.4 Optimization

The optimization problem is solved using the mini-batch gradient descent algorithm for each of the model parameters, $\boldsymbol{\theta}^k$, described in Algorithm 3.1. We update the weight vector using an adaptive learning rate $\boldsymbol{\theta}^k_{\tau+1} = \boldsymbol{\theta}^k_{\tau} - \eta\nabla(\boldsymbol{\theta}^k)$ where η is the learning rate at the current iteration. More specifically,

$$\boldsymbol{\theta}^k \leftarrow \boldsymbol{\theta}^k - \eta\big[\partial\frac{\mathcal{L}(\boldsymbol{\theta}^k)}{\boldsymbol{\theta}^k} + \lambda_1 \sum_{l \in \mathcal{G}_{t_i}} \alpha^{t_i}_{k,l}(\boldsymbol{\theta}^k - \boldsymbol{\theta}^l) - \lambda_2(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^k) + \lambda_3\boldsymbol{\theta}^k\big] \tag{3.5}$$

$$\hat{\boldsymbol{\theta}} \leftarrow \hat{\boldsymbol{\theta}} - \eta\lambda_2(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^k) \tag{3.6}$$

For each city, we update the global model parameters $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^k$ in an alternate manner. The spatio-temporal correlation graph and the weights on edges are precomputed only using the training set.

3.5 STAPLE Model for Precursor Mining

The precursor documents are identified based on their estimated probabilities from the learned model of each city as described in Algorithm 3.2 from lines 3 to 11. We also discover precursors for each city in the past time window from the neighboring cities (lines 12 to 17). For instance, in Figure 3, the precursors for the protest event in city A are also explored in the news articles geolocated at its neighboring cities, B to D using the learned model vectors for each of the cities. If the estimated probability is above a certain threshold, we select it into the precursor candidate pool for the target event in city A. The time complexity is dependent on the number of examples in the training dataset and the number of historical documents for each event. Based on the estimated probability of instances for events, the precursor documents and entities can be selected in linear time

Algorithm 3.2 STAPLE for Precursor Discovery

```
1:  Input: Dataset $(X^k, Y^k), k \in [1, ..., K]$
2:  Output: the estimated probabilities set $Q$, the
    discovered precursor set $O$.
3:  $\Theta \leftarrow$ call Algorithm 3.1 for training
4:  for city $k$ in $K$ do
5:      for $i \in [1, 2, ..., N_k]$ do
6:          for $\mathcal{X}_h \in X_i^k$ do
7:              for $\mathbf{x}_{hj} \in \mathcal{X}_h^k$ do
8:                  estimate $p_{hj}$ using $\boldsymbol{\theta}^k$
9:                  if $p_{hj} \geq 0.5$ then
10:                     $O_i^k \leftarrow \mathbf{x}_{hj}, Q_i^k \leftarrow p_{hj}$
11:         estimate $P_i^k$ by Avg($p_{hj}$)
12:         get "neighbor" cities for $k$ from $\mathcal{G}_{t_i}$
13:         for city $l$ in the "neighbor" cities do
14:             for $h = [1, ..., H]$ do
15:                 for $\mathbf{x}_{hj} \in \mathcal{X}_h^l$ do
16:                     $p_{hj} = \sigma(\boldsymbol{\theta}^l, \mathbf{x}_{hj})$
17:                     $O_i^k \leftarrow \mathbf{x}_{hj}$ if $p_{hj} \geq \xi$
    return $O, Q$
```

with respect to the number of historical documents for each event (lines 7 to 10).

4 Experimental Setup

4.1 Datasets We evaluated our models on event encoded data from six countries. Among them, three countries, Columbia (CO), Paraguay (PY), and Venezuela (VE) are from a labeled set called Gold Standard Report (GSR) [14] from January 2014 to April 2015. The GSR is a manually curated dataset that records civil unrest events from the ten most significant news outlets as ranked by the International Media and Newspapers in each country. An example of a recorded event is given by its city, state, country, date, and event type. The other three datasets, Pakistan (PK), Iran (IR) and Afghanistan (AF) are from the ICEWS dataset [4] which stands for Integrated Crisis Early Warning System. It contains news articles published all over the world with the goal of evaluating national and international crisis events. Events are automatically identified and extracted by the BBN ACCENT event encoder. In our experiments, we only use data from 2015 and 2016. Each news article is labeled in one of the 20 categories. We use "protest" events as our positive examples and no protest days (or other types of events) as negative examples. We observe that the event distribution varies significantly across cities, with large cities having relatively more event occurrences compared to small cities. Statistics about these datasets are shown in Table 1. We use protests in the aforementioned Latin American and Middle East/Asian countries to rigorously evaluate the performance of our framework and also consider other countries (e.g., France) to provide case studies of how our framework works because these countries feature more well-known protests (e.g., climate protests in France, see Table. 3).

Table 1: Datasets used in our experiments. CO: Columbia, PY: Paraguay, VE: Venezuela, PK: Pakistan, IR: Iran, AF: Afghanistan

	CO	**PY**	**VE**	**PK**	**IR**	**AF**
# news	8,386	7,879	9,390	64,868	38,113	27,786
# events	604	971	1,911	1,291	1,084	713

4.2 Experimental Protocol We learn 300-dimensional representations for documents and 100-dimensional embeddings for words. Each document is represented by concatenating the document and entity embeddings. The entity embedding is an aggregation of each entity within the document. The GSR and ICEWS datasets record protest events on a given day at a specific location (city level). To evaluate our proposed model, for each protest event, we extract all the published reports (news articles) for up to two weeks before the occurrence of the specific event of interest. This ordered collection of per-day news documents to the protest day are considered as positive super bags. For negative examples, we identify consecutive sets of days within our studied time periods for each city when no protest event was reported.

4.3 Comparative Methods (1) Multi-Instance SVM (MI-SVM [2]): The MI-SVM model extends the notion of a margin from individual patterns to bags. The margin is defined between the "most positive" instance of the positive bag and the "least negative" instance of the negative bag. We collapse the news articles from the different historical days into one bag and apply this standard MIL formulation based on the SVM in scikit-learn [1]. **(2) Relaxed MIL Model (rMIL$^{\textbf{avg}}$ [18])**: We estimate the probability of each instance in a bag using a logistic function. We use the average of instance-level probabilities to model the probabilities for bags. **(3) Nested MIL (nMIL [13])**: This approach applies a nested level of multi-instance learning for the event forecasting problem. It learns a "global" model for all cities in a country. **(4) Transfer model (STAPLE-tx)**: This method is a simpler variant of the proposed model. Event forecasting for each city is considered as one task. We first apply our objective function to cities with "rich"

[1] http://scikit-learn.org/stable/modules/svm.html

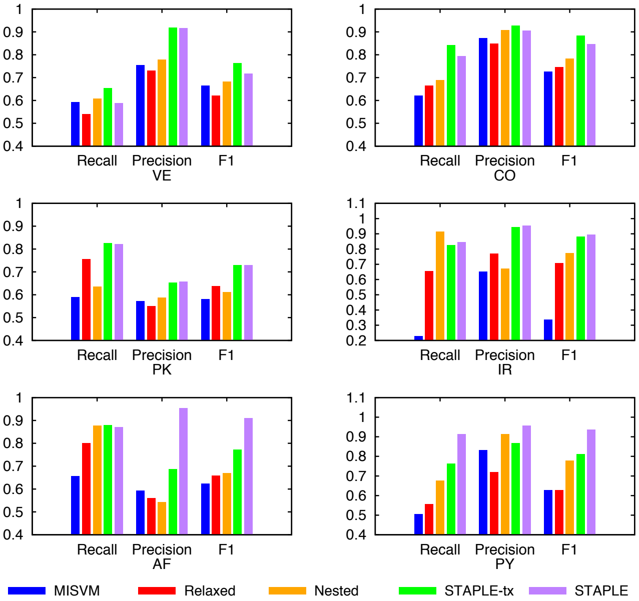


Figure 5: Prediction performance (Recall, Precision, and F1 scores) on six datasets for the comparison methods.

datasets (more event examples). After learning models for these source cities, we incorporate these models into an average model with mean and standard deviation μ, σ and transfer them to the cities that have "sparse" datasets [15]. The target cities will then learn their models by assuming that their model parameters are drawn from a Gaussian distribution $N(\boldsymbol{\mu}, \sigma)$ where $\mathcal{S}$ denotes the set of source cities:

$$\boldsymbol{\mu} = \frac{1}{K-1}\sum_{k\in\mathcal{S}} \boldsymbol{\theta}^k, \ \sigma = \sqrt{\frac{1}{|\mathcal{S}|-1}\sum_{k\in\mathcal{S}}(\boldsymbol{\theta}-\boldsymbol{\mu})^2} \tag{4.7}$$

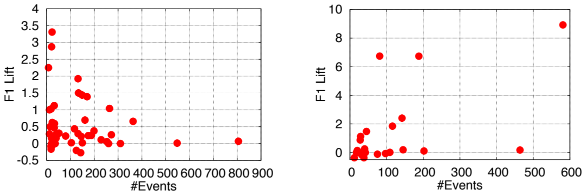
5 Experimental Results

We perform a comprehensive empirical study to evaluate the performance of the proposed models in terms of event recall, event precision, and F1 score. We also analyze the quality of precursors with quantitative metrics and detailed case studies that highlight the strengths of the **STAPLE** model.

5.1 How well does STAPLE forecast? Figure 5 shows the prediction performance of **MI-SVM**, **rMIL$^{\text{avg}}$**, **nMIL** and **STAPLE** methods in terms of recall, precision, and F1 scores in the six datasets considered. The model hyperparameters, λ_1, λ_2 and, λ_3 are determined using a validation set for each country. m_0 and p_0 are set to 0.5 following [13]. For other comparison models, hyperparameters are chosen following their criterion in their papers. The **STAPLE** method outperforms the best state-of-the-art approach (**nMIL**) by 16% to 35% for AF, IR, and PK and 5% to 15% for CO and PY, in term of F1 scores. For CO and VE, the **STAPLE-tx** model outperforms other methods in terms of F1 score. The results in Table 2 demonstrate that the methods that take into account the entity embeddings (second column) perform better than that of using the document embeddings only (first column) for AF and IR datasets but not for PK in terms of F1 score.

Table 2: F1 evaluation for the proposed methods (Text Embeddings vs Text+Entity Embeddings) .

		Text	Text+Entity
Data	**Model**	F1	F1
AF	**STAPLE-tx**	0.772	**0.783**
	STAPLE	0.910	0.910
IR	**STAPLE-tx**	0.880	**0.893**
	STAPLE	0.895	**0.904**
PK	**STAPLE-tx**	0.729	0.725
	STAPLE	0.730	0.713

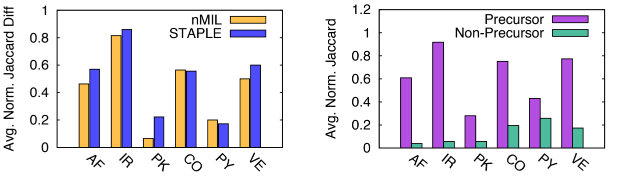


(a) ICEWS (b) GSR

Figure 6: F1 lift per city using **STAPLE** compared to state-of-the-art model, **nMIL**. X-axis denotes the number of events at the city level. Y-axis denotes the F1 lift from the STAPLE model.

5.2 Does the Spatio-Temporal Event Correlation Graph help? Figure 6 depicts the improvements of F1 scores of the **STAPLE** model compared to the **nMIL** model ($F1$ Lift $= \frac{\textbf{STAPLE}_{F1} - \textbf{nMIL}_{F1}}{\textbf{nMIL}_{F1}}$) versus the number of events per city. We observe that the most significant improvements are for cities such as Qom and Kerman in Iran which are relatively small cities in their respective countries.

5.3 How good are the precursors? Figure 7 shows the average normalized Jaccard Index of precursor documents and non-precursor documents (with respect to the target document) discovered by the **STAPLE** and **nMIL** models. The normalized Jaccard index is computed as the pairwise Jaccard index, scaled relative to each event of interest. It is clear that the precursor documents have higher text-based similarity to the target events compared to the non-precursor documents as in



(a) **nMIL** vs. **STAPLE** (b) Precusor vs. The rest

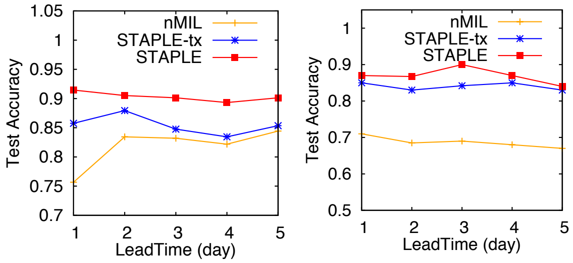
Figure 7: Comparison of the average normalized Jaccard index for precursor news and non-precursor news articles. (a) Shows the difference between Jaccard scores for the precursor and non-precursor documents for **STAPLE** and **nMIL** models. (b) Shows average normalized Jaccard index for precusor and non-precursor documents for **STAPLE**.

Figure 7b. Figure 7a shows that the **STAPLE** model discovers more semantically related documents to the target events compared to the **nMIL** model [13] for most of the countries.

5.4 Case Studies Table 3 demonstrates a case study on precursor story lines identified by the **STAPLE** approach. The key strength of **STAPLE** is its ability to leverage correlated events occurring across different cities. For each protest event, **e**, in a city A, we form a neighborhood city set based on the spatio-temporal correlation graph. Then for each city in the graph, we estimate the probabilities of news in the past week using its model for event **e**. As long as the probability is above a certain threshold, we select the news to be a precursor candidate for this event: $O_e^A \leftarrow x$ if $p(x) \geq \xi$. We studied ICEWS dataset from France as a case study. In the case of a protest event in Paris, France; thousands of activists started a protest pleading leaders to stop global warming near the site of a former terror attack. Our model successfully captured the following related events leading up to this protest: A week earlier, people in Toulouse marched for the devastating attacks in Paris. Many countries announced they will be attending the upcoming Climate Change Summit (COP21) in Paris. Two days before, Australia kicked off climate rallies ahead of these global talks. One day before, activists claimed the need to march for climate change amid terror threats.

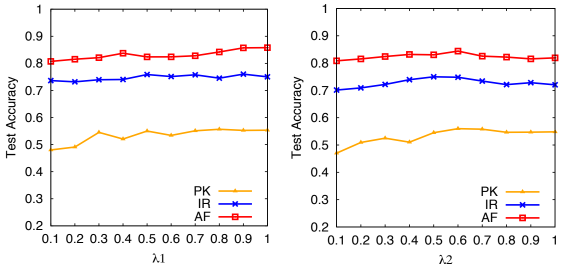
In all cases, our model captured the key change points in the precursor story lines for protest events and the key entity names (such as government officials) that played a crucial role in the development of these events.

5.5 How early can the STAPLE model forecast? Leadtime indicates the number of days in advance that the model makes predictions and historical days denotes the number of days over which the news articles are extracted as input to the prediction algorithms. In order to study the changes in performance with and without the spatio-temporal event correlation graph constraints, we present the accuracy (ACC) score with varying lead times from 1 to 5 days for the **STAPLE** and the **nMIL** models in Figure 8. Due to space constraints, we only depict results for PY and IR. The **STAPLE** model is stable and consistently performs better compared to others with varying values of leadtime.



(a) **PY** (b) **IR**

Figure 8: Test accuracy for two countries over leadtime ranging from 1 to 5 days. x-axis denotes the leadtime and y-axis denotes the accuracy score over the test dataset.



(a) λ_1 (b) λ_2

Figure 9: Sensitivity analysis on hyper-parameters λ_1 and λ_2 ($= \lambda_3$). X-axis represents the varying values of λ_1, λ_2 and Y-axis denotes the test accuracy.

5.6 Sensitivity Analysis Figure 9 illustrates the accuracy of the proposed **STAPLE** model on three test dataset by varying the hyper parameters of λ_1 and λ_2 ($= \lambda_3$). λ_2 and λ_3 are chosen to be the same according to our parameter search result. The performance of different values is relatively stable and is shown for only: AF, PK, and IR countries due to the space constraints.

6 Related Work

Spatial Event Forecasting Predicting future events of interest (e.g., protests) from large, heterogeneous, open source feeds is an important and active area of research [21]. Over the years, supervised and

Table 3: Automatically discovered precursor news stories and relevant entities for one protest event of interest.(i) Precursors for an event in one city are inferred across other cities as well. (ii) Key entities participating in the events provide explanatory power to the precursors. (iii) Probability of protest events gradually increases with the accumulation of precursor events.

Date	Location	Precursor News Summary	Entity	Prob
2015-11-22	Toulouse	**P1**.More than 10,000 people marched Saturday in the French city of Toulouse for peace and against "barbarity" a week after the devastating attacks in the capital.	French, Toulouse	0.81
2015-11-23	Paris	**P1**.Wealthy governments and other donors need to invest more to reduce carbon emissions stemming from agriculture, said a study issued ahead of U.N. climate talks in Paris next week.	UN, Paris	0.68
2015-11-24	Paris	**P1**. China and US have vowed to join hands with France and other parties to work toward success at the UN climate summit in Paris.	China, US, UN, Paris	0.81
2015-11-25	Paris	**P1**. The international community must secure a binding deal against climate change at key UN talks in Paris next week, German Chancellor Angela Merkel said Wednesday.	German, Angela Merkel, UN, French, Aquino, Paris	0.83
2015-11-26	Paris	**P1**. It comes days ahead of a major UN climate summit in Paris which aims to forge an international deal to stop global warming. **P2**. Pope Francis visited the world's poorest continent to issue a clarion call for the COP21.	UN, Pope Francis, COP21, Paris	0.90
2015-11-27	Paris	**P1**. Leaders from Russia Germany, and Europe may meet around the upcoming UN climate change conference in Paris. **P2**. Australia kicks off climate rallies ahead of global talks.	Russia, Germany, Europe, Australia, Paris	0.94
2015-11-28	Paris	**P1**. Activists plan to join arms and form a "human chain" in Paris on Sunday to urge action on global warming, in a muted rally after attacks on the city.	UN, Paris	0.92
2015-11-29	Paris	**Protest in Paris, France**: Around 4,500 activists had earlier linked hands in a peaceful protest near the site of the deadliest of the attacks, pleading for leaders to curb global warming.		

unsupervised learning approaches have been developed to tackle this problem in different domains. Advanced techniques which use a combination of sophisticated features, such as topic related keywords, as input to support vector machines, LASSO, and multi-task learning approaches [17, 22] have also been studied. Ramakrishnan *et al.* [14] designed a comprehensive framework (EMBERS) for predicting civil unrest events in different locations, using a combination of machine learning models ingested with heterogeneous input sources ranging from social media to satellite images. Laxman *et al.* [7] designed a generative model for categorical event prediction in streaming data by identifying frequent episodes. A recent work [19] has studied a multi-modal transfer learning method (FLORAL) to transfer knowledge from a city where there is sufficient multi-modal data and labels, to other cities and locations.

Precursor Discovery In the multiple instance learning (MIL) paradigm [23, 9], labels are associated with sets of instances commonly referred to as *bags* or *groups* instead of individual instances. Individual instance-level labels are unknown or missing. MIL turns out to be a natural fit for the precursor mining problem because the labels are only associated with the events of interest but not the precursor documents in history. Traditional MIL formulations make strong assumptions, e.g., that the aggregation function over instance labels is a noisy-OR function; i.e., the positive bags contain at least one positive instance and the negative bags contain only negative instances. A recent work [6] has developed instance-level predictions from group labels (GICF) which allows for general aggregation functions. Other multiple instance learning approaches and applications are surveyed in [1]. A nested multi-instance learning (nMIL) framework [13] has been proposed to forecast civil unrest events and detect precursor news articles for these events. However, this approach does not account for spatial dependencies and does not perform satisfactorily in the presence of limited amount of data.

Representation Learning In practice, finding good feature representations to model news articles is not a trivial problem. Traditionally, the bag-of-words representation allows for easy interpretation but requires preprocessing and feature selection. Several researchers have developed efficient and effective neural network based representations for language models [3, 12]. Entity recognition has also been widely applied in natural language processing tasks [10, 5]. Most societal events are related to or even caused by known entities such as persons, organizations, and geolocations. In this paper, we combine document level embeddings with recognized entity embeddings for better explanation of precursor story lines and event forecasts.

The methods proposed in this paper can be viewed as complementary to the prior work discussed above, casting the spatial event forecasting and precursor discovery problems into a multi-instance learning framework with a fusion penalty based on spatio-temporal correlation graphs and augmented representations.

7 Conclusion

We presented **STAPLE**, a multi-task spatio-temporal correlation graph model based on a two-level multi-instance learning (MIL) framework for precursor mining coupled with event forecasting. Multiple models for cities are jointly learned together and proven to be effective at both forecasting events and discovering precursors. The richness of the identified precursor events demonstrates that **STAPLE** will be a useful tool for a better understanding of event happenings. For future work, we plan to study narrative generation and entity knowledge graph extraction from with multi-source datasets.

Acknowledgements

This work was partially funded by the Office of Naval Research under contract N00014-16-C-1054 and by the U.S. Department of Homeland Security under Grant Award Number 2017-ST-061-CINA01. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Office of Naval Research or the U.S. Department of Homeland Security.

References

[1] Jaume Amores. Multiple instance classification: Review, taxonomy and comparative study. *Artificial Intelligence*, 201:81–105, 2013.

[2] Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, pages 561–568, 2002.

[3] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003.

[4] Elizabeth Boschee, Jennifer Lautenschlager, Sean O'Brien, Steve Shellman, James Starz, and Michael Ward. Icews coded event data, 2016.

[5] Jenny Rose Finkel, Trond Grenager, and Christopher Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *ACL*, pages 363–370, 2005.

[6] Dimitrios Kotzias, Misha Denil, Nando de Freitas, and Padhraic Smyth. From group to individual labels using deep features. In *KDD*, pages 597–606, 2015.

[7] Srivatsan Laxman, Vikram Tankasali, and Ryen W. White. Stream prediction using a generative model based on frequent episodes in event sequences. In *KDD*, pages 453–461, 2008.

[8] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *ICML*, pages 1188–1196, 2014.

[9] Yu-Feng Li, Ju-Hua Hu, Yuan Jiang, and Zhi-Hua Zhou. Towards discovering what patterns trigger what labels. In *AAAI*, pages 1012–1018, 2012.

[10] Yankai Lin, Zhiyuan Liu, Maosong Sun, Yang Liu, and Xuan Zhu. Learning entity and relation embeddings for knowledge graph completion. In *AAAI*, pages 2181–2187, 2015.

[11] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781, 2013.

[12] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. *CoRR*, 2013.

[13] Yue Ning, Sathappan Muthiah, Huzefa Rangwala, and Naren Ramakrishnan. Modeling precursors for event forecasting via nested multi-instance learning. In *KDD*, pages 1095–1104, 2016.

[14] Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, and et al. "Beating the news" with EMBERS: Forecasting civil unrest using open source indicators. In *KDD*, pages 1799–1808, 2014.

[15] Michael Rosenstein, Zvika Marx, Tom Dietterich, and Leslie Pack Kaelbling. Transfer learning with an ensemble of background tasks. In *NIPS Workshop on Inductive Transfer*, 2005.

[16] Ben Tan, Evan Wei Xiang, Qiang Yang, and Erheng Zhong. Multi-transfer: Transfer learning with multiple views and multiple sources. In *SDM*, pages 243–251, 2013.

[17] Xiaofeng Wang, Matthew S. Gerber, and Donald E. Brown. Automatic crime prediction using events extracted from twitter posts. In *SBP*, pages 231–238, 2012.

[18] Xinggang Wang, Zhuotun Zhu, Cong Yao, and Xiang Bai. Relaxed multiple-instance SVM with application to object discovery. *CoRR*, 2015.

[19] Ying Wei, Yu Zheng, and Qiang Yang. Transfer knowledge between cities. In *KDD*, pages 1905–1914. ACM, 2016.

[20] Pei Yang and Wei Gao. Multi-view discriminant transfer learning. In *IJCAI*, pages 1848–1854, 2013.

[21] Chao Zhang, Guangyu Zhou, Quan Yuan, Honglei Zhuang, Yu Zheng, Lance Kaplan, Shaowen Wang, and Jiawei Han. Geoburst: Real-time local event detection in geo-tagged tweet streams. In *SIGIR*, pages 513–522, 2016.

[22] Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. Multi-task learning for spatio-temporal event forecasting. In *KDD*, pages 1503–1512, 2015.

[23] Zhi-Hua Zhou and Jun-Ming Xu. On the relation between multi-instance learning and semi-supervised learning. In Zoubin Ghahramani, editor, *ICML*, volume 227, pages 1167–1174, 2007.