Precursor Learning for Crowd Forecasting

YUE NING*, Virginia Tech
SATHAPPAN MUTHIAH, Virginia Tech
HUZEFA RANGWALA, George Mason University
DAVID MARES, University of California at San Diego
NAREN RAMAKRISHNAN, Virginia Tech

Forecasting the activities of large crowds is an important and challenging problem in many domains. Mass gatherings underlie civil disobedience activities and as such run the risk of turning violent, causing damage to both property and people. From the perspective of human analysts and policy makers, forecasting algorithms must not only make accurate predictions but must also provide supporting evidence, e.g., the causal factors related to the event of interest. Understanding of such causal factors is crucial for instance to understanding how events will unfold and, in particular, whether they could lead to violence. We propose a nested multiple instance learning based approach, **nMIL**, that jointly tackles the problem of identifying evidence-based precursors from a stream of news articles and forecasts events and their characteristics into the future. Using data from three countries in Latin America, we demonstrate how our approach is able to consistently identify news articles considered as precursors for protests and violence. Our empirical evaluation demonstrates the strengths of our proposed approach in filtering candidate precursors, in forecasting the occurrence of events with a lead time advantage, and in accurately predicting the characteristics of civil unrest events.

Additional Key Words and Phrases: Event Forecasting, Precursor Learning, Multi-Instance Learning

ACM Reference format:
Yue Ning, Sathappan Muthiah, Huzefa Rangwala, David Mares, and Naren Ramakrishnan. 2016. Precursor Learning for Crowd Forecasting. 1, 1, Article 1 (January 2016), 24 pages.
DOI: 0000001.0000001

1 INTRODUCTION

Large public crowd gatherings are common in all forms of society and some of them can lead to violence, involving damage to both property and people. Examples of such crowd gatherings include political rallies, protests, and commemorative events. When a crowd turns violent it generates economic, political and social costs, in addition to the emotional and physical (including death) consequences for individuals directly involved in the violence. Each of the parties who support the right to peaceful gatherings (e.g., government and police officials, community organizations) seek to develop better insights into the triggers that can foment violence in hopes of reducing the risk of violence. Efforts to decrease the probability of a violent gathering without understanding the dynamics that differentiate violent

*The corresponding author

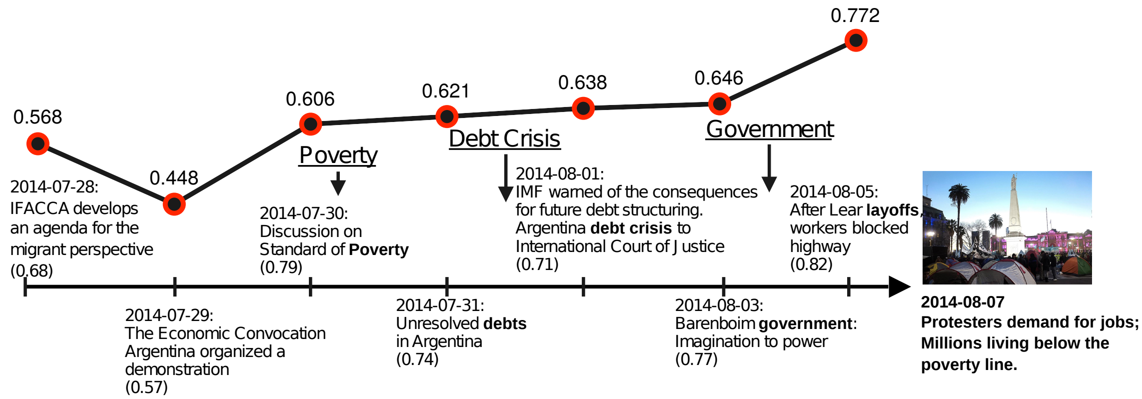


Fig. 1. Precursor story line for a protest event in Argentina. The x-axis is the timeline. The dots above with numbers are the probabilities for each day that the model generated for the target event. Each precursor document is titled in the timeline.

from non-violent events can lead to measures that instead increase that probability. For example, deploying a significant show of force with police and the military at the first sign of a protest can make the protesters feel intimidated and frustrated rather than protected. That frustration can, as we suggest, build into anger and increase the likelihood of violence during the next such gathering. In the past, open source data (e.g., social media and news feeds) have been proven to serve as surrogates in forecasting a broad class of events, e.g., disease outbreaks (Achrekar et al. 2011), election outcomes (O'Connor et al. 2010)(Tumasjan et al. 2010), stock market movements (Bollen et al. 2011) and protests (Ramakrishnan et al. 2014). While many of these works focus on predictive performance, there is a critical need to develop methods that also yield insight by identifying precursors to events of interest.

This paper focuses on the problem of identifying precursors (evidence) for forecasting significant societal events, specifically protests and violent crowd behavior. Modeling and identifying the precursors for a given event is useful for human analysts and policy makers as it discerns the underlying reasons behind the civil unrest movement. In particular, the objective of this paper is to study and forecast protests and violence across different cities in Latin American countries (Argentina, Brazil, Colombia, Paraguay, Mexico, and Venezuela). 6000 news outlets are tracked daily across these countries with the goal of forecasting protest occurrences with at least one day of lead time. From the news feeds, we also aim to identify the specific news articles that can be considered as precursors for the target event.

Figure 1 shows an example of precursors identified by our model. On the right of the timeline is a news report about a protest event in Argentina. The connected dots denotes the generated probabilities of a protest event over the days leading upto this protest. From this example, we find that within 10 days before the event, there are multiple precursor events identified as highly probable leading indicators of a protest. Most significant societal events are a consequence of several factors that affect different entities within communities and their relationships with each other (or the government) over time. In this specific example, the leading precursor was an article commenting on standards of living in Argentina and rising poverty levels. The International Court of Justice also delivered a verdict on the debt crisis. All these factors led to the final protest involving the general population across the country demanding better work opportunities.

We formulate the precursor identification and forecasting problem in a novel multiple instance learning algorithm (MIL) setting. Multiple instance learning algorithms (Andrews et al. 2002; Zhou and Xu 2007) are a class of supervised learning techniques that accept labels for groups of instances, but where labels for individual instances are not available. In our formulation, instances denote news articles and while class labels are not associated with individual news articles, a group of news articles in the days leading to a protest are attached with a label (indicating the occurrence of a protest). We further extend the standard MIL formulation by introducing a nested structure, wherein we group news articles published in a given day at the first level and then group the collection of individual days at the second level. This nested MIL approach allows for modeling the sequential constraints between the news articles published on different days and also provides a probabilistic estimate for every news article and the collection of news article. This estimate is significant because it indicates for a given news article the probability of it signaling a protest event. Recall that in our datasets we do not have any training labels to associate a protest per news article.

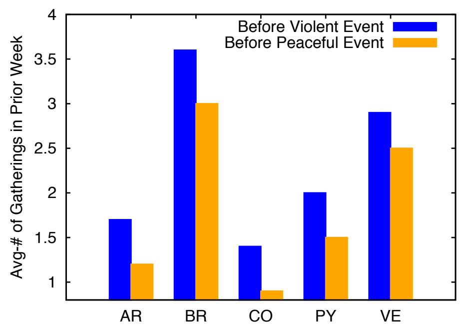


Fig. 2. Average number of gatherings (in prior week) that precede violent and non-violent events. AR: Argentina, BR: Brazil, CO:Colombia, PY: Paraguay, VE:Venezuela.

The modeling of precursors is not crucial to distinguish protests that might turn violent from those that do not. Figure 2 demonstrates the average number of public gatherings that have occurred before both violent and non-violent events in five South American countries. Before the occurrence of a violent event, more protests occur on average in the prior week in comparison to a non-violent event for all the countries. Inspired by this observation, we leverage the **nMIL** model to develop new methods that forecast the occurrence of violent crowd behavior in advance. In particular, by integrating the correlation between the past protest events and future violent protest events, the model forecasts outbreaks of crowd violence using historical web data in both spatial and temporal aspects.

The previous published conference paper (Ning et al. 2016) presented a novel nested framework of multi-instance learning for event forecasting and precursor mining. This paper improves upon the paper in the following ways:

(1) **Tailor the nested multi-instance model for forecasting violent crowd events.** This extension is built on the hypothesis that violent crowd behavior tends to have a qualitatively different set of trigger events signaling the occurrence in the future. The framework is advantageous over computer vision techniques, e.g. (Hassner et al. 2012), that only detect events (not forecast them) and which require the first images of violence to be published.

(2) **Capture the characteristics before violent events and non-violent events**. By taking advantage of the event occurrence in the prior time window, it distinguishes the violence from normal events and forecasts the violent event in the near future.

(3) **Identify precursor events for violent crowd behavior**. This framework also automatically identifies significant precursor events, such as unsatisfied protests, for violent crowd behavior.

Table 1. Notation used in this paper.

Variable	Meaning
$\mathcal{S}$	a set of n "super bags" and their labels, $\mathcal{S} = \{\mathbb{S}, Y\}$, in our dataset
N	number of superbags in our dataset
$\mathbb{S}$	a "super bag" which is also an ordered set of t "bags", $\mathbb{S} = [\mathbf{X}_i], i \in \{1, ..., t\}$
$\mathbf{X}_i$	a bag which is a set of instances, $\mathbf{X}_i = \{\mathbf{x}_{ij}\}, j \in \{1, ..., n_i\}$ with n_i=$\lvert\mathbf{X}_i\rvert$
$\mathbf{x}_{ij}$	the j-th instance in set $\mathbf{X}_i$, a V-dimension vector, $\Re^{V\times 1}$
Y	event label $\in \{-1, +1\}$ of super bag $\mathbb{S}$
V	violence label $\in \{-1, +1\}$ of super bag $\mathbb{S}$
P	estimated probability $\in [0, 1]$ for a super bag
P_i	the probability $\in [0, 1]$ of bag i in a super bag to be positive
p_{ij}	the probability of an instance $\mathbf{x}_{ij}$ in bag $\mathcal{X}_i$ in a super bag to be positive w.r.t an event
p_{ij}^{v}	the probability of an instance $\mathbf{x}_{ij}$ in bag $\mathcal{X}_i$ in a super bag to be positive w.r.t a violent event
O	a set of precursor documents for an event
C	multi-class label of super bag $C \in \{1, 2, ..., K\}$
$\mathbf{w}$	model parameter in the **nMIL** model, a V-dimension vector $\Re^{V\times 1}$.
$\mathbf{v}$	model parameter in the $\mathbf{nMIL^{vio}}$ model, a V-dimension vector $\Re^{V\times 1}$.

(4) **Application and evaluation with comprehensive experiments.** We evaluate the proposed methods for violent events using news data collected from in five countries of Latin America: Argentina, Brazil, Colombia, Paraguay, and Venezuela. For comparison, we implement other multi-instance algorithms, and validate the effectiveness and efficiency of the proposed approach. Specifically, we compare the violence warnings generated by the proposed method to the online system, EMBERS (Muthiah et al. 2016; Ramakrishnan et al. 2014). We also perform qualitative and quantitative analysis on the precursors for violence inferred by our model.

The rest of this paper is organized as follows. We describe the problem formulation in Section 2 and then introduce our proposed model based on multi-instance learning in Section 3. This section is then followed by experiments, presented in Section 4, and evaluations on real world datasets, presented in Section 5. We discuss the related work in Section 6. Finally, we conclude with a summary of the research in Section 7.

2 PROBLEM STATEMENT

Given a collection of streaming media sources (e.g., news feeds, blogs and social network streams), the objectives of our study are two-fold: 1) to develop a machine learning approach to forecast the occurrence of an event of interest in the near future. Specifically, we focus on forecasting protests or civil unrest movements in Latin America from a daily collection of published news articles. 2). besides forecasting the protest, we aim to identify the specific news articles from the streaming news outlets that can be considered as supporting evidence for further introspection by an intelligence analyst. We refer to these identified articles as *precursors* for a specific protest.

Figure 3 provides an overview of our proposed approach and problem formulation. Here, we show groups of news articles collected daily, five days prior to the specific protest event (being forecast). Within our proposed MIL-based formulation, each news article is an individual instance, the collection of news articles published on a given day is a bag, and the ordered collection of bags (days) is denoted as a super-bag (explained in detail later). For this study, each individual news article is represented by a distributed representation for text derived using a framework such as text

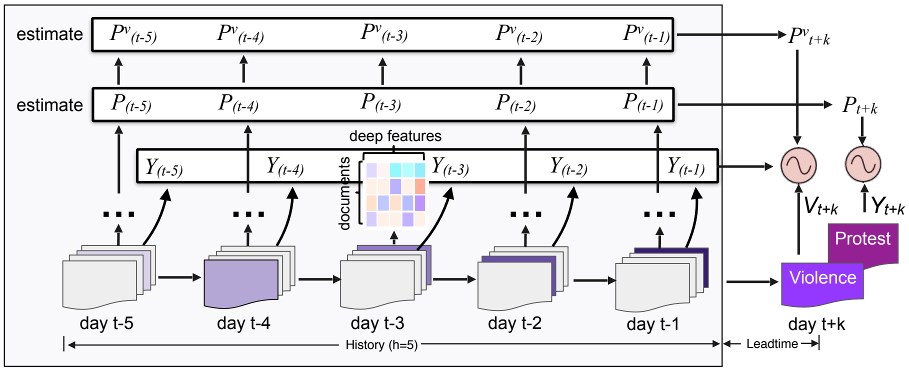


Fig. 3. Overview of proposed approach to event forecasting and precursor discovery.

embedding (Le and Mikolov 2014). Figure 3 shows that for certain days within the collection we attempt to identify news articles (highlighted) that are considered as precursors from the entire collection of input news articles used for forecasting the occurrence of a specific target.

2.1 Formal Definition and Notation

For a given protest event e occurring on day $t+k$, we assume that for each day before the event we are tracking a multitude of news sources. We represent the collection of n_i news articles published on a given day i by $\mathbf{X}_i = \{\mathbf{x_{i,1}} \ldots \mathbf{x_{i,n_i}}\}$, where the j-th news article is represented by $\mathbf{x_{ij}}$. The ordered collection of news articles for the protest event up to day t can be represented as a super-bag, $\mathbb{S}_{1:t} = \{\mathbf{X}_1, \ldots, \mathbf{X}_t\}$. The occurrence of the protest event at time $t+k$ is denoted by $\mathcal{Y}_{t+k} \in \{-1, +1\}$ where 1 denotes a protest and -1, otherwise.

The forecasting problem can be formulated as learning a mathematical function $f(\mathbb{S}_{1:t}) \rightarrow \mathcal{Y}_{t+k}$ that maps the input, an ordered collection of news articles extracted per day to a protest indicator k days in the future from the day t. To identify the news articles considered as precursors (evidence), we aim to estimate a probability for each news article on any given day that signifies the occurrence of a given protest. For a news article $\mathbf{x_{ij}}$, we denote this estimated probability value by p_{ij}. As such, given the collection of news articles we identify the precursor set as the ones with p_{ij} greater than a fixed threshold τ. We represent this precursor set of documents as a subset of the original super-bag, given by $O = \{\mathbf{x_{ij}} \in \mathbb{S}_{1:t} \mid p_{ij} > \tau\}$. As a secondary objective, we aim to forecast the occurrence of an event with a long lead time i.e., large values of k. Table 1 captures the notation and definitions used in this study.

3 METHODS

We first provide our intuition behind formulating the precursor discovery and forecasting problem as a multiple instance learning algorithm. Parallel to the standard multiple instance learning algorithms we have a group of news articles (bags) with labels available only for the entire bag (i.e., leading to a protest); and one of the objectives is to train a classifier to predict the bag-level label. In addition to predicting the group-level labels, we also care about predicting the labels for individual news articles (instances) since they signify the precursor. Various MIL formulations extend the basic definition with a similar motivation, i.e., to estimate the key instances within a bag or provide instance-level labels. However, our problem setting has a two-level grouping structure with sequential constraints, i.e., we capture news articles per day (bags) and group the days to form a super-bag with labels only available at the super-bag level. As such, we propose a nested multiple instance learning formulation for predicting the super bag level labels (forecast) and then estimate the bag-level and instance-level probabilities for identifying association of the bag and instance with the event, respectively. We developed various extensions of our proposed approach to tie the different sequential and group constraints.

3.1 Nested MIL model (nMIL)

We model the instance level probability estimates p_{ij} for a news article j on day i to associate with a targeted event e with a logistic function. These probability estimates indicate how related the specific instance is to the target event, e. Higher the probability value, the more related the document is to the target event and most probably represents a precursor that contains information about causes of the target event.

$$p_{ij} = \sigma(\mathbf{w}^T \mathbf{x}_{ij}) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}_{ij}}}. \tag{1}$$

Here, $\mathbf{w}$ denotes the learned weight vector for our model. The probability for a day (or bag) is then modeled as the average of probability estimates of all instances in a day (Kotzias et al. 2015). Hence, for each bag:

$$P_i = A(\mathbf{X}_i, \mathbf{w}) = \frac{1}{n_i} \sum_{j}^{n_i} p_{ij}, \tag{2}$$

where A is an aggregation function.

We then model the probability of a super-bag $\mathbb{S}$ (associated with an event e) being positive as the average of the probability of all t bags within the super bag to be positive (related to the target event). Thus:

$$P = A(\mathbb{S}, \mathbf{w}) = \frac{1}{t} \sum_{i}^{t} P_i \tag{3}$$

For a given super bag $\mathbb{S}$, as all the t bags within it are temporally ordered, the probability estimates for a given bag (day) is assumed to be similar to its immediate predecessor. This consistency in consecutive bag probabilities is modeled by minimizing the following cross-bag cost as below:

$$g(\mathbf{X}_i, \mathbf{X}_{i-1}) = (P_i - P_{i-1})^2 \tag{4}$$

Finally, given a set of true labels Y for the super bags, we can train our model by minimizing the following cost function w.r.t to $\mathbf{w}$:

$$J(\mathbf{w}) = \frac{\beta}{N} \Bigg[\sum_{\mathbb{S} \in \mathcal{S}} f(\mathbb{S}, Y, \mathbf{w}) + \sum_{\substack{\mathbb{S} \in \mathcal{S}; \\ \mathbf{X}_i, \mathbf{X}_{i-1} \in \mathbb{S}}} \frac{1}{t} \sum_{i=1}^{t} g(\mathbf{X}_i, \mathbf{X}_{i-1}, \mathbf{w}) + \sum_{\substack{\mathbb{S} \in \mathcal{S}; \mathbf{X}_i \in \mathbb{S} \\ \mathbf{x}_{ij} \in \mathbf{X}_i}} \frac{1}{t} \sum_{i=1}^{t} \frac{1}{n_i} \sum_{j=1}^{n_i} h(\mathbf{x}_{ij}, \mathbf{w}) \Bigg] + \lambda R(\mathbf{w}). \tag{5}$$

Here,

- $f(\mathbb{S}, Y, \mathbf{w}) = -\mathbf{I}(Y = 1) \log P - \mathbf{I}(Y = -1) \log(1 - P)$ is the negative log-likelihood function that penalizes the difference between prediction and the true label for super bag $\mathbb{S}$ where $\mathbf{I}(\cdot)$ is the indicator function.
- $g(\mathbf{X}_i, \mathbf{X}_{i-1}, \mathbf{w})$ is the cross-bag cost defined in Equation. 4

- $h(\mathbf{x}_{ij}, \mathbf{w}) = max(0, m_0 - sgn(p_{ij} - p_0)\mathbf{w}^T\mathbf{x}_{ij})$ represents the instance level cost. Here, sgn is the sign function; m_0 is a crucial margin parameter used to separate the positive and negative instances from the hyper line in the feature space; p_0 is a threshold parameter to determine positiveness of instance.
- $R(\mathbf{w})$ is the regularization function.
- β, λ are constants that control the trade-offs between the loss function and regularization function.

*Cross-bag Similarity (**nMIL**$^\Delta$).* The cross-bag similarity $g(,)$ in the above equation does not allow for sudden changes in the day-level probabilities caused due to newer events happening on the current day. We update the cost function across days (bags) (Equation 4) as follows:

$$g(\mathbf{X}_i, \mathbf{X}_{i-1}) = \Delta(\mathbf{X}_i, \mathbf{X}_{i-1})(P_i - P_{i-1})^2 \tag{6}$$

The objective function above allows for label information to spread over the manifold in the feature-space. As such, we compute $\Delta(,)$ as the pairwise cosine similarity between the news articles in $\mathbf{X}_i$ and $\mathbf{X}_{i-1}$. Since we do not have ground truth labels for the bag level (day) we make this consistency assumption that estimated probabilities for consecutive days should be similar if the news articles have similarity in the feature space as well. This model is referred by **nMIL**$^\Delta$and allows for sudden changes in how events unfold.

3.2 Sequential Model (nMIL$^\Omega$)

The basic **nMIL** models assume that there exists a single weight vector across all the days (bags) within a super bag. To model the sequential characteristics of the articles published across consecutive days, we extend this formulation by learning individual weight vectors for each of the historical days. Assuming t days within a super bag $\mathbb{S}$ we learn a weight vector for each individual day represented as $\Omega = [\mathbf{w}_1, \ldots, \mathbf{w}_t]$; where $\mathbf{w}_i$ is the weight vector learned for day i. In this setting, the individual weight vectors are still learned together in a joint fashion akin to multi-task learning approaches (Caruana 1997). However, the probability of a news article j on day i will be given by $p_{ij} = \sigma(\mathbf{w}_i^T \mathbf{x}_{ij})$. This formulation is called **nMIL**$^\Omega$ and given by:

$$J(\Omega) = \frac{\beta}{N}\Big[\underbrace{\sum_{\mathbb{S}\in\mathcal{S}} f(\mathbb{S}, Y, \Omega)}_{\text{empirical loss}} + \underbrace{\sum_{\substack{\mathbb{S}\in\mathcal{S};\\ \mathbf{X}_i, \mathbf{X}_{i-1}\in\mathbb{S}}} \frac{1}{t}\sum_{i=1}^{t} g(\mathbf{X}_i, \mathbf{X}_{i-1}, \mathbf{w}_i)}_{\text{sequential loss}} + \underbrace{\sum_{\substack{\mathbb{S}\in\mathcal{S};\mathbf{X}_i\in\mathbb{S}\\ \mathbf{x}_{ij}\in\mathbf{X}_i}} \frac{1}{t}\sum_{i=1}^{t}\frac{1}{n_i}\sum_{j=1}^{n_i} h(\mathbf{x}_{ij}, \mathbf{w}_i)}_{\text{unsupervised loss}}\Big] + \lambda R(\Omega) \tag{7}$$

Just like the multi-task learning algorithms, the regularization term $R(\Omega)$ can be modified to capture the various relationship-based constraints. However, in this study we ignore these specialized approaches focusing only on the MIL paradigm.

3.3 Optimization

We perform online stochastic gradient decent optimization to solve our cost function and test our model on new data to predict super bag label. For every iteration in our algorithm, we randomly choose a super-bag $(\mathbb{S}, Y)$ from the training dataset $\mathcal{S}$ by picking an index $r \in \{1, \ldots, n\}$ using a standard uniform distribution. Then we optimize an approximation based on the sampled super-bag by:

$$J(\mathbf{w};\mathbb{S}) = \beta\Big[f + \frac{1}{t}\sum_{i}^{t} g_i + \frac{1}{t}\sum_{i}^{t}\frac{1}{n_i}\sum_{j}^{n_i} h_{ij}\Big] + \lambda R(\mathbf{w}) \tag{8}$$

The gradient of the approximate function is given by:

$$\begin{aligned}\nabla J(\mathbf{w}) = \frac{\partial J(\mathbf{w};\mathbb{S})}{\partial \mathbf{w}} &= \lambda \mathbf{w} - \frac{Y-P}{P(1-P)}\frac{\beta}{t}\sum_{i}^{t}\frac{1}{n_i}\sum_{k}^{n_i} p_{ij}(1-p_{ij})\mathbf{x}_{ij} \\ &+ \frac{\beta}{t}\sum_{i}^{t} 2(P_i - P_{i-1})\Big[\frac{1}{n_i}\sum_{j}^{n_i} p_{ij}(1-p_{ij})\mathbf{x}_{ij} - \frac{1}{n_{i-1}}\sum_{j}^{n_{i-1}} p_{\delta j}(1-p_{\delta j})\mathbf{x}_{\delta j}\Big] \\ &- \frac{\beta}{t}\sum_{i}^{t}\frac{1}{n_i}\sum_{j}^{n_i} sgn(p_{ij}-p_0)\mathbf{x}_{ij}(o_{ij})\end{aligned} \tag{9}$$

where $\delta = i-1, o_{ij} = I(sgn(p_{ij}-p_0)\mathbf{w}\mathbf{x}_{ij} < m_0)$. We update the weight vector using a varied learning rate and $\mathbf{w}' = \mathbf{w} - \eta\nabla(\mathbf{w})$ using mini-batch stochastic gradient descent where η is the learning rate at current iteration.

3.4 Violence Forecaster

In this extended problem, there are two categories for the target events: *violent* and *non-violent* events. We introduce a new model parameter, $\mathbf{v}$, for violent crowd events, with the same dimension as $\mathbf{x}_{ij}$. Likewise, the probability of a news article related to a violent event is defined as:

$$p_{ij}^{v} = \sigma(\mathbf{v}^T\mathbf{x}_{ij}) = \frac{1}{1+e^{-\mathbf{v}^T\mathbf{x}_{ij}}} \tag{10}$$

The probability of a violent protest event is modeled as a joint probability of violence and protest. Applying the Bayes rule, we get:

$$\gamma = P(V=1,Y=1) = P(V=1|Y=1)\times P(Y=1) = \Big(\frac{1}{t}\sum_{i=1}^{t}\frac{1}{n_i}\sum_{j}^{n_i} p_{ij}^{v}\Big)\Big(\frac{1}{t}\sum_{i=1}^{t}\frac{1}{n_i}\sum_{j}^{n_i} p_{ij}\Big) \tag{11}$$

Given a set of true labels Y (protest) and V (violent protest) for the super bags, we also know if any protest event occurs ($Y_i = 1/0$) on each history day i in the same city before the target event. We can train our model, $\mathbf{nMIL^{vio}}$, by minimizing the following cost function with respect to $\mathbf{w}$ and $\mathbf{v}$ as:

$$\begin{aligned}\min_{\mathbf{w},\mathbf{v}} \frac{\beta}{N}\sum_{\mathbb{S}\in\mathcal{S}} \Big[L(\mathbf{w}|\mathbb{S}) + L(\mathbf{v}|\mathbb{S}) + (\gamma - \frac{1}{t}\sum_{i=1}^{t} Y_i)^2 + \frac{1}{t}\sum_{i=1}^{t} g(\mathbf{w},\mathbf{X}_i,\mathbf{X}_{i-1}) + \frac{1}{t}\sum_{i=1}^{t}\frac{1}{n_i}\sum_{j=1}^{n_i} h(\mathbf{x}_{ij},\mathbf{w})\Big] \\ + \frac{\lambda_1}{2}||\mathbf{w}||^2 + \frac{\lambda_2}{2}||\mathbf{v}||^2\end{aligned} \tag{12}$$

where $L(*)$ represents the negative log-likelihood loss, $g(*) = (P_i - P_{i-1})^2$ is a squared loss function for two consecutive days requiring two days' probabilities to be close, and $h(*) = \max(0, m_0 - sgn(p_{ij}-p_0)\mathbf{w}^T\mathbf{x}_{ij})$ is the hinge loss function at the instance level.

$$L(\mathbf{w}|\mathbb{S}) = -\sum_{m=0}^{1} I(Y=m)\log P(Y=m) \tag{13}$$

In particular, violence is an attribute or a result of protest events. Protest events can end either peacefully or violently. The probability of a violent protest event is conditional on the probability of a protest event. Thus the negative log-likelihood for violent events is calculated as follows:

$$L(\mathbf{v}|\mathbb{S}) = -I(V=1,Y=1)\log P(V=1,Y=1) - I(V=0,Y=1)\log P(V=0,Y=1) \tag{14}$$

From our observations, violent protest events usually follow a sequence of protest events. Thus, the probabilities of violent protest events are assumed to be highly related with the number of protest events that occur before these violent events ($(\gamma - \frac{1}{t}\sum_{i=1}^{t} Y_i)^2$). For instance, if there are seven protest events in the same city in one week, it is highly probable that some protest will turn to violence because the growing anger and dissatisfaction tend to make people resort to violence. We develop an alternating minimization algorithm that can be applied to achieve a solution of Eq. 12. Specifically, when $\mathbf{w}$ is fixed, all $\mathbf{v}$'s can be solved by a stochastic gradient descent algorithm as:

$$\Delta(\mathbf{v}) = \beta\Big[-\frac{V-\gamma}{\gamma(1-\gamma)}\frac{\partial\gamma}{\partial\mathbf{v}} + 2(\gamma - \frac{1}{t}\sum_{i=1}^{t} Y_i)\frac{\partial\gamma}{\partial\mathbf{v}}\Big] + \lambda_2\mathbf{v} \tag{15}$$

Likewise, when $\mathbf{v}$ is fixed, all $\mathbf{w}$'s derivatives have a formulation as follows:

$$\begin{aligned}\Delta(\mathbf{w}) =&\beta\Big[-\frac{Y-P}{P(1-P)}\frac{\partial P}{\partial\mathbf{w}} - \frac{V-\gamma}{\gamma(1-\gamma)}\frac{\partial\gamma}{\partial\mathbf{w}} + 2(\gamma - \frac{1}{t}\sum_{i=1}^{t} Y_i)\frac{\partial\gamma}{\partial\mathbf{w}}\\ &+\frac{1}{t}\sum_{i=1}^{t} 2(P_i - P_{i-1})(\frac{\partial P_i}{\mathbf{w}} - \frac{\partial P_{i-1}}{\mathbf{w}}) - \frac{1}{t}\sum_{i=1}^{t}\frac{1}{n_i}\sum_{j=1}^{n_i} sgn(p_{ij}-p_0)\mathbf{x}_{ij}I_{ij}\Big] + \lambda_1\mathbf{w}\end{aligned} \tag{16}$$

where I_{ij} is the indicator of $(sgn(p_{ij}-p_0)\mathbf{w}^T\mathbf{x}_{ij} \le m_0)$. We alternatively updating $\mathbf{w}$ and $\mathbf{v}$ via gradient descent toward convergence. A complete algorithm is described in Algorithm 1.

3.5 Multiclass Classification

We also extend our developed **nMIL** formulations to solve general purpose multi-class classification problems (**nMIL-MC**) rather than binary classification problems. Within our domain, each labeled event is manually attached with a categorical attribute (event population) denoting the subgroup or community of people who participated in the protest event (e.g., teachers, doctors, etc.)

For multiclass classification problems, we train one-versus-rest classifiers for each of the classes learning a separate weight vector per class. When classifying a super bag to a specific event population we first forecast the binary protest indicator label for a super bag. Next, we apply the multi-class classification only on the predicted positive examples.

3.6 Precursor discovery

In the **nMIL** models, each super bag consists of an ordered set of bags and each bag represents the documents in one day in the city for which we are forecasting a protest event. We present in Algorithm 2 the steps to identify news articles as precursors based on their estimated probability given by $\hat{y}_{tm} > \tau_1$ (protest), $\hat{v}_{tm} > \tau_2$ (violent protest).

Algorithm 1 Model Learning for **nMIL** and **nMIL$^{\mathbf{vio}}$**

procedure **nMIL** MODEL LEARNING
  Input: training examples, $\mathcal{S} = \{(\mathbb{S}_r, Y_r)\}$.
  Output: model parameter, $\mathbf{w}$.
  Initialize $\mathbf{w}$
  for $\tau = 1$ to T **do**
    for super bag $(\mathbb{S}_r, Y_r)$ **do**
      update $\mathbf{w}$ ▷ Solving Eq. 9
  return $\mathbf{w}$
procedure **nMIL$^{\mathbf{vio}}$** MODEL LEARNING
  Input: training examples, $\mathcal{S} = \{(\mathbb{S}_r, Y_r, V_r)\}$
  Output: model parameters, $\mathbf{v}$
  Pre-compute $\{Y_i | (i = t - H, ..., t)\}$ for each event.
  Initialize $\mathbf{w}, \mathbf{v}$
  for $\tau = 1$ to T **do**
    for super bag $(\mathbb{S}_r, Y_r, V_r)$ **do**
      Fix $\mathbf{v}$, update $\mathbf{w}$ ▷ Solving Eq. 16
      Fix $\mathbf{w}$, update $\mathbf{v}$ ▷ Solving Eq. 15
  return $\mathbf{v}$

Algorithm 2 Precursor Discovery in **nMIL** and **nMIL$^{\mathbf{vio}}$**

Input:testing examples, $\mathcal{S} = \{(\mathbb{S}_r, Y_r, V_r)\}_{r \in n^+}$
Output: precursors for protests and violent protests: $\{(O_r, O_v)\}_{r \in n^+}$
$\hat{\mathbf{w}} \leftarrow$ **nMIL** model learning, $\hat{\mathbf{v}} \leftarrow$ **nMIL$^{\mathbf{vio}}$** model learning
for super bag $(\mathbb{S}_r, Y_r, V_r)$ **do**
  $O_r = [], O_v = []$
  for t = 1,2,...,h (history days) **do**
    $p_t = [], p_t^v = []$
    for $\mathbf{x}_{tm} \in \mathrm{X}_t$ **do**
      $\hat{y}_{tm} = \sigma(\hat{\mathbf{w}}^T \mathbf{x}_{tm}), \hat{v}_{tm} = \sigma(\hat{\mathbf{v}}^T \mathbf{x}_{tm})$
      if $\hat{y}_{tm} > \tau_1$ **then**
        $p_t \leftarrow (m, \hat{y}_{tm})$
      if $\hat{v}_{tm} > \tau_2$ **then**
        $p_t^v \leftarrow (m, \hat{v}_{tm})$
    sort (p_t, p_t^v) in descending order
    $O_r, O_v \leftarrow$ m where m in top(p_t), top(p_t^v) by threshold τ_1, τ_2
return $\{(O_r, O_v)\}_{r \in n^+}$

4 EXPERIMENTAL EVALUATION

4.1 Datasets

The experimental evaluation was performed on news documents collected from around 6000 news agencies between July 2012 to December 2014 across three countries in South America, viz. Argentina, Brazil, and Mexico. For Argentina and Mexico, the input news articles were primarily in Spanish and for Brazil, the news articles were in Portuguese.

The ground truth information about protest events, called the gold standard report (GSR) is exclusively provided by MITRE (Ramakrishnan et al. 2014). The GSR is a manually created list of civil unrest events that happened during the period 2012-2014. A labeled GSR event provides information about the geographical location at the city level, date, type and population of a civil unrest news report extracted from the most influential newspaper outlets within the country of interest. These GSR reports are the target events that are used for validation of our forecasting algorithm. We have no ground truth available for verifying the validity of the precursors.

Argentina: We collected data for Argentina from newspaper outlets including *Clarin* and *Lanacion* for the period of July 2010 to December 2014. There are multiple protest events in Argentina during this period. For instance, people protested against the government and utility/electricity-providing companies because of heatwaves in Dec. 2013.

Brazil: For Brazil, we obtained data from news agencies including the three leading news agencies in Brazil; *O Globo, Estadao,* and *Jornal do Brasil* from November 2012 to September 2013. During this period Brazil faced several mass public demonstrations across several Brazilian cities stemming from a variety of issues ranging from transportation costs, government corruption, and police brutality. These mass protests were initiated due to a local entity advocating for free public transportation. This period had an unusually high social media activity and news coverage and is also known as the "Brazilian Spring" [1].

Mexico: For Mexico, we tracked news agencies including the top outlets: *Jornada, Reforma, Milenio* from January 2013 to December 2014. Over 619 days, we noticed 71 news articles per day on average. There were more than 2000 protest events in this two-year period with major unrest movements in 2013 led by teachers and students demanding education reforms by protesting against the government.

For $\mathbf{nMIL^{vio}}$ model evaluation on violent protest events, we collect news from top agencies for these five countries based on the distribution of violent protest events: Argentina, Brazil, Colombia, Paraguay, and Venezuela.

4.2 Experimental Protocol

The GSR signifies the occurrence of a protest event on a given day at a specific location. To evaluate the MIL-based forecasting and precursor discovery algorithms, for each protest event we extract all the published news articles for up to 10 days before the occurrence of the specific event. This ordered collection of per-day news documents up to the protest day are considered as positive super bags. For negative samples, we identify consecutive sets of five days within our studied time periods for the different countries when no protest was reported by the GSR. The ordered collection of per-day news documents not leading to a protest are considered as negative super bags for the nMIL approach. For any news article (i.e., an individual instance) within a positive/negative super-bag we have no label (or ground truth). As part of the precursor discovery algorithm, we estimate a probability for an individual instance to signal a protest (by showing evidence). It is important to note that the GSR linked news article for a protest is never used for training purposes. Having identified the positive and negative samples, we split our datasets into training and testing partitions and perform 3-fold cross-validation. A single run of the model on a machine with 4 cores and 16 GB memory takes about 250 seconds.

We study the performance of forecasting models with varying lead time and varying historical days. Lead time (l) indicates the number of days in advance the model makes predictions and historical days (h) denotes the number of days over which the news articles are extracted as input to the prediction algorithms. As an example, if l is set to 1, then the model forecasts if a protest event is planned for the next day. Setting the historical days, h, to 5 denotes that we use news from five days before the current day to make the forecast. We varied l from 1 to 5 and h from 1 to 10 and trained 50 different models for the different approaches to study the characteristics of the developed approaches

[1] http://abcnews.go.com/ABC_Univision/brazilian-spring-explainer/story?id=19472387

with varying lead time and historical days. For event forecasting, we evaluate the performance by standard metrics including precision, recall, accuracy and F1-measure.

4.3 Comparative Approaches

We compare the proposed **nMIL** models to the following approaches:

- **SVM**: We use the standard support vector machine formulation (Cortes and Vapnik 1995) by collapsing the nested grouping structure and assigning the same label for each news article as its super-bag (for training). During the prediction phase, the SVM yields the final super-bag prediction (forecast) by averaging the predicted label obtained for each of the instances.
- **MI-SVM** (Andrews et al. 2002): The MI-SVM model extends the notion of a margin from individual patterns to bags. Notice that for a positive bag the margin is defined by the margin of the "most positive" instance, while the margin of a negative bag is defined by the "least negative" instance. In our case, we collapse the news articles from the different historical days into one bag and apply this standard MIL formulation.
- **Relaxed-MIL** ($\mathbf{rMIL^{nor}}$) (Wang et al. 2015): Similar to the MI-SVM baseline, we collapse the news articles into one bag. However, unlike the MI-SVM formulation the $\mathbf{rMIL^{nor}}$ model can provide a probabilistic estimate for a given document within a bag to be positive or negative.
- **Modified Relaxed-MIL** ($\mathbf{rMIL^{avg}}$): This approach is similar to the $\mathbf{rMIL^{nor}}$, except we compute the probability of a bag being positive by taking average of estimate of each instance in the bag rather than using the Noisy-OR model discussed above.
- **GICF** (Kotzias et al. 2015): This model optimizes a cost function which parameterized the whole-part relationship between groups and instances and pushes similar items across different groups to have similar labels.

4.4 Feature Description

In practice, finding good feature representations to model the news articles is not a trivial problem. Traditionally the bag-of-words representation allows for easy interpretation but also requires pre-processing and feature selection. Several researchers have developed efficient and effective neural network representations for language models (Bengio et al. 2003; Mikolov et al. 2013a,b). Specifically, we learn deep features for documents by taking advantage of the existing doc2vec model. For each document, we generate a 300 dimension distributed representation for training with a contextual window size of 10 in an unsupervised version (Le and Mikolov 2014). We compared the performance of deep features with traditional TF-IDF features but the results showed little difference. Thus, we only report the evaluation of models with deep features.

4.5 Violence Alert Metrics.

The forecasting alerts for violence generated by the model and the real events are structured records as follows:

```
Date:2014-07-01, Type: Violent protest.
Location: Brazil, Sao Paulo, Sao Paulo.
```

The quality score for a forecast involved evaluations based on time and location given by:

$$\mathsf{QS} = 2 * (\mathsf{DS} + \mathsf{LS})$$

where DS, LS denote the date score and location score respectively.

$$\mathsf{DS} = 1 - \min(|d_e - d_p|, 7)/7$$

Table 2. Event forecasting performance comparison based Accuracy (Acc) and F-1 score w.r.t to state-of-the-art methods. The proposed nMIL, nMIL$^{\Delta}$, nMIL$^{\Omega}$ method outperform state-of-the-art methods across the three countries.

	Argentina		Brazil		Mexico	
	Acc	F-1	Acc	F-1	Acc	F-1
SVM	0.611(±0.034)	0.406(±0.072)	0.693(±0.040)	0.598(±0.067)	0.844(±0.062)	0.814(±0.091)
MI-SVM	0.676(±0.026)	0.659(±0.036)	0.693(±0.040)	0.503(±0.087)	0.880(±0.025)	0.853(±0.040)
rMIL^{nor}	0.330(±0.040)	0.411(±0.092)	0.505(±0.012)	0.661(±0.018)	0.499(±0.009)	0.655(±0.025)
rMIL^{avg}	0.644(±0.032)	0.584 (±0.055)	0.509(±0.011)	0.513(±0.064)	0.785(±0.038)	0.768(±0.064)
GICF	0.589(±0.058)	0.624(±0.048)	0.650(±0.055)	0.649 (±0.031)	0.770(±0.041)	0.703(±0.056)
nMIL	**0.709**(±0.036)	0.702(±0.047)	**0.723**(±0.039)	0.686(±0.055)	**0.898**(±0.031)	**0.902**(±0.030)
nMIL$^{\Delta}$	0.708(±0.039)	**0.714**(±0.034)	0.705(±0.048)	**0.698**(±0.045)	0.861(±0.014)	0.868(±0.014)
nMIL$^{\Omega}$	0.687(±0.038)	0.680(±0.045)	0.713(±0.028)	0.687(±0.038)	0.871(±0.013)	0.879(±0.014)

where d_e is the event date and d_p is the predicted date for the event. If the predicted date of the event is the same as the actual date of the event, then DS is 1.

Location score has many ways of definition. In our problem, location is in terms of a triples of (country, state, city). Comparing a true event with a predicted event, we obtain a score at these three levels:

$$\mathsf{LS} = \frac{1}{3}l_1 + \frac{1}{3}l_1 l_2 + \frac{1}{3}l_1 l_2 l_3$$

where l_1 is the country-level score, l_2 is the state level score, and l_3 is the city level score. We selected a set of cities based on their scales. Then we built training and testing examples for these cities and the location score is only calculated for the selected cities.

Other typical evaluation metrics for classification such as accuracy (ACC) and area under curve (AUC) score are also used in our experiments. True positive examples are the true violent events and the model predicts correctly. True negative examples are the true non-violent events or no-event and model predicts correctly.

5 RESULTS

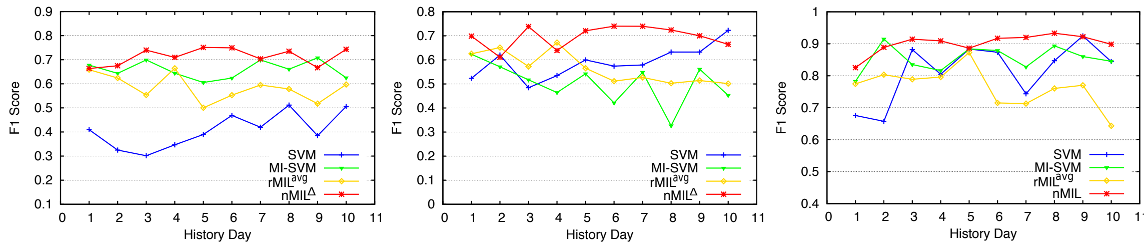
In this section, we evaluate the performance of the proposed models to answer a few questions as follows:

- How well can the proposed methods forecast protest and violent protest events?
- How early can the proposed methods forecast protest and violent protest events?
- How well can the **nMIL** model forecast event populations?
- How sensitive is the **nMIL** model to the hyper parameters?
- How good are the precursors? Do they tell a story for the target events?
- Why do crowds turn violent?

We first show the results for the basic **nMIL** model for protest events and then the **nMIL^{vio}** model for violent protest events.

5.1 How well does nMIL forecast protests?

5.1.1 Comparative Evaluation. Table 2 reports the prediction performance of the **nMIL** approach in comparison to other baseline approaches for the task of forecasting protests. Specifically, we use set $\beta = 3.0$, $\lambda = 0.05$, $m_0 = 0.5$ and $p_0 = 0.5$ (β, λ chosen by sensitivity analysis, m_0, p_0 by default setting in hinge loss) and report the average accuracy



(a) Argentina **(b)** Brazil **(c)** Mexico

Fig. 4. Forecasting evaluation on 3 countries with respect to F1 score for SVM, rMIL^{nor}, rMIL^{avg}, and nMIL. X-axis is the number of historical days used in the training process. Y-axis shows the average F1 score of 10 runs of experiments.

Table 3. F1-measure for rMIL^{avg} and nMIL$^{\Delta}$ models on Argentina, Brazil, and Mexico with history days from 1 to 5.

	Country	Argentina					Brazil					Mexico				
	History Days	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
Leadtime 1	**rMIL^{avg}**	0.719	0.714	0.690	0.710	0.705	0.717	0.692	0.696	0.662	0.680	0.815	0.803	0.789	0.796	0.873
	nMIL$^{\Delta}$	**0.745**	**0.735**	**0.722**	0.691	**0.716**	**0.734**	**0.768**	**0.721**	**0.735**	**0.717**	**0.842**	**0.868**	**0.863**	**0.884**	**0.884**
Leadtime 2	**rMIL^{avg}**	0.659	0.624	0.554	0.665	0.500	0.695	0.651	0.573	0.672	0.565	0.846	0.875	0.860	0.878	0.912
	nMIL$^{\Delta}$	**0.664**	**0.675**	**0.740**	**0.710**	**0.751**	**0.699**	0.611	**0.738**	0.639	**0.721**	0.825	**0.889**	**0.914**	**0.909**	0.886
Leadtime 3	**rMIL^{avg}**	0.674	0.606	0.622	0.543	0.578	0.694	0.682	0.620	0.715	0.622	0.819	0.787	0.808	0.750	0.853
	nMIL$^{\Delta}$	0.649	**0.669**	0.560	**0.669**	**0.737**	**0.687**	0.639	**0.674**	**0.717**	**0.742**	**0.856**	**0.903**	**0.884**	**0.909**	**0.900**
Leadtime 4	**rMIL^{avg}**	0.656	0.558	0.588	0.556	0.476	0.729	0.712	0.720	0.628	0.621	0.809	0.822	0.798	0.878	0.772
	nMIL$^{\Delta}$	**0.676**	**0.693**	**0.670**	**0.712**	**0.631**	**0.754**	0.584	**0.736**	**0.735**	**0.725**	**0.872**	**0.888**	**0.894**	**0.916**	**0.874**
Leadtime 5	**rMIL^{avg}**	0.669	0.676	0.590	0.567	0.575	0.710	0.588	0.616	0.548	0.570	0.828	0.845	0.810	0.733	0.889
	nMIL$^{\Delta}$	0.626	0.676	**0.687**	**0.773**	**0.737**	0.683	**0.665**	**0.657**	**0.697**	**0.735**	**0.833**	**0.937**	**0.878**	**0.935**	**0.931**

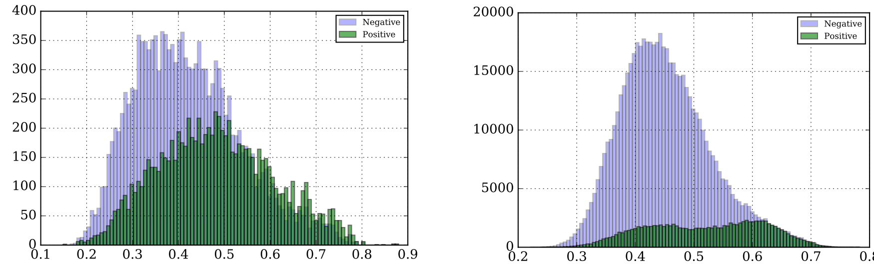
and F1 score along with standard deviation for predicting protests across multiple runs of varying historical days with lead time set to 1.

We observe that the **nMIL** approaches outperform the baseline approaches across all the three countries. The **rMIL^{nor}** approach performs poorly because the the noisy-OR aggregation function associating the bag-level labels to instance-level labels forces most of the news articles within the positive bags to have probability values close to 1. However, given the large collection of news articles available per day only a subset of them will provide a signal/evidence for a protest. For Argentina, the **nMIL** and **nMIL$^{\Delta}$** approaches outperformed the best baseline (**MI-SVM**), by **7%** and **8%** with respect the average F1 score, respectively.

Figure 4 shows the changes to F1 score for the proposed **nMIL** approach in comparison to **SVM**, **MI-SVM** and **rMIL^{avg}** for different number of historical days that are used in training with lead time set to 2. We trained 10 different models that use different number of historical days respectively varying from 1 to 10. These results show that the methods that utilize the nested structure (**nMIL**, **nMIL$^{\Delta}$**) within the multi-instance learning paradigm generally performed better than others. Moreover, the proposed **nMIL** models performed well consistently across different countries with different number of history days.

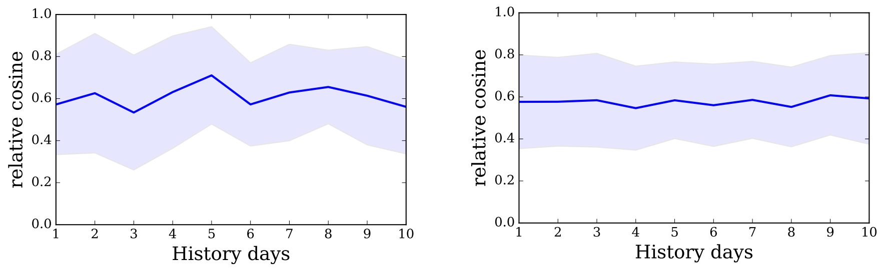
5.2 How early can nMIL forecast?

In order to study the changes of performance with and without the nested structure, we show the F1 score with varying lead times and historical days from 1 to 5 for **rMIL^{avg}** and **nMIL$^{\Delta}$** models in Table 3, respectively.



(a) Argentina **(b)** Mexico

Fig. 5. Estimated probabilities for negative examples (purple) and positive examples (green) for Argentina and Mexico.



(a) Argentina **(b)** Mexico

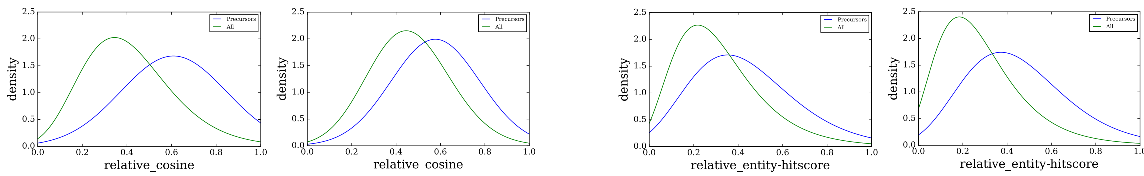
Fig. 6. Mean of relative cosine values w.r.t target events in history days for Argentina and Mexico.

We observe that with larger lead time (i.e., forecasting earlier than later), the **nMIL** model does not necessarily lose forecasting accuracy, but is sometimes even better. This can be explained by the fact that several times protests are planned a few days in advance and that civil unrest unfold as a series of actions taken by multiple participating entities over a sequence of days. As the lead time increases, F1 score for forecasting initially drops and then increases back. This behavior is also noted in prior work by Ramakrishnan et. al (Ramakrishnan et al. 2014), which includes protest related data from these countries. In comparison to the **nMIL** model, the **rMIL$^{\mathbf{avg}}$** approach, (which collapses the sequential structure encoded within the history of days) seems to perform inconsistently with increasing lead time.

5.3 Do the precursors tell a story?

5.3.1 Quantitative Evaluation. Figures 5a and 5b show the distribution of the estimated probabilities for instances within positive and negative super bags for Argentina and Mexico, respectively. The instances within the negative super bags show lower probability estimates by the proposed model and the instances within the positive super bags show higher probability estimates. For Mexico, fewer instances within the positives are assigned high probabilities indicating strength of the proposed model to identify and rank the precursors.

Relative cosine similarity is computed as the pairwise normalized cosine similarity, scaled relative to each event. Figures 6a and 6b show the average cosine similarity value for the precursor documents (probability estimate greater than 0.7) with the target GSR documents. For Argentina, we observe that on average, the documents on day 5 have the highest semantic similarity to the target event documents (GSR). The documents on day 3 and day 10 have lower similarity compared to the target event.



(a) Argentina **(b)** Mexico **(c)** Argentina **(d)** Mexico

Fig. 7. The figures on top depict the distribution of relative cosine similarity for all documents (green line) and for precursor documents (blue line) with probability greater than 0.7. The figures in the bottom row depict the distribution of relative entity hit score for all documents (green line) and for precursor documents (blue line) with probability greater than 0.7.

In order to investigate the relationship between the semantic similarity and the estimated probability by the proposed models, we compare the distribution of relative cosine similarity and relative entity hit score of the precursor documents with the target GSR documents with respect to bag of words features. Entity words in each news document are extracted by an enrichment tool for natural language processing. The relative entity hit score is calculated as the the intersection of entity set of precursor document and the target event divided by the relative minimum length of these two sets.

Figures 7a and 7b show the fitted Gaussian distribution of relative cosine similarities for all documents (green lines) and precursor documents (blue lines) for Argentina and Mexico, respectively. Figures 7c and 7d show the distribution of relative entity hit score for Argentina and Mexico, respectively. These distribution figures demonstrate that the proposed model assigns higher probability to news articles with higher semantic similarity to the GSR articles representing the protests events. These results show the strength of our proposed models in identifying the precursor articles.

Case Studies. We present findings about the identified precursors based on the probability estimate by **nMIL** across three observed protests. In Figure 1, we present a protest event against government in Argentina, and the selected precursors before its occurrence with their estimated probabilities. The titles of news reports as precursors are shown in the timeline.

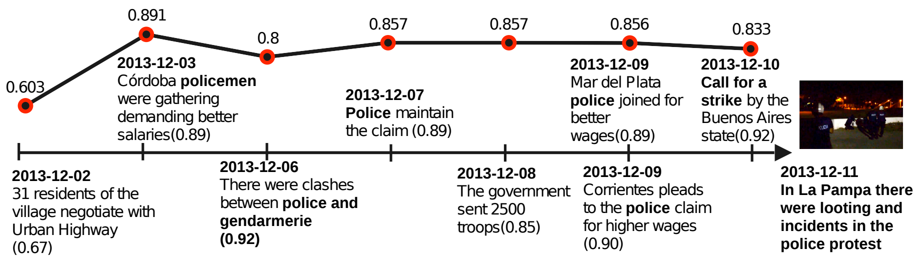
In Figures 8a and 8b, we present story lines by precursors that were discovered for two different protest events in Argentina and Mexico, respectively. Figure 8a showcases the story line about a protest event in Argentina in December 2014. In this case, the police were protesting against government for better salaries. Before this event, clashes between police and gendarmerie (military policy) had occurred leading to the involvement of several policemen from different parts of the country. The text from news articles demonstrate the tense situation between the police and government in La Pampa, Argentina identified as precursors.

Figure 8b shows another story line of a continuous protest event in Mexico regarding the infamous case of 43 missing students [2]. The resulting outrage triggered constant protests which were identified by our proposed model. The figure shows a timeline of how the events turned violent leading up to the burning of congressional offices and depicts how different communities joined the movement.

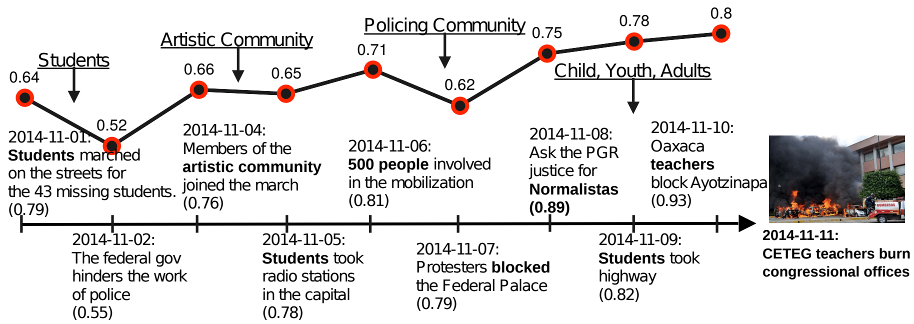
5.4 Can nMIL forecast event populations?

We also evaluated the performance of our **nMIL** approaches for predicting the event populations by solving a multi-class classification problem. In Table 4 we depict the weighted-average F1 score for event populations (here, with

[2] https://en.wikipedia.org/wiki/2014_Iguala_mass_kidnapping



(a) A continuous police protest in Argentina against government for better salary. In the beginning, policemen at Cordoba were requesting for better salaries. Later on, police in Catamarca were involved in clashes with gendarmerie. Three days before the target event, the government sent out troops and more and more police joined for the same purpose. One day before the event, Buenos Aires state call for a strike.



(b) Protester in Mexico burned the congressional offices for justice for the missing students. In the beginning, students were marching for justice. Gradually, more communities such as artistic and policing community joined the event. Later on, children, youth, adults, students and teachers blocked traffic for protest.

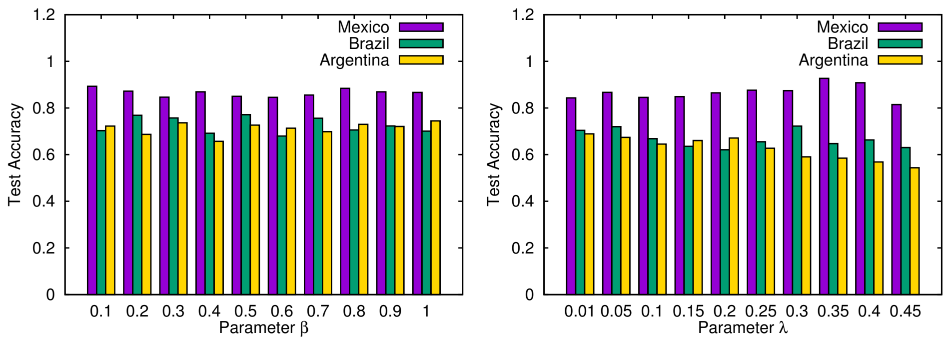
Fig. 8. Case Studies.

Table 4. Multi-Class F1-Measure for $rMIL^{avg}$ and nMIL models on Argentina and Mexico with historical days from 1 to 5.

	History Days	1	2	3	4	5	Average(Variance)
Argentina	**$rMIL^{avg}$**	0.512	0.512	0.473	0.417	0.457	0.474(1e-3)
	nMIL	0.523	0.552	0.515	0.485	0.537	**0.524(7e-4)**
Mexico	**$rMIL^{avg}$**	0.576	0.526	0.447	0.547	0.493	0.518(3e-3)
	nMIL	0.570	0.583	0.560	0.615	0.545	**0.575(7e-4)**

categories such as *Government, Wages, Energy, Others* drawn from the GSR). Due to space limitations, we only depict the performance of weighted average F1 score on event population across 1 to 5 historical days with lead time of 1.

The proposed multi-class **nMIL** model outperforms the multi-class **$rMIL^{avg}$** model. On average, for event population, **nMIL** outperformed **$rMIL^{avg}$** by 10.5% and 10.6% for Argentina and Mexico, respectively.



(a) β in **nMIL** **(b)** λ in **nMIL**

Fig. 9. Sensitivity analysis on β and λ. x-axis represents the varying values for the parameter and y-axis denotes the test accuracy.

Table 5. Violent event forecasting performance comparison based on Accuracy (ACC) and AUC score w.r.t to state-of-the-art methods. The proposed nMIL^{vio} method outperforms state-of-the-art methods across the five countries with 2 weeks historical data.

	Argentina		Brazil		Colombia		Paraguay		Venezuela	
	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC	ACC	AUC
MI-SVM	0.3065	0.5719	0.2210	0.5152	0.3170	0.5635	0.2808	0.5093	0.3058	0.5179
rMIL^{avg}	0.6312	0.5515	0.4617	0.5182	0.6312	0.5515	0.5694	0.5965	0.4376	0.4889
nMIL	0.6688	0.5682	0.5520	0.5223	0.6799	0.6452	0.6498	0.5890	0.5592	0.5698
nMIL-3C	0.5435	0.5508	0.2949	0.5227	0.6263	0.6637	0.5306	0.5869	0.2344	0.4569
nMIL^{vio}	**0.8043**	**0.7123**	**0.7620**	**0.5396**	**0.7911**	**0.6808**	**0.8924**	**0.6607**	**0.5835**	**0.5940**
MI-SVM	baseline	baseline	baseline	baseline	baseline	baseline	baseline	baseline	baseline	baseline
rMIL^{avg}	105.96%	-3.57%	108.90%	0.58%	99.13%	-2.13%	102.81%	17.12%	43.09%	-5.61%
nMIL	118.25%	-0.65%	149.76%	1.38%	114.51%	14.50%	131.46%	15.66%	82.85%	10.03%
nMIL-3C	77.37%	-3.69%	33.41%	1.46%	97.61%	17.78%	88.99%	15.23%	-23.36%	-11.79%
nMIL^{vio}	**162.46%**	**24.54%**	**244.76%**	**4.74%**	**149.58%**	**20.81%**	**217.87%**	**29.72%**	**90.80%**	**14.70%**

5.5 How sensitive is nMIL to parameters?

There are three main parameters in the proposed **nMIL** model, which are the regularization parameter λ, weight for super bag loss β and threshold for instance level hinge loss m_0. Figures 9a and 9b illustrate the performance of the proposed **nMIL** by varying β and λ, respectively. The test accuracy for different values of λ and β is relatively stable.

5.6 How well does the nMIL^{vio} model achieve the overall accuracy for violence?

Table 5 lists the comparative performance in terms of Accuracy (ACC) and AUC scores for five countries in South America. In the bottom part of the table, the lift/drop percentage of the proposed model is presented comparing to the **MI-SVM** as a baseline. Here **nMIL-MC** is the model that consider three classes, *1). violent protest, 2). non-violent protest, 3). no protest*, in the nMIL framework. The accuracy computes the fraction of correct predictions for both positive and negative classes. The AUC is a common evaluation metric for binary classification problems with imbalance. It will be close to 1 when the true positive rate increases quickly. The proposed model outperforms other state-of-the-art methods for all datasets with lead time equal to one day. With a portion of training samples changing from 10% to 100%,

Table 6. AUC scores of the proposed method and the best baseline with lead time from 1 to 4 for violent events.

Dataset	Method	Lead time 1	2	3	4
Argentina	**nMIL**	0.568	0.608	0.610	0.656
	nMIL^{vio}	**0.712**	**0.674**	**0.646**	**0.689**
Brazil	**nMIL**	0.522	0.519	0.507	0.573
	nMIL^{vio}	**0.540**	**0.584**	**0.540**	**0.613**
Colombia	**nMIL**	0.645	0.549	0.693	0.627
	nMIL^{vio}	**0.681**	**0.619**	**0.735**	0.614
Paraguay	**nMIL**	0.589	0.670	0.596	0.593
	nMIL^{vio}	**0.661**	**0.758**	**0.635**	**0.692**
Venezuela	**nMIL**	0.570	0.597	0.609	0.563
	nMIL^{vio}	**0.594**	**0.628**	**0.642**	**0.588**

Table 7. Quality scores of the proposed method and the delivery from online system, EMBERS (Ramakrishnan et al. 2014) for violent events. DS and LS are over 1; QS is over 4.

Dataset	Methods	DS	LS	QS
Argentina	**EMBERS**	0.83	0.69	3.04
	nMIL^{vio}	**0.99**	**0.93**	**3.84**
Brazil	**EMBERS**	0.85	0.81	3.32
	nMIL^{vio}	**0.99**	**0.99**	**3.96**
Colombia	**EMBERS**	0.82	0.75	3.14
	nMIL^{vio}	**0.94**	**0.99**	**3.86**
Paraguay	**EMBERS**	0.89	0.76	3.3
	nMIL^{vio}	**0.95**	**1**	**3.9**
Venezuela	**EMBERS**	0.82	0.8	3.24
	nMIL^{vio}	**0.93**	**0.99**	**3.84**

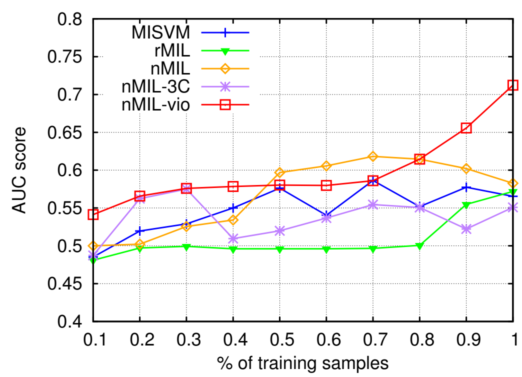


Fig. 10. AUC scores of models for the Argentina dataset.

Figure 10 shows the AUC scores for the **nMIL^{vio}** model and other state-of-the-art models for Argentina dataset. Given the space limitation, we only show this result on Argentina. In general, the AUC scores increase when the number of training samples is increased. With the full set of training samples, the proposed **nMIL^{vio}** model outperforms **MI-SVM**, **rMIL^{avg}**, and **nMIL** methods by 26%, 25% and 22%, respectively.

5.7 How early does the nMIL^{vio} model predict for violence?

Table 6 shows the AUC performance of the proposed method **nMIL^{vio}** in comparison to the best baseline method, **nMIL**, with lead time varying from 1 to 4. For each value of the lead time we train a model where X_r is a super bag containing t historical days and Y_r indicates if a violent protest event happened on day $t + k$. Notice that lead time k indicates the model predicts k days in advance. For Brazil, Colombia, Paraguay and Venezuela, it is noted that a shorter lead time (k=1) does not necessarily imply a better predictive performance compared to a longer lead time (k=3, 4).

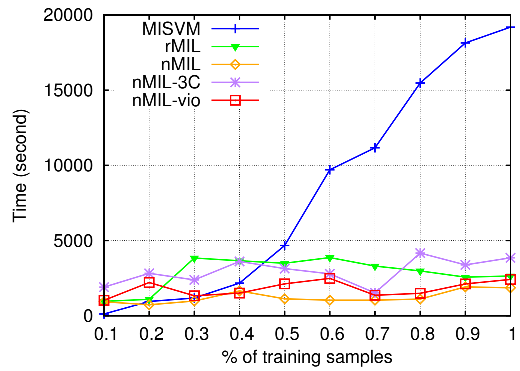


Fig. 11. Computation time of models for the Argentina dataset.

5.8 How does the nMIL$^{\text{vio}}$ fare compared to the online system, EMBERS?

Table 7 presents the performance with respect to quality scores for **EMBERS** delivered system and the proposed methods. The proposed model, **nMIL$^{\text{vio}}$**, outperforms the online system in event data and event location scores for all the five datasets with an average performance improvement of 21%.

5.9 Why do crowds turn violent?

Table 8 presents case studies on two violent protests and one non-violent protest. The detected precursor events are reported before these protest events. We select each precursor event by setting a threshold for its probabiliy $p_{ij}^{v} > 0.5$. The top words are selected from the precursor document of that day based on their frequency. We can make three observations about these results.

First, the occurrence of keywords such as "police", "teargas" in the precursors for violent protests suggest that greater forms of control and authority often beget violence. This reflects ongoing understanding that violence often has at its roots distrust between authorities and citizens. Second, we observe that before a violent protest event there have been other protests and strike events (even if peaceful) among the precursors. Words such as "protest", "block", "march", "strike" tend to appear more frequently in news articles preceding a violent protest than a non-violent protest. These findings suggest that a rapidly rising sequence of peaceful events can produce an emotional state among protesters that increases the likelihood of violence in the next protest. We suggest that emotional state is frustration. Given the myriad of challenges confronting civil society in Latin America today, fortunately frustration is not sufficient to produce violence. But the pattern of an increase in a set of words and events is intriguing. Third, analysis of precursor events indicates that they are occurring across a few cities or states, not just in the locale that will subsequently experience violence. One might have thought that protesters would be most affected by what happens locally but this data suggests that those protesters prone to violence reflect upon their national and not just local experiences when formulating their grievances and developing their frustrations. This correlation needs to be explored carefully because it both limits the responsibility of local authorities for potential violence and suggests that locally focused tactics to lower the risk of violence will be of little value.

Table 8. Precursor events and word distributions discovered by the proposed model in past seven days for violent protest events and one non-violent protest event. Related keywords are manually highlighted (from Google Translate). Note: ungrammatical sentences are due to shortcomings of translation; all analysis in this paper is performed in the native language.

Day	Precursor Events	Top Words
Day T-8	**P1**. A group of about 50 people hold a manifestation in Copacabana Beach.	street
Day T-5	**P1**. In the early afternoon, about 150 teachers protest in the center of the city.	protest teachers
Day T-3	**P1**. Workers of the petrochemical complex in Rio De Janeiro make a protest. **P2**. Several activists representative the commission of human rights participate in a protest	protest activist
Day T-1	**P1**. Police user tear gas against protesters in Tijuca, northern Zon of Rio de Janeiro	police
2014-07-14	**Violent Protest**: About 400 workers protest for delayed salary in Rio de Janeiro. Manifestation ends in clash with police.	
Day	**Precursor Events**	**Top Words**
Day T-8	**P1**. The rain of Sao Paulo didn't prevent women march on street for equal rights	street
Day T-5	**P1**. Protest in front of the central railway station of Brazil. **P2**. About 25 students and their fathers participate a protest	protest students
Day T-3	**P1**. Thousands of protesters gather together in the Cinelndia, downtown Rio.	protest
Day T-2	**P1**.The demonstration against the Dilma government and corruption in Belo Horizonte. **P2**. About 3000 people, according to estimate of the military police, participate in the protest	government police national
2015-03-17	**Violent Protest**:A group of protesters close the runway in the marginal Tiete, burn tires and garbage bags.	
Day	**Precursor Events**	**Top Words**
Day T-8	**P1**. A young guy died from an attempt of kidnapping.	criminal justice
Day T-7	**P1**. Senate start debate on the reform of the political penal code and the criminal justice commission matter **P2** Deputy of the New Alliance Party arrive to the session of the plenary of the local congress,	economic maintain human
Day T-5	**P1**. The workers against Congress gather in May Square.	organization
Day T-2	**P1**. The judge rejected the proposition on one of the accused criminals who took the property of Lugano	financial property
Day T-1	**P1**. Men in police disguise assaulted the house of a doctor. **P2**. After two hours of chaos, workers lead a protest	medical police
2014-05-07	**Non-Violent Protest**: About 500 people gathered last night at the Plaza San Martin to demand the authority to enforce the security of the citizens.	

5.10 How fast does the nMIL^{vio} run?

Figure 11 presents the computation time of the proposed method and state-of-the-art methods on a Dell server with Intel Xeon CPU, 80-core, 504 GB memory based Ubuntu 12.04.5 operating system.

The **MI-SVM** algorithm is computationally more expensive when the number of training examples is large. Other probability based MIL methods are relatively stable when the training set is varied from 10% to 100%.

6 RELATED WORK

Event Detection and Forecasting. Event detection and forecasting from online open source datasets has been an active area of research in the past decade. Both supervised and unsupervised machine learning techniques have been developed to tackle different challenges. Linear regression models use simple features to predict the occurrence time of future events (Arias et al. 2014; Bollen et al. 2011; He et al. 2013; O'Connor et al. 2010). Advanced techniques use a combination of sophisticated features such as topic related keywords, as input to support vector machines, LASSO and multi-task learning approaches (Ritterman et al. 2009; Wang et al. 2012). Ramakrishnan *et al.* (Ramakrishnan et al. 2014) designed a framework (EMBERS) for predicting civil unrest events in different locations by using a wide combination of models with heterogeneous input sources ranging from social media to satellite images. Zhao *et al.* (Zhao et al. 2015b) combine multi-task learning and dynamic features from social networks for spatial-temporal event forecasting. Generative models have also been used in (Zhao et al. 2015a) to jointly model the temporal evolution in semantics and geographical burstiness within social media content. Laxman *et al.* (Laxman et al. 2008) designed a generative model for categorical event prediction in event streams using frequent episodes. However, few existing approaches provide evidence and interpretive analysis as support for event forecasting.

Identifying Precursors. Identifying precursors for significant events is an interesting topic and has been used extensively for interpretive narrative generation and in storytelling algorithms (Hossain et al. 2012). A nested Multi-Instance learning framework (Ning et al. 2016) which is also the basis of this paper has been proposed to tackle the event forecasting and precursor mining problem. Rong *et al.* (Rong et al. 2015) developed a combinational mixed Poisson process (CMPP) model to learn social, external and intrinsic influence in social networks.

Multiple Instance Learning. In the multiple instance learning (MIL) paradigm, we are given labels for sets of instances commonly referred as *bags* or *groups*. However, individual instance-level labels are unknown or missing. The bag-level labels are assumed to be an association function (e.g., OR, average) of the unknown instance level labels. One approach to MIL adapts support vector machines (SVMs) by: (i) modifying the maximum margin formulation to discriminate between bags rather than individual instances (Andrews et al. 2002), and (ii) developing kernel functions that operate directly on bags (Gartner et al. 2002). Other multiple instance learning approaches and various applications are found in a detailed survey (Amores 2013). Specifically, the generalized MIL (Weidmann et al. 2003) formulation assumes the presence of multiple concepts and a bag is classified as positive if there exists instances from every concept. Relevant to our work, besides predicting bag labels, Liu *et al.* (Liu et al. 2012) seek to identify the key instances within the positively-labeled bags using nearest neighbor techniques. Recent work (Kotzias et al. 2015) has focused on instance-level predictions from group labels (GICF) and allowed for the application of general aggregation functions with applications to detecting sentiments associated with sentences within reviews.

The methods proposed in this paper can be viewed as complementary to prior work, casting the forecasting and precursor discovery problems within novel extensions of multiple instance learning.

7 CONCLUSION

We have introduced a nested multi-instance learning framework for forecasting protest events and its adaptation for violent crowd behavior, and identifying precursor events that lead to violence. Most event forecasting approaches focus on early detection of events or predictive accuracy using spatial and temporal dependencies in the underlying data. In contrast, we present an application of multi-instance learning that ingests historical online web data to provide probabilistic evidence of a forecast and further distinguishes violent protest events in the near future from other non-violent events.

We empirically evaluated the strengths of our developed method on open source news datasets from five Latin American countries. Through extensive evaluation and analysis, we illustrate the strong forecasting performance of the proposed methods for violence prediction. We also show qualitatively via several case studies, the characteristics of identified precursors for both violent and non-violent protest events. In the future, we plan to study the patterns of change in protest events turning to violence and other societal factors that contribute to the evolution of violent protest events.

ACKNOWLEDGMENTS

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

REFERENCES

H. Achrekar, A. Gandhe, R. Lazarus, Ssu-Hsin Yu, and Benyuan Liu. 2011. Predicting Flu Trends using Twitter data. In *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on.* 702–707.

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

Stuart Andrews, Ioannis Tsochantaridis, and Thomas Hofmann. 2002. Support vector machines for multiple-instance learning. In *Advances in neural information processing systems*. 561–568.

Marta Arias, Argimiro Arratia, and Ramon Xuriguera. 2014. Forecasting with Twitter Data. *ACM Trans. Intell. Syst. Technol.* 5, 1, Article 8 (Jan. 2014), 24 pages.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.* 3 (March 2003), 1137–1155.

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

Rich Caruana. 1997. Multitask Learning. *Mach. Learn.* 28, 1 (July 1997), 41–75.

Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Mach. Learn.* 20, 3 (Sept. 1995), 273–297.

Thomas Gartner, Peter A. Flach, Adam Kowalczyk, and Alex J. Smola. 2002. Multi-Instance Kernels. In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*. San Francisco, CA, USA, 179–186.

T. Hassner, Y. Itcher, and O. Kliper-Gross. 2012. Violent flows: Real-time detection of violent crowd behavior. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops.* 1–6.

Jingrui He, Wei Shen, Phani Divakaruni, Laura Wynter, and Rick Lawrence. 2013. Improving Traffic Prediction with Tweet Semantics. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI '13).* 1387–1393.

M. Shahriar Hossain, Patrick Butler, Arnold P. Boedihardjo, and Naren Ramakrishnan. 2012. Storytelling in Entity Networks to Support Intelligence Analysts. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '12).* New York, NY, USA, 1375–1383.

Dimitrios Kotzias, Misha Denil, Nando de Freitas, and Padhraic Smyth. 2015. From Group to Individual Labels Using Deep Features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15).* New York, NY, USA, 597–606.

Srivatsan Laxman, Vikram Tankasali, and Ryen W. White. 2008. Stream Prediction Using a Generative Model Based on Frequent Episodes in Event Sequences. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '08).* New York, NY, USA, 453–461.

Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. *CoRR* abs/1405.4053 (2014).

Guoqing Liu, Jianxin Wu, and Zhi-Hua Zhou. 2012. Key Instance Detection in Multi-Instance Learning.. In *ACML (JMLR Proceedings)*, Steven C. H. Hoi and Wray L. Buntine (Eds.), Vol. 25. 253–268.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient Estimation of Word Representations in Vector Space. *CoRR* abs/1301.3781 (2013).

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013b. Distributed Representations of Words and Phrases and their Compositionality. *CoRR* abs/1310.4546 (2013).

Sathappan Muthiah, Patrick Butler, and others. 2016. EMBERS at 4 Years: Experiences Operating an Open Source Indicators Forecasting System. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. New York, NY, USA, 205–214.

Yue Ning, Sathappan Muthiah, Huzefa Rangwala, and Naren Ramakrishnan. 2016. Modeling Precursors for Event Forecasting via Nested Multi-Instance Learning. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. New York, NY, USA, 1095–1104.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series.. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Naren Ramakrishnan, Patrick Butler, Sathappan Muthiah, and others. 2014. "Beating the News" with EMBERS: Forecasting Civil Unrest Using Open Source Indicators. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*. New York, NY, USA, 1799–1808.

Joshua Ritterman, Miles Osborne, and Ewan Klein. 2009. Using prediction markets and Twitter to predict a swine flu pandemic. In *In Proceedings of the 1st International Workshop on Mining Social*.

Yu Rong, Hong Cheng, and Zhiyu Mo. 2015. Why It Happened: Identifying and Modeling the Reasons of the Happening of Social Events. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. New York, NY, USA, 1015–1024.

A. Tumasjan, T.O. Sprenger, P.G. Sandner, and I.M. Welpe. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. 178–185.

Xiaofeng Wang, Matthew S. Gerber, and Donald E. Brown. 2012. Automatic Crime Prediction Using Events Extracted from Twitter Posts. In *Proceedings of the 5th International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction (SBP'12)*. Berlin, Heidelberg, 231–238.

Xinggang Wang, Zhuotun Zhu, Cong Yao, and Xiang Bai. 2015. Relaxed Multiple-Instance SVM with Application to Object Discovery. *CoRR* abs/1510.01027 (2015).

Nils Weidmann, Eibe Frank, and Bernhard Pfahringer. 2003. A Two-Level Learning Method for Generalized Multi-instance Problems. In *The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. 468–479.

Liang Zhao, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2015a. Spatiotemporal Event Forecasting in Social Media. In *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015*. 963–971.

Liang Zhao, Qian Sun, Jieping Ye, Feng Chen, Chang-Tien Lu, and Naren Ramakrishnan. 2015b. Multi-Task Learning for Spatio-Temporal Event Forecasting. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. New York, NY, USA, 1503–1512.

Zhi-Hua Zhou and Jun-Ming Xu. 2007. On the relation between multi-instance learning and semi-supervised learning.. In *ICML (ACM International Conference Proceeding Series)*, Zoubin Ghahramani (Ed.), Vol. 227. 1167–1174.