Forecasting violent events in the Middle East and North Africa using the Hidden Markov Model and regularized autoregressive models

KSM Tozammel Hossain, Shuyang Gao, Brendan Kennedy, Aram Galstyan and Prem Natarajan

Abstract
This paper focuses on forecasting Military Action-type events by both state and non-state actors. Here we demonstrate that the dynamics of these types of events can be adequately described by a Hidden Markov Model (HMM) where the hidden states correspond to different operational regimes of an actor, and observations correspond to event frequency—and the HMM effectively predicts events with different lead times. We also demonstrate that one can enrich statistical time series-based methods that work only on historical data by exploiting predictive signals in real-time external data streams. We demonstrate the superior predictive power of the proposed models with evaluation of recent data capturing activities over two groups, ISIS and the Syrian Arab Military, two countries, Syria and Iraq, and two cities, Aleppo and Mosul. We also present an approach to converting predictions of the proposed models to real-world warnings.

Keywords
Event forecasting, Hidden Markov Model, autoregressive models, external signals

1. Introduction
There has been significant recent interest in modeling and predicting violent events, such as Military Action by state actors, or terrorist attacks by non-state actors, collectively referred to as MANSA events. There are certain characteristics of MANSA event dynamics that make them particularly challenging to model. For instance, it is well established that the dynamics of terrorist attacks have distinctly non-Poissonian characteristics. In particular, the inter-event duration distribution (which is exponential for the Poisson process) has been shown to be heavy tailed and bursty for a number of different event types. Thus, we need different mechanisms for adequately reproducing and predicting activity patterns with highly non-Poissonian statistics.

In this paper we study the problem of forecasting violent events using historical and external open source data (e.g., news articles, blogs, and tweets), given that the historical data may not be up-to-date. Specifically, we focus on predicting violent events in the Middle East and North Africa (MENA) region over a year from 1 August 2016 to 30 September 2017. For evaluation we use manually curated violent events with rich features—such as actor, target, time, and location—as well as news articles data collected over the MENA region by Arabia Inform.

It has been previously demonstrated that the bursty dynamics of terrorist activity can be well-captured by an appropriately designed $d$-state Hidden Markov Model
(HMM), where a hidden state characterizes a specific operational mode of an organization. The simplest setting of \( d = 2 \) corresponds to the case where the dynamics are coarsely quantized as low-activity and high-activity regimes, respectively. Here we show that even a simple two-state HMM can be used to adequately describe the daily patterns of violent events by ISIS, the Syrian Arab Military, and other actors, providing better predictive capabilities over simple baseline models.

Another important challenge for developing high-fidelity models for MANSA events is the availability of reliable, up-to-date historical data for generating real-time predictions. Indeed, recent studies, such as Raghavan et al. and Porter et al., try to predict the number of terrorist attacks at time \( t + 1 \), assuming that one has access to the historical data up to time \( t \). This assumption does not hold in many realistic situations. Indeed, in a typical scenario, one usually has access to historical data up to time \( t - \tau \), where \( \tau \) is a scenario-dependent time lag, so one has to make a prediction without relying on the most recent historical data. Figure 1 illustrates the realistic settings for event forecasting.

Here we address this shortcoming of existing models by proposing to use additional (surrogate) data sources to compensate for the lack of most recent event data. In particular, we focus on a scenario where in addition to historical event counts, we also have a time-stamped set of documents that contains potentially relevant information about events. Our results indicate that the signals extracted from streaming news sources can indeed lead to more accurate forecasts.

The rest of the paper is organized as follows: Section 2 discusses relevant research on event forecasting and Section 3 presents models that we exploit for forecasting MANSA events. Finally, we present an evaluation of our models in Section 4 and discuss our findings in Section 5.

2. Related work

There has been a significant interest in modeling the activities of terrorist groups. Enders and Sandler proposed a threshold autoregressive (TAR) model to study both short- and long-run spurts in terrorist activities. Dugan et al. suggested group-based trajectory analysis techniques (Cox proportional hazards model or zero-inflated Poisson model) to identify regional terrorism trends with similar developmental paths. More recently, Porter et al. suggested the two-component self-exciting hurdle model (SEHM) and Raghavan et al. proposed a \( d \)-state HMM for describing the activity profile of terrorist groups.

Developing a precise model for the dynamic behavior of time series is a challenging problem and an essential one for the success of forecasting methods. Researchers have extensively studied and used time series analysis in many domains, such as finance, epidemiology, geophysics, and sociology. A popular strategy for analyzing time series data is using classical autoregressive models, such as AR, ARMA, ARIMA, and ARMAX. Autoregressive models are widely used in intrusion detection, detecting denial-of-service (DoS) attacks, and network monitoring. These models assume that the underlying data-generating process is linear, that is, the value at a time point is a linear combination of the past values. However, real-world time series exhibit volatility and nonlinearity. A way to deal with the problem of volatility is to employ ARCH and GARCH, which are extensions of classical autoregressive models.

The generation of temporal features from text corpora for event forecasting is a diverse practice in the prediction of civil unrest, crime, political violence, and epidemics. Using datasets of social media or news articles, domain-relevant information is typically extracted.

![Figure 1. Overview of event forecasting without recent historical data. The proposed model takes historical data and indicators from external sources as inputs, and the model makes forecasts without recent historical data. GSR: gold standard report.](image-url)
using expert-generated keywords as a starting point. Techniques that generate features from social media text using some form of supervised learning—keyword counting, manual document filtering, document classification, etc.—include work in spatio-temporal forecasting of civil unrest by Zhao et al.\textsuperscript{29,30} using keywords to filter relevant information from social media posts. In the same domain, Compton et al.\textsuperscript{31} use keywords and geographical terms to filter Twitter posts, performing manual annotation on a small set of tweets in order to produce detailed forecasts of the demographic, spatial, and temporal information of civil unrest events.

Emphasizing the role news articles can play as precursors to particular events, Ning et al.\textsuperscript{32} propose a nested, multi-task learning approach to discover news articles that have a high impact on future event outcomes—whether or not a protest event occurs in a certain city. In this model, documents are represented as bag-of-words or a similarly unsupervised method of representation.

Forecasting military events has gained attention in recent years, as datasets have become more available. Zammit-Mangion et al.\textsuperscript{33} apply a point process model to conflict events from the Afghan War Diary. Yonamine\textsuperscript{34} models military events in Afghanistan using the Autoregressive Fractionally Integrated Moving Average (ARFIMA) to predict time series of district-level event counts. For a comprehensive review of datasets and models for the prediction of political violence, we refer the reader to Schrodt et al.\textsuperscript{1}.

3. Models

The intuition behind time series model is that when events are correlated in time, then given a sequence of events, one can learn patterns of past events that are useful for predicting future events. Time series prediction techniques use historical data about events (with optional surrogate data) to learn a model of the process that produced these events. The model can, in turn, be used to predict new events. In this section, we describe how we apply two types of models—the HMM and autoregressive models—to address the challenge of modeling events executed by military and non-state actors.

3.1. Hidden Markov Models

We first present the HMM-based approach for modeling terrorist activities. In our context, the key idea of the HMM is that the current number of events (e.g., terrorist activities) depends on the past history of events through \( K \) dominant hidden states, which represent different operational phases of the terrorist activities. For example, the hidden states of a two-state HMM correspond to “low-activity” and “high-activity” processes, as shown in Figure 2. The process transitions probabilistically between low-activity and high-activity states. While in a particular state, the process outputs some events according to a state-dependent probability distribution.

Let \( Y = (y_1, y_2, \ldots, y_T) \) be the observed sequence of events, for example, the daily number of terrorist attacks, and \( Z = (z_1, z_2, \ldots, z_T) \) be the underlying states of the process giving rise to the events \( Y \). Here \( T \) denotes the length of the time series, that is, the sequence of events. A HMM is described by a set of hidden states \( \{S_1, S_2, \ldots, S_N\} \), transition probabilities between the states \( \eta_{ij} = P(z_t = S_j | z_{t-1} = S_i) \), initial probabilities of the states \( \pi_i = P(z_1 = S_i) \), and the emission probabilities of events conditioned on the hidden state \( \phi_i(k) = P(y_t = k | z_t = S_i) \). The hidden states \( Z \) are discrete-valued random variables. A transition between the states is Markovian, that is, the future state is conditionally independent of the past states given the current state. In our problem setting, we consider the emission probabilities of events to be a continuous value from one of four possible distributions: Poisson, Gaussian, geometric, or Hurdle.
We propose RARE—regularized autoregression with exogenous variables—for predicting terrorist activities. RARE is based on the ARX model—the autoregressive model with external variables—and Lasso. The key idea is to use penalized regression (e.g., Lasso) for selecting autoregressive terms as well as covariates. The model is robust to the absence of historical data and requires limited history for prediction.

Let $\mathcal{Y} = (y_1, y_2, \ldots, y_T)$ be the observed sequence of events over $T$ time units, that is, the length of the time series. Formally, RARE($p, k$) defines an autoregressive model with $p$ autoregressive lags and $k$ external variables. Given the observed series of events $\mathcal{Y} = (y_1, y_2, \ldots, y_T)$, the predicted value $y_t$ at time point $t$ is expressed as follows:

$$y_t = \mu_y + \sum_{i=1}^{p} \alpha_i y_{t-i} + \sum_{j=1}^{k} \beta_j x_{j,t} + w_t$$  \hspace{1cm} (1)$$

Here $\mu_y$ is a constant, $\alpha_i$ is the autoregressive (AR) coefficient at lag $i$, $\beta_j$ is the regression coefficient for external variable $x_j$, and $w_t \sim \mathcal{N}(0, \sigma^2)$ is the white noise at time point $t$. The model exploits $\ell_1$-regularization for selecting $k$ external variables. We estimate the model parameters $\mu_y, \alpha, \beta$ by minimizing the following objective function:

$$\sum_{t=1}^{T} \left( y_t - \mu_y - \sum_{i=1}^{p} \alpha_i y_{t-i} - \sum_{j=1}^{k} \beta_j x_{j,t} \right)^2 + \lambda_\alpha ||\alpha||_1 + \lambda_\beta ||\beta||_1$$  \hspace{1cm} (2)$$

3.1.1. Estimating HMM Parameters. The unknown parameters of the proposed HMM are $H = \{\pi, \eta, \phi\}$. No analytical solution exists for this model that maximizes the probability of the observed sequence (i.e., likelihood). Hence, we applied an Expectation Maximization (EM)-based algorithm (also known as Baum–Welch reestimation) to estimate the parameters of the model.

3.1.2. Predicting with the HMM. To predict the number of new events, we adopt a sliding window approach. We teach our model with data determined by a user-defined time window (e.g., four months), estimate the expected number of events for a gap period (e.g., one month), and forecast for the next one month. The expected number of events at time $t$ given $z_{t-1}$ is as follows:

$$\bar{y}_t = \sum_{j}^{N} \eta_{z_{t-j}} \cdot E[S_j]$$

where $E[S_j]$ is the expected number of events at state $S_j$.

3.2. Autoregressive models

We propose RARE—regularized autoregression with exogenous variables—for predicting terrorist activities. RARE is based on the ARX model—the autoregressive model with external variables—and Lasso. The key idea is to use penalized regression (e.g., Lasso) for selecting autoregressive terms as well as covariates. The model is robust to the absence of historical data and requires limited history for prediction.

Let $\mathcal{Y} = (y_1, y_2, \ldots, y_T)$ be the observed sequence of events over $T$ time units, that is, the length of the time series. Formally, RARE($p, k$) defines an autoregressive model with $p$ autoregressive lags and $k$ external variables.

**Algorithm 1.** Generator($\eta, \pi$) for HMM

**Input:** A set of parameters.

**Output:** Number of domain registrations.

1. Choose the initial state $z_1 \sim \text{Mult}(\pi)$
2. Draw each row of $\eta_1$ using Dir($\alpha$) > Transition matrix for a user-defined $\alpha$
3. Choose the emission probability distribution $\phi \in \{\text{Poisson, Gaussian, Geometric, Hurdle Geometric}\}$
4. For each time $1 \leq t \leq T$ do
5. If not the 1st day then
6. $z_t \sim \text{Mult}(\eta_{z_{t-1}})$
7. Draw $y_t \sim \phi_{z_t}$

geometric. The generative process for the model is shown in Algorithm 1.
\[ y_t = \frac{1}{W} \sum_{i=1}^{W} y_{t-i} \] (4)

### 3.3. Evaluation of time series models

We use three error measures for quantitative evaluation of our time series models: (a) mean absolute error (MAE); (b) root mean squared error (RMSE); and (c) mean absolute scaled error (MASE). These measures are defined as follows in terms of forecasting error, \( e_t = y_t - \hat{y}_t \), at time point \( t \), where \( y_t \) and \( \hat{y}_t \) are the true and predicted values, respectively.

- **MAE**:
  \[ \text{MAE} = \frac{1}{T} \sum_{t=1}^{T} |e_t| \]

- **RMSE**:
  \[ \text{RMSE} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} |e_t|^2} \]

- **MASE**:
  \[ \text{MASE} = \frac{\frac{1}{T} \sum_{t=1}^{T} |e_t|}{\frac{1}{T-1} \sum_{t=2}^{T} |y_t - y_{t-1}|} \]

### 4. Experiments

We now present a case study for the proposed models using data on military and non-state actor events in the MENA region. Our goals are to answer the following questions.

1. Can the HMM capture latent structures in activities executed by various actors?
2. How do the proposed models perform with MANSA events at actor, country, and city levels?
3. Which external signals are good indicators for forecasting MANSA events?
4. How can we generate warnings given predicted event counts? How does the model perform in terms of quantitative evaluation of generated warnings?

#### 4.1. Datasets

The ground truth information about MANSA events, called the gold standard report (GSR) is exclusively provided by the Center for Analytics at New Haven. The GSR is a manually created list of MANSA events by domain experts. Each event in the dataset has 22 different attributes: actor, actor status, approximate location, causalities, country, earliest reported date, encoding comment, event date, event id, event subtype, event type, first reported link, gold standard source link, latitude, longitude, news source, other links, revision date, state, target, target name, and target status. While much care had been taken to address the attribution and duplication problem in the manual event documentation step, we also remove any duplicates in preprocessing steps using the these attributes.

For evaluation we use ground truth time series of daily event counts based on manually extracted, structured reports on events, at actor, city, and country level (see Table 1). We use two actors—ISIS and the Syrian Arab Army, two countries—Syria and Iraq, and two cities—Aleppo and Mosul. In addition, we use surrogate data, which is generated from Arabic news articles originating from MENA countries.

In order to generate potentially predictive signals, we apply a temporal topic-based feature extraction approach to Arabia Inform news articles, a corpus of news documents originating from MENA countries (see Figure 3), over a time span co-occurring with our GSR event time series. We consider the subset of the corpus that has at least one of our countries of interest (Iraq, Syria, Saudi Arabia, Lebanon, Yemen, Jordan) “tagged” as part of the meta-data provided from each document’s URL. As the corpus consists mostly of articles published in Egypt, and thus the majority of articles have “Egypt” as a tagged location, we exclude articles about places in Egypt. This largely Arabic corpus has approximately 20,000 documents per day, including a variety of topics spanning entertainment, politics, reporting articles, and general purpose news items.

#### 4.1.1. Topic-based temporal feature generation

To learn latent shifts in the news corpus that possess information
about our events of interest, we chose two topic modeling
techniques: firstly, we train Latent Dirichlet Allocation
(LDA) models with 100, 150, and 200 topics on the whole
corpus and aggregate (see below for details) the posterior
distributions of each topic over a given day’s documents.
Secondly, we pre-train a LDA model on a set of 10,000
Arabic news articles—reporting MANSA events—which
were used to generate the ground truth event dataset, and
repeat the temporal feature generation on the entire Arabia
Inform corpus, inferring topic posterior distributions
over the same corpus. We use the Mallet LDA package, performing light stemming and stop-word removal as pre-
processing. We found the Mallet LDA package to produce
more coherent and consistent topics compared to the
Corex and Gensim LDA packages. As the news arti-
cles in our dataset are predominantly in Arabic (90%), we
perform light stemming, as Arabic is a highly inflecting
language and the development of a proper Arabic lem-
matizer is still an active area of research. For our experi-
ments, we did not achieve significant results with the first
method (not pre-trained), and thus present only our find-
ings with the pre-trained topic model.

In order to generate daily features given a trained topic
model and a set of time-stamped documents, we denote $D_t$ as the set of Arabia Inform documents on day $t$, $N_t$ as the
number of documents on day $t$, and $d_{t,i}$ as the $i$th docu-
ment on day $t$, represented by a bag-of-words. Then we
have $v_{t,i}$ as the $k$-dimensional vector of estimated topic
distributions for a learned LDA topic representation with $k$
topics, applied to document $d_{t,i}$. $v_{t,i,j}$ represents the $j$th coordinate of the feature vector $v_{t,i}$ of document $d_{t,i}$. Finally, to generate a temporal feature vector using a doc-
ument corpus, we denote the $j$th coordinate of the feature
vector $V_t$, for $j \in k$:

$$V_{t,j} = \frac{1}{N_t} \sum_{i=1}^{N_t} v_{t,i,j}$$

4.2. Structures in actors’ activities using a two-state
Hidden Markov Model

We hypothesize that the number of activities performed by
an actor might have hidden structures (e.g., high-activity
and low-activity periods), which may not be well-captured
using a simple counting process, such as the Poisson pro-
cess. We employ the HMM for capturing hidden structures
in activities by various actors in our dataset, such as ISIS,
the Syrian Arab Military, the Iraqi Military, and the
Russian Military. Figure 4 illustrates the result of the
HMM for ISIS with various settings. Specifically, the data-
sets contain daily counts of terrorist events by ISIS in Iraq
and Syria. We used the initial 80% of the data for training,
using the Baum–Welch algorithm to estimate the model
parameters. Here we report results using the Gaussian
observation model, so that the total number of parameters
is four: two transition probabilities, from L state to H and
vice versa, and four parameters for the observation model
(two for each hidden state).

Figure 4 depicts the model learned via the Baum–
Welch method. We observe that both hidden states have
significant inertia, for example, the actor is more likely to
stay in the same hidden state than transition to a new one.
Also, what is perhaps more important, is that the rate of
events (as characterized by the mean of the Gaussian
model) differs significantly between the states: The aver-
age number of attacks per day is 13.4 when the actor is in
high-activity state $H$, compared to 5.7 in the low-activity
state $L$. Figure 4(b) shows the prediction error of the model
for different choices of observation models with different
lead times.

Next, we focus on the task of reconstructing the hidden
trajectory of the actor. Toward that goal, we run the
Viterbi algorithm, which returns a single (maximum a pos-
teriori) hidden state sequence that best explains the
observed counts. Figure 5 shows the event count together
with the reconstructed hidden dynamics. Remarkably,
even this simple two-state model is able to capture the spurts in the activity.

4.3. Predictive performance of HMM, ARIMA, and RARE models

The GSR represents the occurrence of an event on a given day at a specific location by a specific actor. As GSR is typically lagged (e.g., by a month), and thus it poses a challenge for the prediction algorithm. In our evaluation settings, we assume a gap of a month between the last day of the training period and the first day of the testing period. We keep the test period to be a month, as the GSR is updated each month. For the RARE model, we use topic-based temporal features as external signals (see Section 4.1). Before applying temporal features, we first align them using correlation analysis with the GSR: we determine the lag where the maximum correlation occurred between a temporal feature and the GSR, and use the lag for alignment. We tested the RARE model with 50 and 100 external features and $p = 30$.

Figure 6(a) and Table 2 illustrate the models’ predictions and performance measures for ISIS activities over the month of January 2017, respectively. Here the models are trained with the data from 1 August 2016 to 30 November 2016, and the month of December 2016 is considered as the gap period; (b) forecasting of Syrian Arab Military activities over the period (March 2017) wherein the models are trained with the data from 1 August 2016 to 31 January 2017, and the month of February 2017 is considered as the gap period. For both settings, the HMM with two hidden states and Gaussian emission probability is used, and the ARIMA and RARE models are identified using a grid search over parameters.
predictions and performance measures for Syrian Arab Military activities over the month of March 2017, respectively. Here we assume the absence of a GSR in February 2017, and the models are trained with the data from 1 August 2016 to 31 January 2017. The RARE model clearly outperforms the other models in terms of capturing the trends and performance measures.

We also evaluate our models for country-level event activities. We use six months of the GSR as training data starting from 1 August 2016, and use the model for predicting over a month, where there is a gap of a month between training and forecasting spans. We then shift the training period by a month and repeat the forecasting up to the month of September 2017. Tables 4 and 5 show the comparison between methods with different average performance metrics with average over seven months: mean absolute error (MAE), root mean squared error (RMSE), and mean absolute scaled error (MASE).
Figure 7. Forecasting events in Syria and Iraq using the Hidden Markov Model (HMM), autoregressive integrated moving average (ARIMA) model, regularized autoregression with exogenous variables (RARE) model, and a base rate model. (a), (b) Models forecast activities in Syria and Iraq over the period August 2017 and May 2017, respectively. For both settings, the HMM with two hidden states and Gaussian emission probability is used, and the ARIMA and RARE models are identified using a grid search over parameters.

Figure 8. Forecasting events in Aleppo and Mosul using the Hidden Markov Model (HMM), autoregressive integrated moving average (ARIMA) model, regularized autoregression with exogenous variables (RARE) model, and a base rate model. (a), (b) Models forecast activities in Syria and Iraq over the period May 2017. For both settings, the HMM with two hidden states and Gaussian emission probability is used, and the ARIMA and RARE models are identified using a grid search over parameters.

Table 6. Forecasting of MANSA events in Aleppo using the Hidden Markov Model (HMM) and autoregressive models (autoregressive integrated moving average (ARIMA) and regularized autoregression with exogenous variables (RARE)). Methods are compared in terms of different performance metrics with average over seven months: mean absolute error (MAE), root mean squared error (RMSE), and mean absolute scaled error (MASE).

| Method          | MAE  | RMSE | MASE |
|-----------------|------|------|------|
| Base rate       | 3.69 | 3.88 | 3.01 |
| HMM_Gaussian    | 1.56 | 1.78 | 1.32 |
| ARIMA           | 3.40 | 3.74 | 2.22 |
| RARE            | 1.57 | 1.83 | 1.24 |

Table 7. Forecasting of MANSA events in Mosul using the Hidden Markov Model (HMM) and autoregressive models (autoregressive integrated moving average (ARIMA) and regularized autoregression with exogenous variables (RARE)). Methods are compared in terms of different performance metrics with average over seven months: mean absolute error (MAE), root mean squared error (RMSE), and mean absolute scaled error (MASE).

| Method          | MAE  | RMSE | MASE |
|-----------------|------|------|------|
| Base rate       | 4.37 | 4.89 | 2.07 |
| HMM_Gaussian    | 3.85 | 4.52 | 1.77 |
| ARIMA           | 3.64 | 4.17 | 1.59 |
| RARE            | 4.25 | 4.77 | 2.03 |
Table 8. Representative features selected by the regularized autoregression with exogenous variables model with a training set for ISIS activities from 1 August 2016 to 30 November 2016.

| Topic 0   | Topic 9   | Topic 38  | Topic 45  | Topic 48   |
|----------|-----------|-----------|-----------|-----------|
| Damascus Forces | Lattakia Forces | General Security | Forces Government Capture | Forces Big Brotherhood |
| Eastern Aleppo | Syrian conflict Sham | Syrian Region Prosecution | East North Necessity Accused | Islamic Support Security |
| Al Gouta Syrian conflict | Islamic Support | Shadow Areas Police | Rights Al-Qaeda Areas Police | Rights Eastern Aleppo |
| Syrian conflict East | Islamic Support | Shadow Areas Police | Rights Eastern Aleppo | Rights Eastern Aleppo |
| Rights East | Islamic Support | Shadow Areas Police | Rights Eastern Aleppo | Rights Eastern Aleppo |
| West Rights | Al-Qaeda Areas Police | Director Work | Loyal group | Loyal group |
| Human Clashes | Front Work | Major General | Insurgents Syrian conflict | Insurgents Syrian conflict |
| Insurgents Loyal | Nationalities Entities | Elements | Insurgents Syrian conflict | Insurgents Syrian conflict |
| Bombing Human | Syrian conflict Including | Investigation Group | Insurgents Syrian conflict | Insurgents Syrian conflict |
| Factions Insurgents | Syrian conflict Including | Investigation Group | Islamic Ocean | Islamic Ocean |
| Islamic Ocean | The Kurds Operation | Group | Islamic Ocean | Islamic Ocean |

Figure 9. Evaluation of warnings generated using the base rate model for two types of events in Syria and Iraq from 1 March 2017 to 30 September 2017. (Color online only.)
country-level event data, we use six months of the GSR as training data starting from 1 August 2016, and use the model for predicting over a month, where there is a gap of a month between training and forecasting spans. We then shift the training period by a month and repeat the forecasting up to the month of September 2017. Tables 6 and 7 show the comparison between methods over different metrics across seven months. We observed that the HMM and RARE model perform better compared to other models for Aleppo, but the ARIMA model outperformed others for Mosul. The reason could be the sparsity in the city-level events. Figures 8(a) and (b) illustrate the models’ predictions over activities in Aleppo and Mosul for the months of August 2017 and May 2017, respectively.

Table 9. Representative features selected by the regularized autoregression with exogenous variables model with a training set for Syrian Arab Military activities from 1 August 2016 to 31 January 2017.

| Topic 0   | Topic 9   | Topic 41 | Topic 47        |
|-----------|-----------|-----------|----------------|
| Damascus  | Forces    | Saudi Arabia | National | Iraq |
| Forces    | Brive     | Border     | Iraq       |
| Eastern   | Aleppo   | Sanafir    | North     |
| Al Gouta  | Syrian conflict | Alliance | Kirkuk |
| Syrian conflict | East | United Nations | Baquba |
| Rights    | Islamic   | Demarcation | Security |
| West      | Rights    | The kingdom | Capture |
| Human     | Clashes   | The Houthis | Diyala |
| Insurgents| Loyal     | Party      | Reporter |
| Bombing   | Human     | agreement  | Mosul     |
| Factions  | Insurgents| Countries  | East      |
| Islamic   | Ocean     |            | Security  |

Figure 10. Evaluation of warnings generated using the Hidden Markov Model for two types of events in Syria and Iraq from 1 March 2017 to 30 September 2017. (Color online only.)
4.4. Important predictors in forecasting MANSA events

The RARE model identifies a subset of autoregressive variables and external variables that are predictive of the target, which is the number of events occurring each day. We analyze the topics selected by the algorithm. As an example, Tables 8 and 9 show some of the features identified by the RARE model with ISIS and Syrian Arab Military activities, respectively. We observe that many of identified topics are meaningful and relevant to the events associated with ISIS and the Syrian Arab Military.

4.5. Warning generation

The proposed models essentially forecast event counts, but an intelligence analyst may need more details about the events for better understanding and dissemination. We propose a two-phase algorithm for generating real-world warnings. We transfer these event counts for each model to meaningful warnings with sampling each event detail field from its corresponding empirical distribution of the fields. To see the efficacy of this approach, we generate warnings at the country level (Syria and Iraq) for two different types of events (military action and non-state actor events) over the months from March to September 2017. For each event count, we use six trials for generating six different sets of warnings. We match the generated warnings against GSR events using the Hungarian matching algorithm as well as other numerical and string matching algorithms. If a warning occurs within seven days of the corresponding true event, then a warning is included for further analysis in terms of various metrics. Figures 9–11 illustrate the evaluation of warnings generated by the base rate model, the HMM, and the RARE model in terms of precision, recall, and quality score. Each box in the plots

Figure 11. Evaluation of warnings generated using the regularized autoregression with exogenous variables model for two types of events in Syria and Iraq from 1 March 2017 to 30 September 2017. (Color online only.)
represents 50% of the data, and each vertical red line denotes the median. We can see that the RARE model performs better than the others in terms of precision, and performs slightly better than the base rate model in terms of warning quality.

5. Discussion

We explore state-based (HMM) and autoregressive (ARIMA and RARE) models for generating event forecasts with external indicators. We observe that both the HMM and RARE model perform quite well with a reasonable amount of data (actor and country-level events), while performance deteriorates when events are sparse. When event density is low and event type is rare, it poses a challenge to our proposed models for predicting events in such settings. Some of the countries (e.g., Saudi Arabia and Yemen) and most of the cities in our dataset have low event density, for which the HMM and the autoregressive models seem inadequate. In addition, there are some event types that are rare, such as some epidemic disease that do not occur so often compared to flu epidemics. For these rare events, the HMM and the autoregressive models may not work well. To address these problems we need predictive models that would take the elaborate event context in external sources into account.

In this study we model each actor independent of others, although actors interact with each other in a real-world scenario. It would be interesting to pursue modeling actors with more than two operational states as well as interactions between multiple actors.

This study explores an external source (Arabia Inform news articles) for event forecasting. Our methods can be extended to deal with signals from additional sources—such as Twitter and blogs—which we plan to explore in the future. It also possible to develop models that consider each of the data sources separately and that select subsets of external signals from each group for prediction. In addition to event count prediction models, we plan to explore models that not only forecast events but also identify the precursors to events in external sources.

Acknowledgments

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 the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon.

Funding

This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA).

ORCID iD

KSM Tozammel Hossain https://orcid.org/0000-0003-0136-2145

References

1. Schrodt PA, Yonamine J and Bagozzi BE. Data-based computational approaches to forecasting political violence. In: Subrahmanian VS (ed) Handbook of computational approaches to counterterrorism. New York: Springer, 2013, pp. 29–162.
2. Raghavan V, Galstyan A and Tartakovsky AG. Hidden Markov models for the activity profile of terrorist groups. Ann Appl Stat 2013; 7: 2402–2430.
3. Arabia Inform.http://arabiainform.com (accessed 30 December 2017).
4. Porter M, White G, et al. Self-exciting hurdle models for terrorist activity. Ann Appl Stat 2012; 6: 106–124.
5. Ward MD, Metternich NW, Carrington C, et al. Geographical models of crises: evidence from ICEWS. Adv Des Cross Cult Activ 2012; pp. 429–438.
6. Lewis E, Mohler G, Brantingham P, et al. Self-exciting point process models of civilian deaths in Iraq. Secur J 2012; 25: 244–264.
7. Enders W and Sandler T. The effectiveness of antiterrorism policies: a vector-autoregression-intervention analysis. Am Polit Sci Rev 1993; 87: 829–844.
8. Enders W and Sandler T. Is transnational terrorism becoming more threatening? A time-series investigation. J Conflict Resolut 2000; 44: 307–332.
9. Dugan L, LaFree G and Piquero A. Testing a rational choice model of airline hijackings. Criminology 2005; 43: 1031–1065.
10. LaFree G, Morris N and Dugan L. Cross-national patterns of terrorism: comparing trajectories for total, attributed and fatal attacks, 1970–2006. Br J Criminol 2009; 50: 622–649.
11. Lendasse A, De Bodt E, Wertz V, et al. Non-linear financial time series forecasting-application to the bel 20 stock market index. Eur J Econ Soc Syst 2000; 14: 81–91.
12. Chakraborty P, Khadivi P, Lewis B, et al. Forecasting a moving target: ensemble models for ILI case count predictions. In: Zaki MJ, Obradovic Z, Tan PN, Banerjee A, Kamath C and Parthasarathy S (eds.) Proceedings of the 2014 SIAM international conference on data mining, Philadelphia, PA, 24–26 April 2014, pp.262–270. SIAM.
13. Wang Z, Chakraborty P, Mekaru SR, et al. Dynamic Poisson autoregression for influenza-like-illness case count prediction. In: Cao L, Zhang C, Joachims T, Webb GI, Margineantu DD and Williams G (eds.) Proceedings of the 21th ACM SIGKDD international conference on knowledge
27. Ulfelder J. A multimodel ensemble for forecasting onsets of
28. Chen L, Hossain KT, Butler P, et al. Syndromic surveillance
26. Minhas S, Ulfelder J and Ward MD. Mining texts to effi-
16. Box-G, Jenkins G, Reinsel G, et al. Time series analysis:
19. Douc R, Moulines E and Stoffer D. Nonlinear time series:
20. Ramakrishnan N, Butler P, Muthiah S, et al. ‘Beating the
17. Prado R and West M. Time series: modeling, computation,
25. Wang M and Gerber MS. Using Twitter for next-place predic-
24. Python A, Illian J, Jones-Todd C, et al. A Bayesian approach
efficiently generate global data on political regime types. In:
18. Viinikka J, Debar H, Mé L, et al. Processing intrusion detec-
23. Wang X, Gerber MS and Brown DE. Automatic crime pre-
22. Kallus N. Predicting crowd behavior with big public data. In:
21. Muthiah S, Butler P, Khandpur RP, et al. Embers at 4 years: Experiences operating an open source indicators forecasting
20. Ramakrishnan N, Butler P, Muthiah S, et al. ‘Beating the news’ with EMBERs: forecasting civil unrest using open source indicators. In: Macskassy SA, Perlich C, Leskovec J, Wang W and Ghani R (eds.) Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining, New York, NY, 24–27 August 2014, pp.1799–1808.
19. Douc R, Moulies E and Stoffer D. Nonlinear time series: theory, methods and applications with R examples. Boca Raton, FL: CRC Press, 2014.
18. Viinikka J, Debar H, Mé L, et al. Processing intrusion detection alert aggregates with time series modeling. Informat Fusion 2009; 10: 312–324.
17. Prado R and West M. Time series: modeling, computation, and inference. Boca Raton, FL: CRC Press, 2010.
16. Box G, Jenkins G, Reinsel G, et al. Time series analysis: forecasting and control. 4th ed. Hoboken, NJ: John Wiley Sons, 2016.
14. Shumway R and Stoffer D. Time series analysis and its applications: with R examples. 3rd ed. New York: Springer Science Business Media, 2011.
15. Box-Steffensmeier J, Freeman J, Hitt M, et al. Time series analysis for the social sciences. New York: Cambridge University Press, 2014.
13. Wang X, Gerber MS and Brown DE. Automatic crime prediction using events extracted from twitter posts. In: Yang SJ, Greenberg AM and Endsley MR (eds.) Proceedings of the International conference on social computing, behavioral-cultural modeling, and prediction, College Park, MD, 3–5 April 2012, pp.231–238. Springer.
12. Ning Y, Muthiah S, Rangwala H, et al. Modeling precursors for event forecasting via nested multi-instance learning. In: Krishnapuram R, Shah M, Smola AJ, Aggarwal CC, Shen D and Rastogi R (eds.) proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, CA, 13–17 August 2016, pp.1095–1104. ACM.
11. Zammit-Mangion A, Dewar M, Kadirkamanathan V, et al. Point process modelling of the Afghan war diary. Proc Natl Acad Sci 2012; 109: 12414–12419.
10. Yonamine JE. Predicting future levels of violence in Afghanistan districts using gdelt. Unpublished manuscript, 2013.
9. Zhaor L, Chen F, Lu CT, et al. Spatiotemporal event forecasting in social media. In: Venkatasubramanian S and Ye J (eds.) proceedings of the 2015 SIAM international conference on data mining, Vancouver, BC, 30 April–2 May 2015. pp.963–971. SIAM.
8. Zhao L, Sun Q, Ye J, et al. Multi-task learning for spatio-temporal event forecasting. In: Zaki MJ, Obradovic Z, Tan PN, Banerjee A, Kamath C and Parthasarathy S (eds.) Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, Sydney, NSW, 10–13 August 2015. pp.1503–1512. ACM.
7. Compton R, Lee C, Xu J, et al. Using publicly visible social media to build detailed forecasts of civil unrest. Secur Informatic 2014; 3 (1):4.
6. Seabold S and Perktold J. Statsmodels: econometric and statistical modeling with large corpora. In: McCullum PC, McCallum A, Dyer J (eds.) Proceedings of the 22nd ACM SIGKDD conference on knowledge discovery and data mining, San Francisco, CA, 13–17 August 2016, pp.963–971. SIAM.
5. Tiibhirini R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol 1996; 58: 267–288.
4. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol 1996; 58: 267–288.
3. Seabold S and Perktold J. Statsmodels: econometric and statistical modeling with python. In: Walt S and Millman J (eds.) proceedings of the 9th Python in science conference, volume 57, Austin, TX, 28 June–3 July 2010, p.61.
2. Rabiner L. A tutorial on hidden Markov models and selected applications in speech recognition. Proc IEEE 1989; 77: 257–286.
1. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B Methodol 1996; 58: 267–288.
Author biographies

KSM Tozammel Hossain is a Postdoctoral Research Associate in the Machine Intelligence and Data Science (MINDS) group at the Information Sciences Institute, University of Southern California (USC). He earned his doctoral degree in computer science at Virginia Tech. His research interests broadly lie in developing and applying data science and machine learning techniques to solve problems stemming from bioinformatics, social networks, and social science.

Shuyang Gao has recently received his PhD in computer science in the MINDS group at the Information Sciences Institute, USC. He has a BS in computer science from Fudan University, Shanghai, China. His research focuses on information theory and machine learning.

Brendan Kennedy is a PhD student in the Department of Computer Science at the USC. He received his BS in computer science at Gonzaga University. His research interests are in the areas of computational social science, interpretable machine learning, and natural language processing.

Aram Galstyan is the Director of the MINDS group at the Information Sciences Institute, USC. He is also a research associate professor at the USC Computer Science department. His work focuses on various problems at the intersection of machine learning, information theory, and statistical physics. He is the Co-PI for IARPA’s Mercury program, where he leads the algorithm development effort for machine-based forecasting. He was the PI for the DARPA SMISC project, where he developed methods that helped the USC ISI team achieve perfect accuracy on DARPA’s Social Bot Detection challenge.

Prem Natarajan is the Michael Keston Executive Director of Information Sciences Institute at the USC, a vice dean of the USC Viterbi School of Engineering, and a professor of computer science. At ISI, he leads institute-wide managerial and technical directions, including research, development, and the MOSIS electronic chip brokerage. He also heads teams in his areas of expertise: novel approaches to face, speech, handwriting, and optical character recognition (OCR), along with other deep learning and natural language processing directions.
