Hindawi
Discrete Dynamics in Nature and Society
Volume 2017, Article ID 8180272, 13 pages
https://doi.org/10.1155/2017/8180272

Research Article

Predicting Social Unrest Events with Hidden Markov Models Using GDELT

Fengcai Qiao, Pei Li, Xin Zhang, Zhaoyun Ding, Jiajun Cheng, and Hui Wang

College of Information Systems and Management, National University of Defense Technology, Changsha, Hunan 410073, China

Correspondence should be addressed to Fengcai Qiao; qiaofengcai125@gmail.com

Received 16 October 2016; Accepted 3 April 2017; Published 10 May 2017

Academic Editor: Pasquale Candito

Proactive handling of social unrest events which are common happenings in both democracies and authoritarian regimes requires that the risk of upcoming social unrest event is continuously assessed. Most existing approaches comparatively pay little attention to considering the event development stages. In this paper, we use autocoded events dataset GDELT (Global Data on Events, Location, and Tone) to build a Hidden Markov Models (HMMs) based framework to predict indicators associated with country instability. The framework utilizes the temporal burst patterns in GDELT event streams to uncover the underlying event development mechanics and formulates the social unrest event prediction as a sequence classification problem based on Bayes decision. Extensive experiments with data from five countries in Southeast Asia demonstrate the effectiveness of this framework, which outperforms the logistic regression method by 7% to 27% and the baseline method 34% to 62% for various countries.

1. Introduction

Social unrest events (protests, strikes, demonstration, and occupation) are common happenings in both democracies and authoritarian regimes [1]. Most social unrest events initially intended to be a demonstration to the public or the government. However, in many occasions they often escalate into general chaos, resulting in violent, riots, sabotage, and other forms of crime and social disorder. Take Thailand as an example; a series of political protests and three military coups happened between 1990 and 2015, resulting in the government being deposed, which illustrates the power of the social unrest. Figure 1 depicts the activities that causally preceded the protest against the amnesty bill in Bangkok at August 7, 2013. Anticipating these latent instabilities before they occur and applying preventive strategies to avoid them have important ramifications such as prioritizing citizen grievances for the decision makers, issuance of travel warnings for the tourism industry, and insight into how citizens express themselves for the social scientist, which has motivated many social and data science researchers to focus on revealing the patterns contained in these events and further the prediction of future latent social unrest.

Last century, most researchers conducted the prediction work using human-coded data, including WEIS [2] and COPDAB [3]. In the last two decades, several small-scale vertical machine-readable datasets [4, 5] and large scale code event datasets like ICEWS [6] and GDELT (Global Data on Events, Location, and Tone) [7] appeared, fueling the development of computation methods for the analysis and prediction of social unrest. It is worth mentioning that the GDELT dataset, with its tremendous amount of event records more than any other event datasets, opens up a new perspective of this research area. So far, there are few works aiming at utilizing GDELT to make predictions about social unrest. Existing works attempted to use linear regression [8], time series forecasting [9], and frequent subgraphs [10, 11] to conduct the prediction work using GDELT. In [12], GDELT and ICEWS are used as data sources to predict unrest in Latin America. Nevertheless, in these works comparatively little attention has been paid to consider the event development stages in the forecasting models with GDELT.

This paper develops a hidden Markov models based framework for leveraging large scale digital history events captured from GDELT to characterize the transitional process of social unrest event evolutionary development. In the

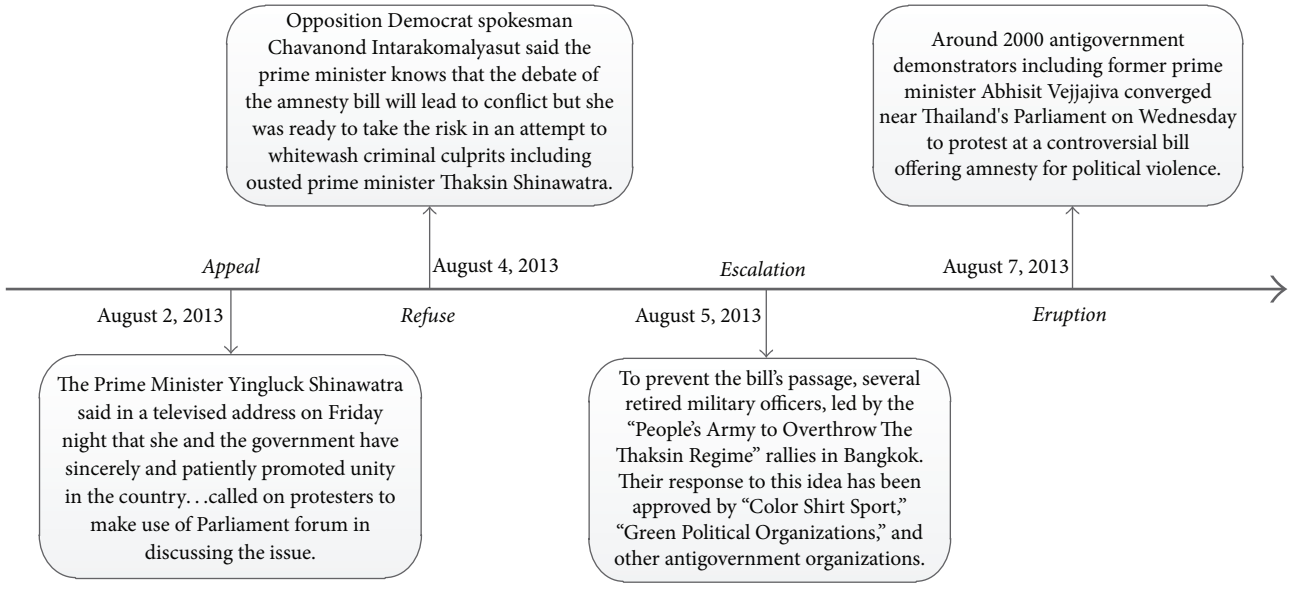


Figure 1: Event development stages before the protest against the amnesty bill at August 7, 2013, in Bangkok, Thailand.

HMM approach discussed by Rabiner [13], the sequencing of observed events can be considered that yield a likely path of hidden states or phases in which the events occur, which is consistent with the concept of event development stage. Our proposed framework utilizes the temporal burst patterns in GDELT event streams to uncover the underlying event development mechanics starting from the prior probability of each stage. Eventually, the social unrest event prediction is formulated as a sequence classification problem. More concretely, our main contributions in this paper to social unrest event prediction with GDELT dataset are four pronged:

(i) First, we identify a sequence or stages of events that potentially lead to a social unrest (like Figure 1). Typical evolution stages of social unrest include appeal, accusation, refuse, escalation, and protest. These sequences are used to train models and eventually predict new social unrest in a country.

(ii) Second, we propose a novel hidden Markov model based framework which contains four major components: ground set extraction, burstiness modeling, HMM training, and event prediction. The ground set contains social unrest events that are significant enough to garner more-than-usual real-time coverage in mainstream news reporting. The temporal burst patterns of GDELT stream are taken as observations and modeled. Then, two HMM models are trained, with one for social unrest prone sequences and one for social unrest free sequences, after which new sequences' likelihoods are calculated and predictions are made by Bayes decision theory to specify the classification rule.

(iii) Third, as mentioned above, research works utilizing GDELT to predict social unrest events are at starting stage, though now GDELT has become the world's largest and high resolution event dataset. To the best of our knowledge, our work is the first to propose a practical HMM based pipeline to take advantage of the GDELT dataset, attempting to explore the predictive power of these massive coded event data. Experimental results demonstrate that the GDELT indeed reflects useful precursor indicators that reveal the causes or development of future events.

(iv) Last, considering that no work has focused on prediction social unrest in Southeast Asia, we conduct extensive experiment evaluations with GDELT event data from five main countries in this area. The proposed framework outperforms the logistic regression method by 7% to 27% and the baseline method 34% to 62% for different countries. Sensitivity analyses reveal the impact of the parameters on the new framework's performance. Moreover, our experiments fill the gap that no previous work had aimed at the district in Southeast Asia for conducting the social unrest event forecasting.

The paper is organized as follows: a coarse introduction of related work is provided in Section 2. Our HMM based social unrest event prediction framework is presented in Section 3. In Section 4, extensive experiments to evaluate the performance of the new model are conducted and analyzed. The work is summarized and conclusions are drawn in Section 5. In the last section, we give Appendix for technical discussion of Section 3.4.

2. Related Work

In this section, we will give a brief introduction of the existing works related to this paper, including researches on analysis of social unrest events and the guide to GDELT dataset.

2.1. Researches on Social Unrest Events. Current researches into the analysis of social unrest events can be categorized into two main types: event detection and event prediction.

Event detection provides users what is going on. It has long been addressed and is an extensively studied topic in the literature. Researchers utilize news or social networks, for example, twitter, as real-time and ubiquitous social sensors to promptly discover new events occurring. Document clustering techniques are used to identify events retrospectively or as the stories arrive [14]. Works like [15–17] focus on extraction patterns (templates) to extract information from text. For a survey on these detection techniques in twitter, we point the readers to [18]. However, these event detection approaches can only uncover events after they have occurred and are unable to predict future events because they all focus on observations that directly reflect currently occurring events, rather than precursor indicators that reveal the causes or development of future events [19].

Event prediction has been explored in a variety of applications, including elections [20, 21], disease outbreaks [22], stock market movements [23, 24], social unrest event prediction [11, 12, 25–31], movie earnings [23], crime [32], and failure prediction [33]. Most recent social unrest event prediction techniques can be categorized into three types: planned event forecasting, classification based prediction, and time series mining. Planned event prediction methods do not need to mine patterns from the previous data. They are based on the hypothesis that protests that are larger will be more disruptive and communicate support for its cause better than smaller protests. Mobilizing large numbers of people is more likely to occur if a protest is organized and the time and place are announced in advance [1, 26, 29]. Classification based prediction incorporates volume features and informative features such as semantic topics to train a classification model and then predicts the occurrence of future events. Several classification methods are utilized such as random forest [27], support vector machines [22], logistic regression [10, 11, 23, 25], and LASSO based logistic regression [12, 28]. Time series based mining uses temporal correlation of relevant features such as tweet volume by adopting appropriate approaches. For example, Achrekar et al. [34] used autoregressive modeling to predict flu trends using twitter data. Radinsky and Horvitz [30] utilized NYT news articles from 1986 to 2007 to build event chain and identify significant increases in the likelihood of disease outbreaks, deaths, and riots in advance of the occurrence of these events in the world.

2.2. The GDELT Dataset. The GDELT Project [7] is a real-time network diagram and database of global human society for open research which monitors the world's broadcast, print, and web news from nearly every corner of every country in over 100 languages and identifies the people, locations, organizations, counts, themes, sources, emotions, counts, quotes, and events driving our global society every second of every day, creating a free open platform for computing on the entire world. Each day the GDELT Project monitors the news media across nearly every corner of the world and compiles a list of over 300 categories of "events" from riots and protests to peace appeals and diplomatic exchanges, recording the details of the event, including its georeferenced location, into a master "event database" of more than a quarter-billion events, dating back to 1979 and updated each morning around 4AM EST. In particular, from 19 February, 2015, GDELT 2.0 has been online which updates every 15 minutes accessing the world's breaking events and reaction in near-real time.

In GDELT event data table, each record has 58 fields (61 fields in GDELT 2.0), capturing information pertaining to a specific event in CAMEO format [35]. In this paper, we use the following nine fields from a record: SQLDATE, MonthYear, EventRootCode, GoldsteinScale, NumMentions, AvgTone, ActionGeo_CountryCode, ActionGeo_Lat, and ActionGeo_Long. SQLDATE and MonthYear are the date the event took place in YYYYMMDD format and YYYYMM format, respectively. EventRootCode defines the root-level category the event code falls under. For example, code 1452 (engaging in violent protest for policy change) has a root code of 14 (PROTEST). This makes it possible to aggregate events at various resolutions of specificity. GoldsteinScale is a numeric score from −10 to +10, capturing the theoretical potential impact that type of event will have on the stability of a country. NumMentions is the total number of mentions of this event across all source documents, which can be used as a method of assessing the importance of an event: the more the discussion of that event is, the more likely it is to be significant. AvgTone is the average tone of all documents containing one or more mentions of this event. The score ranges from −100 (extremely negative) to +100 (extremely positive). ActionGeo_CountryCode is the location of the event, which is a 2-character FIPS10-4 country code for the location. ActionGeo_Lat and ActionGeo_Long are the centroid latitude and centroid longitude of the landmark for mapping.

The dataset is also available on Google Cloud Platform (https://cloud.google.com/) and can be accessed using Google BigQuery. In this paper, we export the following GDELT event data for the experiments from the Google BigQuery (https://bigquery.cloud.google.com/table/gdelt-bq:full.events?pli=1) web service.

3. HMMs-Based Social Unrest Events Prediction

3.1. Framework. Proactive reaction to social unrest events is at first glance closely coupled with social unrest event detection: an unrest event needs to be detected before the government can react to it. However to be precise, not the detection result but the eruption of a social unrest event is the kind of event that should be primarily avoided, which makes a big difference. Hence, it goes without saying that efficient proactive handling of social unrest events requires the prediction of the future level of social unrest, to judge whether the current situation bears the risk of a unrest event or not.

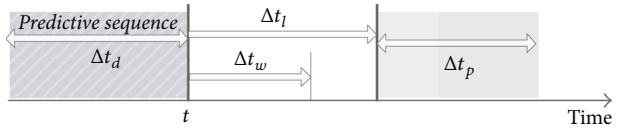


Figure 2: Prediction mechanism of upcoming social unrest events. t: *present time*; Δt_l: *lead time*; Δt_w: *warning time*; Δt_p: *prediction period*; Δt_d: *data window size*.

The basic assumption of our approach is that eruption of social unrest events can be identified by characteristic patterns of the event sequence prior to the happening time point using HMMs. Prediction mechanism of upcoming social unrest events is illustrated in Figure 2. If a prediction is performed at time t, we would like to know whether a social unrest event will occur or not between time $t + \Delta t_l$ and $t + \Delta t_l + \Delta t_p$.

Δt_l usually is called the *lead time*. Δt_l has a lower bound called *warning time* Δt_w, which is determined by the time needed for the specified organization like the government to perform some proactive action, for example, the time needed to make a public statement. Δt_d stands for the length of the data window called *data window size* which contains the predictive sequence of data. The sequence describes the current state of the country or district. The *prediction period* Δt_p is the length of the time interval for which the prediction holds.

Based on above prediction mechanism, our prediction task will resolve around predicting significant social unrest events on the country level and considering that country alone. To accurately predict social unrest events it is crucial to be able to characterize these events' underlying development before the occurrence by utilizing relevant GDELT event records observations. We propose a Hidden Markov Model based framework to characterize the underlying development of these events. Figure 3 illustrates the proposed HMMs-based social unrest event prediction framework, which contains four major components: ground set extraction, burstiness modeling, HMM training, and, last, event prediction.

Formally, denote ER as a basic GDELT event record. ER ("column Name") means the value of a specified column in a record. Denote $D = \{\mathrm{ER}_{c,t}\}_{c\in\Omega,t\in\Gamma}$ as a collection of GDELT event record data split into different countries Ω in time period Γ. The country c and the day t can be filtered by ER (ActionGeo_CountryCode) and ER (SQLDATE), respectively. Since event records ER are being added daily by the hundreds or thousands to the GDELT event table, we aggregate those event records by day, defined as $\mathrm{DAER}_{c,t}$, meaning the daily aggregated event record on the day t in country c. Then a sequence of DAERs is defined as $s = \{\mathrm{DAER}_{c,t}\}_{t\in T\subseteq\Gamma}$, which contains all the daily aggregated event records in country c in the time period $T \subseteq \Gamma$.

3.2. Ground Set Extraction. Ground truth is absolutely vital for the prediction problem. Unfortunately, until now there is no public ground set in the social unrest prediction area. As a result, in this paper we treat GDELT as the Ground Truth for social unrest events. Actually, the generated ground set does reflect the real world happenings well according to our manual inspection (see Figure 5).

For each country, the social unrest events we are interested in predicting are those that are significant enough to garner more-than-usual real-time coverage in mainstream news reporting for the country. That is, there is a significant social unrest event in country c on the day t. In GDELT, root event code 14 can be taken to mean social unrest. More records with event code "14" mean more social unrest event report coverage. For each country c we are interested in, we firstly aggregate the count of event mention with root event code 14 on each day t. Since new events are being added daily by the hundreds or thousands to the GDELT, there is a heterogeneous upward trend in the event mention and what is more than usual in counts changes. As a result, to remove the upward trend in the unrest event mentions, we normalize the mention counts with root code 14 by the average volume of the trailing quarter (90 days). That is, we let

$$M_{c,t} = \frac{\sum \mathrm{ER}_{c,t}\,(\mathrm{NumMentions}) : \mathrm{ER}\,(\mathrm{EventRootCode}) = 14}{(1/90)\sum_{j=t-90}^{t-1}\sum \mathrm{ER}_{c,j}\,(\mathrm{NumMentions}) : \mathrm{ER}\,(\mathrm{EventRootCode}) = 14}, \tag{1}$$

where $M_{c,t}$ is the normalized total count of social unrest event mentions on the day t in country c and ER (NumMentions) is the value of NumMentions of each record. Next we define the average event mention count on each day in country c as

$$\overline{M_c} = \frac{1}{|\Gamma|}\sum_{t\in\Gamma} M_{c,t}, \tag{2}$$

where Γ denotes the set of days in the training set.

To smooth the data we consider a seven-day moving average. By definition, we say that a significant social unrest in country c occurs during the 7-day stretches $t-3, t-2, \ldots, t+2, t+3$ if

$$M'_{c,t} = \frac{1}{7}\sum_{j=t-3}^{t+3}\frac{M_{c,j}}{\overline{M_c}} > \theta. \tag{3}$$

We set $\theta = \overline{M'_c} + 2.576 * \sqrt{(1/|\Gamma|)\sum_{t=1}^{|\Gamma|}(M_{c,t}/\overline{M_c} - \overline{M'_c})^2}$, which is the upper bound of the 99% confidence interval, where $\overline{M'_c} = (1/|\Gamma|)\sum_{t\in\Gamma}(M_{c,t}/\overline{M_c})$. The threshold is chosen so to select only significant social unrest events.

Table 1: The EventRootCode and descriptions.

EventRootCode	Description
10	*Demand*. Demand investigation, policy support, political reform, negotiation, etc.
11	*Disapprove*. Criticize, accuse, complain officially, lawsuit, rally opposition against, etc.
12	*Reject*. Refuse to yield, reject mediation, reject proposal, veto, defy norms, etc.
13	*Threaten*. Threaten nonforce, threaten political dissent, halt negotiations, etc.
14	*Protest*. Rally, strike, hunger strike, boycott, protest violently, riot, etc.

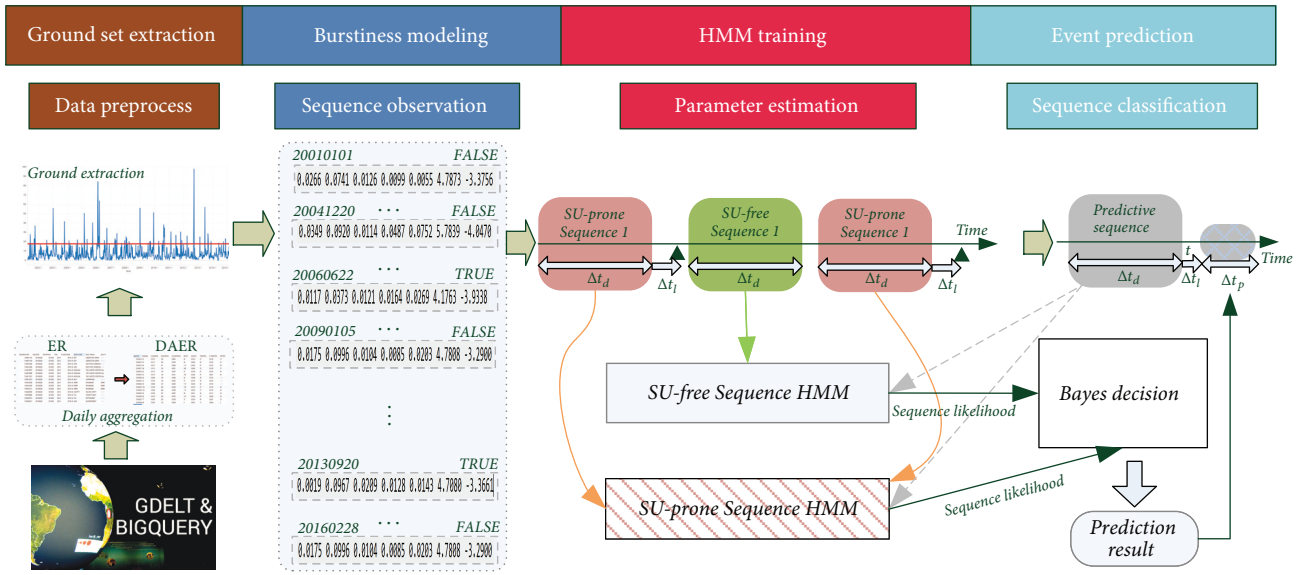


Figure 3: The proposed HMMs-based social unrest event prediction framework: two HMMs are trained, with one for SU-prone sequences and one for SU-free sequences. SU-prone sequences consist of observations within a time window of length Δt_d preceding a social unrest event (▲) by lead time Δt_l. SU-free sequences consist of observations at times when no social unrest event was imminent. t is the time the prediction is performed at. Δt_p is the prediction period.

3.3. Burstiness Modeling. The states of the social unrest event are unobserved but have a close theoretical analog in the concept of development stage that has been explicitly coded in the dataset. Usually, the social unrest event has its breeding development and evolution until the last occurrence, through a longer or shorter life cycle, meaning that it is usually not a sudden outbreak. Typical stages in the events' life cycle often include appeal, accusation, refuse, escalation, and protest. Of course, not every social unrest event will go through all of these stages. Our HMM model characterizes the developments of each significant social unrest event as a sequence of latent states, with a sequence of DAERs being the observations generated by the latent stages.

The GDELT event data captures various types of event owing to the CAMEO event code scheme, with EventRootCode field in the data table. In consideration of the stages of social unrest event, the following event types in Table 1 are added to our observations. The count of each type of those events can reflect signals of social unrest event development. Given a sequence of daily aggregated event records s, denote $r^e_{c,t}$ as the *ratio* of events with event root code e on the day t in country c:

$$r^e_{c,t} = \frac{\sum \mathrm{ER}_{c,t}\,(\mathrm{NumMentions}) : \mathrm{ER}\,(\mathrm{EventRootCode}) = e}{\sum_{j=1}^{20} \mathrm{ER}_{c,t}\,(\mathrm{NumMentions}) : \mathrm{ER}\,(\mathrm{EventRootCode}) = j}, \quad (4)$$

where $e = 10, 11, 12, 13, 14$ and the denominator means 20 event types in GDELT.

The observed variable O should include *ratio*s of the above five event types. In addition, we also add the mean value of ER (AvgTone) denoted as $\mathrm{at}_{c,t}$ and mean value of ER (GoldsteinScale) denoted as $\mathrm{gs}_{c,t}$ to the observation variable. Thus, observation O is a vector with 7 dimensions:

$$O_{c,t} = \left(r^{10}_{c,t}, r^{11}_{c,t}, r^{12}_{c,t}, r^{13}_{c,t}, r^{14}_{c,t}, \mathrm{at}_{c,t}, \mathrm{gs}_{c,t}\right). \quad (5)$$

3.4. HMM Training. Let $S = \{s_i\}$ denote the set of latent states, $1 \le i \le N$. Let $\pi = [\pi_i]$ denote the vector of initial state probabilities. Given a sequence of the above seven-dimension

vector observations O, a standard continuous HMM can be defined as $\lambda = (\pi, A, B)$. A is a $N \times N$ state transition probability matrix, where $A_{i,j} = P(s_j \mid s_i)$ is the transition probability of moving from the latent state s_i to latent state s_j. B is the emission probability matrix. The output probability for each state, $B_{i,t} = b_i(\mathbf{o}_t) = f(\mathbf{o}_t; \kappa_i)$, is a function of the observations $f(\mathbf{o}_t)$ that depends on model parameters κ_i. Here we use a Gaussian mixture output distribution:

$$b_i(\mathbf{o}_t) = \sum_{m=1}^{M} c_{im} \Phi(\mathbf{o}_t; \boldsymbol{\mu}_{im}, \mathbf{U}_{im}), \tag{6}$$

where M is the number of mixture components in the Gaussian mixture and $\sum_{m=1}^{M} c_{im} = 1$.

As shown in Figure 3, the goal of training process is generating both the social unrest event model SU and nonunrest event model $\overline{\text{SU}}$ model, that is, calculating parameters λ_{SU} and $\lambda_{\overline{\text{SU}}}$. We use the Baum-Welch expectation-maximization algorithm [13] for this purpose. The objective of the training algorithm is to optimize the HMM parameters π, A, and B such that the overall training sequence likelihood is maximized. *Sequence likelihood* is defined as the probability that a given HMM model λ can generate observation sequence $O = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_L)^T$:

$$P(O \mid \lambda) = \sum_{\mathbf{s}} \pi_1 b_{s_1}(\mathbf{o}_1) \prod_{t=2}^{L} P(S_t = s_t \mid S_{t-1} = s_{t-1}) b_{s_t}(\mathbf{o}_t), \tag{7}$$

where $\mathbf{s} = [s_t]$ denotes a sequence of latent states of length L. The sum over $\mathbf{s}$ denotes that all possible state sequences are investigated. However, this will result in unacceptable complexity especially when the observation sequence is long. Here we adopt forward algorithm or backward algorithm [13] to solve this issue. Denote the forward variable as α. We have

$$\begin{aligned} &\alpha_t(i) = P(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t, S_t = s_i \mid \lambda) \quad 1 \le i \le N \\ &\alpha_1(i) = \pi_i b_i(\mathbf{o}_1) \\ &\alpha_t(j) = \left[\sum_{i=1}^{N} \alpha_{t-1}(i) a_{ij}\right] b_j(\mathbf{o}_t) \\ &\qquad t = 2, 3, 4, \ldots, L, \ 1 \le j \le N. \end{aligned} \tag{8}$$

Finally, the sequence likelihood can be efficiently computed by

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_L(i). \tag{9}$$

A backward variable β is defined as

$$\begin{aligned} &\beta_t(i) = P(\mathbf{o}_{t+1}, \mathbf{o}_{t+2}, \ldots, \mathbf{o}_L \mid S_t = s_i, \lambda) \quad 1 \le i \le N \\ &\beta_L(i) = 1 \\ &\beta_t(i) = \sum_{j=1}^{N} a_{ij} b_j(\mathbf{o}_{t+1}) \beta_{t+1}(j) \\ &\qquad t = L-1, L-2, \ldots, 1, \ 1 \le i \le N. \end{aligned} \tag{10}$$

The sequence likelihood is

$$P(O \mid \lambda) = \sum_{i=1}^{N} \pi_i b_i(\mathbf{o}_1) \beta_1(i). \tag{11}$$

Using α and β together, let ξ denote the probability that a transition from latent state i to state j takes place at time t:

$$\begin{aligned} \xi_t(i, j) &= P(S_t = s_i, S_{t+1} = s_j \mid O, \lambda) \\ &= \frac{\alpha_{t-1} a_{ij} b_j(\mathbf{o}_t) \beta_t(j)}{P(O \mid \lambda)}. \end{aligned} \tag{12}$$

α, β, and ξ can be used to maximize the model parameters. The entire procedure of computation of α, β, ξ and subsequent maximization of model parameters are iterated until convergence, which will converge at least to a local maximum. Inequivalent to standard HMM that start from a randomly initialized HMM, we initial π and A according to the long history records of GDELT event data, which aims to reduce the randomness initialization of parameters. See Appendix for a more technical discussion.

Finally, we trained two HMM models based on two corresponding sets of sequences, one set from sequences prior to the positive 7-day stretches minus the lead time period and the other negative. Thus, one model characterizes the development process leading to a social unrest event, while the other one characterizes the process that does not lead to a social unrest event.

3.5. Event Prediction. After the training of model parameters, the social unrest event prediction is formalized as a sequence classification problem. For the prediction, an unknown sequence prior to the target 7-day stretch minus the lead time period will be aligned with the above model in each class. The sequence will be classified into the class corresponding to the higher alignment score, higher likelihood. However, likelihood $P(O \mid \lambda)$ gets small very quickly for long sequences, such that limits of double-precision floating point operations are reached. Scaling technique *log-likelihoods* is used for this reason. Besides, different costs should be associated with classification. For example, falsely classifying a SU-prone sequence as SU-free might be much worse than vice versa.

We use Bayes decision theory to specify the classification rule: the unknown sequence of observations O is classified as SU-prone, if

$$\log\left[P(O \mid \lambda_{\text{SU}})\right] - \log\left[P(O \mid \lambda_{\overline{\text{SU}}})\right] > \underbrace{\log\left[\frac{c_{\overline{\text{SU}},\text{SU}} - c_{\overline{\text{SU}},\overline{\text{SU}}}}{c_{\text{SU},\overline{\text{SU}}} - c_{\text{SU},\text{SU}}}\right] + \log\left[\frac{P(\overline{\text{SU}})}{P(\text{SU})}\right]}_{\varepsilon \in (-\infty, \infty)}, \tag{13}$$

where c_{ta} denotes the associated cost for assigning a sequence of type t to class a; for example, $c_{\text{SU},\overline{\text{SU}}}$ denotes the cost for falsely classifying a SU-prone sequence as SU-free. $P(\overline{\text{SU}})$ and $P(\text{SU})$ are constant representing the prior probabilities of SU sequences and $\overline{\text{SU}}$ sequences, respectively. See, for example, [36] for a derivation of the formula.

Figure 4: Mention counts of protest events occurring in Cambodia, Indonesia, Malaysia, Philippines, and Thailand extracted from GDELT between January 1, 2001, and February 29, 2016.

Thus, given the costs of misclassification, the right hand side of this inequality determines a constant threshold on the difference of sequence log-likelihoods, denoted as ε. If the threshold is small more sequences will be classified as SU-prone increasing the chance of detecting SU-prone sequences. On the other hand, the risk of falsely classifying a SU-free sequence as SU-prone is also high. If the threshold increases, the behavior is inverse: more and more SU-prone sequences will not be detected at a lower risk of false classification for SU-free sequences.

4. Experimental Evaluation

This section presents an experimental evaluation of the performance of the proposed HMM based prediction approach based on comprehensive experiments on GDELT event data from five main countries from Southeast Asia.

4.1. Experiment Design

4.1.1. Dataset. Our goal in this paper is to predict the overall level of social unrest using GDELT, and our focus area is distributed across five major nations in Southeast Asia: Thailand, Malaysia, Philippines, Indonesia, and Cambodia. Numerous event records extracted from online media and frequent protests or strikes throughout these countries make them ideal countries to study patterns and signals prior to the happening of social unrest events. As mentioned above, GDELT uses the CAMEO coding system [35], where root event code 14 can be taken to mean social unrest. Figure 4 illustrates the mention counts of protest event occurring in these countries retrieved from GDELT between January 1, 2001, and February 29, 2016. The average counts of protest events per year for each country range from 480 in Cambodia to 1700 in Thailand. In consideration of the quarterly normalization in Section 3.2, the actual training data was from April 1, 2001, to December 31, 2013, and the test data January 1, 2014, to February 29, 2016.

Table 2: Number of positive 7-day stretches in the 778 weeks of our experiment in different countries.

Country	# of positive 7-day stretches	
	Training	Testing
Thailand	95	12
Malaysia	78	8
Philippines	85	9
Indonesia	83	7
Cambodia	88	7

4.1.2. Ground Set. The ground set was generated as the manner described in Section 3.2. Overall across the five countries considered, about 11.5% of 7-day stretches are labeled positive, distributed mostly evenly among the countries. The whole training and testing period include 5448 days and 778 weeks. The number of positive 7-day stretches in the 778 weeks with training and testing period, respectively, on different countries is shown in Table 2. The training period includes 666 7-day stretches while the testing period 112. An example plot of ground set for Thailand is shown in Figure 5 with annotations of news abstract describing the social unrest event in the top ten stretches above threshold.

4.1.3. Comparison Methods. We compare the proposed HMM based social unrest event prediction method with logistic regression *(LogReg)* model and a *baseline* method. The *LogReg* model [32] also treats the event prediction as a classification problem. The input feature here is the sum of event mentions of each type in the predictive sequences during the period of δt_d. The output is 0 if there is no event and 1, if there is one. The *baseline* method considers the probability of historical social unrest event occurrence to be the probability of future social unrest event occurrence. Note that this baseline is also used as the prior parameter in the training process of the HMM models.

4.1.4. Performance Metrics. We evaluate our social unrest event prediction framework using metrics similar to those described in Kallus [27]. We quantify the success of the proposed predictive mechanism and comparison methods based on their balanced accuracy. Let $T_{ct} \in \{0, 1\}$ and $P_{ct} \in \{0, 1\}$, respectively, denote whether a significant social unrest event occurs in country c during the days $t-3$, $t-2$, $t-1$, t, $t+1$, $t+2$, and $t+3$ and whether we predict there to be one. The true positive rate (TPR) is the fraction of positive instances ($T_{ct} = 1$) correctly predicted to be positive ($P_{ct} = 1$) and the true negative rate (TNR) is the fraction of negative instances predicted negative. The balanced accuracy (BACC) is the unweighted average of these:

$$\text{BACC} = \frac{\text{TPR} + \text{TNR}}{2}. \tag{14}$$

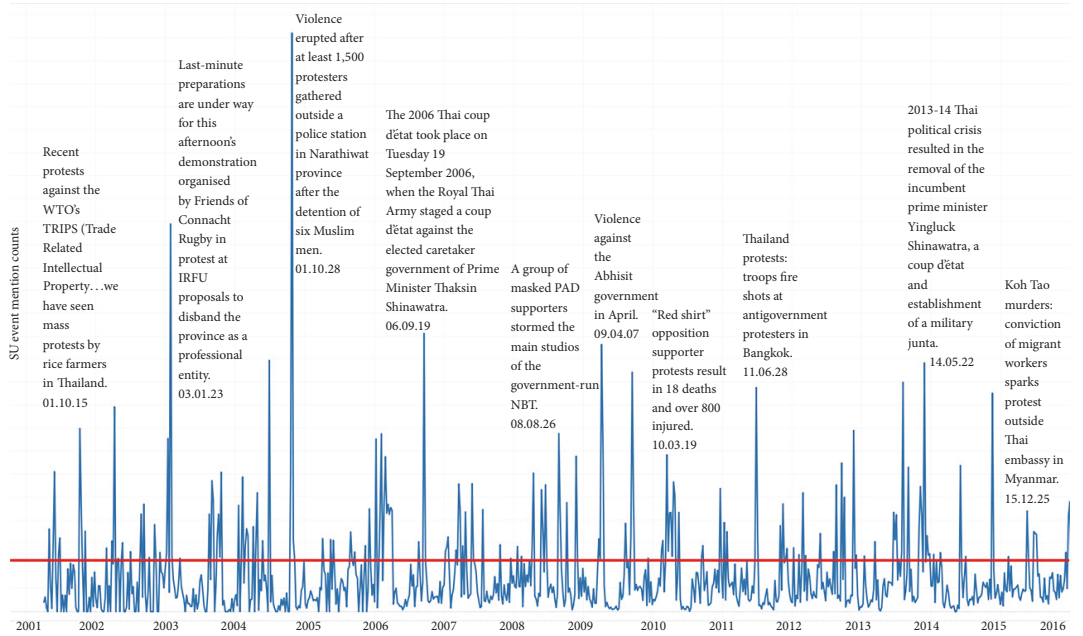


Figure 5: Normalized SU event mention counts of Thailand with annotations for top ten stretches above θ (red line).

BACC, unlike the marginal accuracy, cannot be artificially inflated. In fact, as the unbalanced distribution of positive and negative examples in our dataset, always predicting "no social unrest event" without using any data will yield a nearly 90% marginal accuracy but only 45% balanced accuracy. In fact, a prediction without any relevant data will always yield a BACC of 50% on average by statistical independence.

4.1.5. Parameter Settings. The *baseline* method does not require any parameters and we implemented the *LogReg* method based on its origin. The proposed HMM based prediction method has four prior parameters: prediction period Δt_p, dimension of observation O, the number of latent states N, the number of Gaussian mixtures, and three tunable parameters: lead time Δt_l, data window size Δt_d, and threshold ε. We used a prediction period Δt_p of seven days (one week) in our experiments. The number of latent states and the number of Gaussian mixtures were set as 5 and 3, respectively. The tunable parameters are estimated based on the tenfold cross-validation by maximizing the average balanced accuracy of the five countries. The *lead time* Δt_l, the *data window size* Δt_d, that is, the sequence length, and the threshold ε were set to be 1, 10, and 6, respectively. Finally, we use the open-source HMM toolbox developed by Murphy; see [37], to implement the various HMM functions.

4.2. Event Prediction Results. Figure 6 compares our proposed HMM based prediction method to the *LogReg* model and the *baseline* method based on the BACC metric. In every case in the figure, we note that, for all the five countries, our proposed approach achieved the best overall performance in balanced accuracy, outperforming the *LogReg* model by 27%, 17%, 7%, 15%, and 7% and the *baseline* 62%, 39%, 45%, 43%, and 34% for Thailand, Indonesia, Philippines, Malaysia, and Cambodia, respectively. This is likely because our HMM based prediction framework better captures the features and characterizes the development stages of social unrest events behind the observed sequence data. The poor performance of the *baseline* method, actually close to a totally random model, indicates that focusing solely on the probability of historical social unrest event occurrence is insufficient for the task of social unrest event prediction.

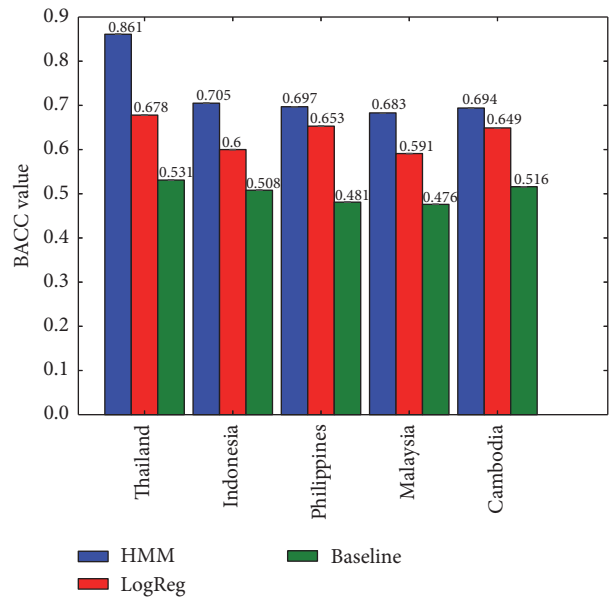


Figure 6: Comparison of our HMM based method with the *LogReg* model and the *baseline* method based on the BACC metric.

Table 3: State transition matrix (25 cells).

	Appeal	Accusation	Refuse	Escalation	Eruption
Appeal	10 → 10	10 → 11	10 → 12	10 → 13	10 → 14
Accusation	11 → 10	11 → 11	11 → 12	11 → 13	11 → 14
Refuse	12 → 10	12 → 11	12 → 12	12 → 13	12 → 14
Escalation	13 → 10	13 → 11	13 → 12	13 → 13	13 → 14
Eruption	14 → 10	14 → 11	14 → 12	14 → 13	14 → 14

One of the advantages of our HMM prediction method is that it allows employing a customizable threshold permitting to control the tradeoff between the true positive rate (TPR) and false positive rate (FPR = 1 − TNR). As we vary the threshold ε, we can monotonically trade off TPR with FPR. The range of achievable such rates for each country is plotted in Figure 7, with the HMM based prediction method and the *LogReg* method in comparison. Also, the HMM based social unrest event prediction model outperforms the *LogReg* model for each individual country, with the areas under the curve (AUC) of HMM method for each country are obviously bigger than that of *LogReg* method for each country. In particular, the prediction task for Thailand achieves best performance obviously. This is probably because Thailand experienced more massive social unrest events and thus, more patterns of development were learned.

4.3. Sensitivity Analysis on Δt_l and Δt_d. Although we set fixed values for parameters in the comparison in last section, the impact of the number of days of lead times, that is, the parameter Δt_l, and the data window size Δt_d on the event prediction performance for each country were also studied. We turned Δt_l from 1 to 10 and chose three data window sizes Δt_d: 10, 20, and 30. The detailed variation tendencies are illustrated in Figure 8, leading to two main observations. Firstly, overall, the prediction balanced accuracy decreases as the days of lead time increase for all the data window sizes in each country, which indicates that the performance is sensitive to the number of days of lead time in the given value interval of parameters. In most cases, the lead time of 1 day achieves best performance. This is consistent with the common understanding that the more close to one the social unrest event is the more probable it may be predicted. Secondly, as shown by these curves, the balanced accuracy and the data window size do not have a relationship with obvious trend. It depends on specified lead time and specified country. For example, for Thailand, $\Delta t_d = 10$ with $\Delta t_l = 1$ performs best while $\Delta t_d = 30$ performs best with other lead times. For other countries, this relationship takes on a different situation. This reflects that we should use different data window sizes with different lead time and different countries to achieve best prediction performance.

5. Discussion

This paper presents a hidden Markov models based framework for leveraging large scale digital history coded events captured from GDELT to utilize the temporal burst patterns in GDELT event streams to uncover the underlying event development mechanics and formulate the social unrest event prediction as a sequence classification problem. Extensive empirical testing with data from five countries in Southeast Asia demonstrated the effectiveness of this framework by comparing it with logistic regression model and the baseline model and the fact that the GDELT dataset does reflect some useful precursor indicators that reveal the causes or development of future events.

We plan to conduct our future work in the following four aspects. First, we will apply this proposed framework to the city level prediction within a country. Second, we want to add other informative data like Twitter and Facebook to enhance the prediction accuracy. In addition, in GDELT 2.0, event mention details and global knowledge graphs [38] are also provided real-timely, which can bring us with detailed insights to the events. Third, we also plan to label a Ground Truth dataset for social unrest events in Asia like the Gold Standard Report (GSR) [12] for Latin American to better evaluate our future methods. Last, in this paper we do not consider the geographical factor which also affects the event coverage. Next we will improve our model to distinguish widespread news coverage from localized coverage.

Appendix

Technical Details

We initialize the HMM parameters π and A according to the long history records of GDELT event data. The input event data stream is sorted by time and every neighbouring two event records form a paired event value. After entering the paired event values in the corresponding cells of the state transition matrix, see Table 3; we count each of the event type pairs in the input stream for each cell and divide by the total to derive the initial state transition A. In addition, we simply count the records of each type to derive the π, that is, initial probabilities of each latent state. The obvious alternative is to treat all state transitions and each element in π as equally probably (0.2 each on a five-state model). However, we consider our initial values, derived from the data, to be more useful in that they are bound to our understanding of the observed events in our data; that is, we consider them as actual pairs of events and, from this empirical understanding, infer their most appropriate hidden state context.

Table 4 shows the initial state transition matrix A and initial state probabilities π for Thailand, from which we can draw two empirical conclusions from the historical event

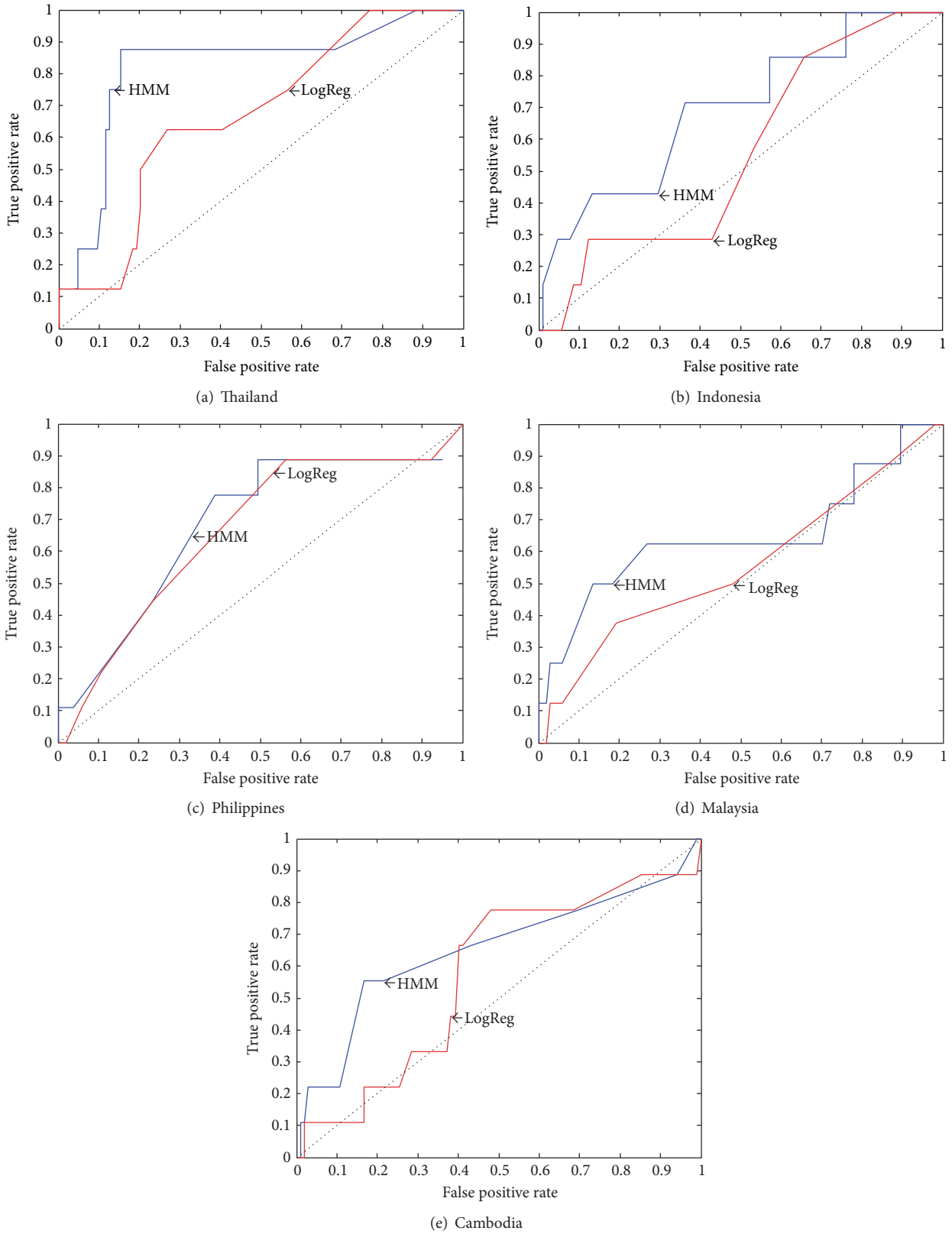
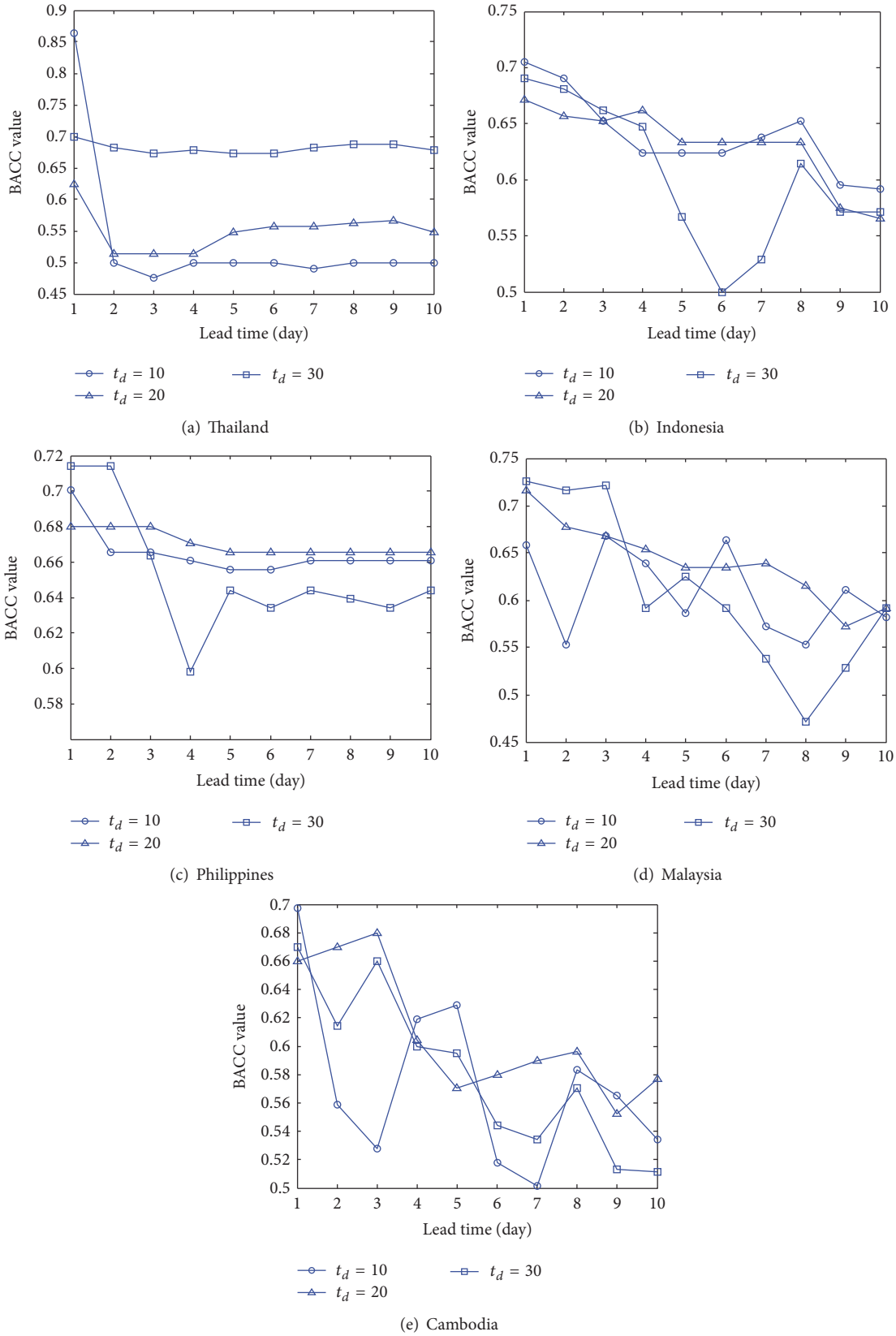


Figure 7: ROC curves for the compared prediction models. The HMM based social unrest event prediction model outperforms the *LogReg* model for each individual country.



(a) Thailand

(b) Indonesia

(c) Philippines

(d) Malaysia

(e) Cambodia

Figure 8: Sensitivity analysis on lead times Δt_l and data window size Δt_d.

Table 4: Initial state transition matrix A and initial state probabilities π for Thailand.

(a) A

0.2781	0.4172	0.1339	0.0877	0.0831
0.0999	0.5758	0.1538	0.0897	0.0808
0.1035	0.3398	0.3347	0.1196	0.1024
0.1117	0.3506	0.1178	0.2810	0.1389
0.1153	0.3362	0.1210	0.0878	0.3397

(b) π

0.1263
0.4586
0.1746
0.1168
0.1237

records. First, the transition probability from state S_i to S_i is the biggest. This reveals the fact that each type of events often lasts for a period of time. Second, for each state, its second biggest transition probability comes from its neighbouring state, which means that the evolution of event stages follows some certain rules, not a random process.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work is supported by the National Science Foundation of China (NSFC) under Grant nos. 71331008 and 61105124.

References

[1] S. Muthiah, B. Huang, J. Arredondo et al., "Planned protest modeling in news and social media," in *Proceedings of the 29th AAAI Conference on Artificial Intelligence (IAAI '15)*, pp. 3920–3927, 2015.

[2] C. A. McClelland, *World-Event-Interaction-Survey: A Research Project on the Theory and Measurement of International Interaction and Transaction*, University of Southern California, 1967.

[3] E. E. Azar, "The conflict and peace data bank (COPDAB) project," *Journal of Conflict Resolution*, vol. 24, no. 1, pp. 143–152, 1980.

[4] D. Bond, J. C. Jenkins, C. L. Taylor, and K. Schock, "Mapping mass political conflict and civil society: issues and prospects for the automated development of event data," *Journal of Conflict Resolution*, vol. 41, no. 4, pp. 553–579, 1997.

[5] S. P. O'Brien, "Crisis early warning and decision support: Contemporary approaches and thoughts on future research," *International Studies Review*, vol. 12, no. 1, pp. 87–104, 2010.

[6] B. Kettler and M. Hoffman, "Lessons learned in instability modeling, forecasting, and mitigation from the darpa integrated crisis early warning system (icews) program," in *Proceeding of the 2nd International Conference on Cross-Cultural Decision Making: Focus*, 2012.

[7] K. Leetaru and P. A. Schrodt, "Gdelt: global data on events, location, and tone 1979–2012," in *ISA Annual Convention*, vol. 2, Citeseer, 2013.

[8] E. Alikhani, *Computational Social Analysis: Social Unrest Prediction Using Textual Analysis of News*, State University of New York at Binghamton, Binghamton, NY, USA, 2014.

[9] J. E. Yonamine, "Predicting future levels of violence in afghanistan districts using gdelt," Inpress.

[10] F. Qiao and H. Wang, "Computational approach to detecting and predicting occupy protest events," in *Proceedings of the 4th International Conference on Identification, Information, and Knowledge in the Internet of Things (IIKI '15)*, pp. 94–97, IEEE, Beijing, China, October 2015.

[11] Y. Keneshloo, J. Cadena, G. Korkmaz, and N. Ramakrishnan, "Detecting and forecasting domestic political crises: a graph-based approach," in *Proceedings of the 6th ACM Web Science Conference (WebSci '14)*, pp. 192–196, ACM, June 2014.

[12] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, and N. Ramakrishnan, "Combining heterogeneous data sources for civil unrest forecasting," in *Proceedings of the IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '15)*, pp. 258–265, Paris, France, August 2015.

[13] L. R. Rabiner, "Tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 1989.

[14] E. Gabrilovich, S. Dumais, and E. Horvitz, "Newsjunkie: providing personalized newsfeeds via analysis of information novelty," in *Proceedings of the 13th International World Wide Web Conference (WWW '04)*, pp. 482–490, ACM, New York, NY, USA, May 2004.

[15] S. Lightner, R. Dave, S. Boddhu, and R. Flagg, "Event detection through text analysis using trained event template models, 2016," US Patent 20,160,019,470.

[16] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, vol. 2, pp. 1003–1011, Association for Computational Linguistics, August 2009.

[17] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, and D. Jurafsky, "Stanford's multi-pass sieve coreference resolution system at the conll-2011 shared task," in *Proceedings of the 15th Conference on Computational Natural Language Learning: Shared Task*, pp. 28–34, Association for Computational Linguistics, 2011.

[18] F. Atefeh and W. Khreich, "A survey of techniques for event detection in Twitter," *Computational Intelligence*, vol. 31, no. 1, pp. 132–164, 2015.

[19] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Spatiotemporal event forecasting in social media," in *Proceedings of the SIAM International Conference on Data Mining (SDM '15)*, vol. 15, pp. 963–971, May 2015.

[20] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: linking text sentiment to public opinion time series," in *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM '10)*, pp. 122–129, Washington, DC, USA, May 2010.

[21] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with Twitter: what 140 characters reveal about political sentiment," in *Proceedings of the 4th International*

AAAI Conference on Weblogs and Social Media (ICWSM '10), pp. 178–185, May 2010.

[22] J. Ritterman, M. Osborne, and E. Klein, "Using prediction markets and twitter to predict a swine flu pandemic," in *Proceedings of the 1st International Workshop on Mining Social Media*, vol. 9, pp. 9–17, 2009.

[23] M. Arias, A. Arratia, and R. Xuriguera, "Forecasting with twitter data," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 1, article 8, 2013.

[24] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.

[25] R. Compton, C. Lee, J. Xu et al., "Using publicly visible social media to build detailed forecasts of civil unrest," *Security Informatics*, vol. 3, no. 1, pp. 1–10, 2014.

[26] S. Muthiah, *Forecasting protests by detecting future time mentions in news and social media [M.S. thesis]*, 2014.

[27] N. Kallus, "Predicting crowd behavior with big public data," in *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, pp. 625–630, International World Wide Web Conferences Steering Committee, Seoul, Korea, April 2014.

[28] J. Cadena, G. Korkmaz, C. J. Kuhlman, A. Marathe, N. Ramakrishnan, and A. Vullikanti, "Forecasting social unrest using activity cascades," *PLoS ONE*, vol. 10, no. 6, Article ID e0128879, 2015.

[29] N. Ramakrishnan, P. Butler, S. Muthiah et al., "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)*, pp. 1799–1808, August 2014.

[30] K. Radinsky and E. Horvitz, "Mining the web to predict future events," in *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM '13)*, pp. 255–264, February 2013.

[31] V. B. Petroff, J. H. Bond, and D. H. Bond, "Using hidden markov models to predict terror before it hits (again)," in *Handbook of Computational Approaches to Counterterrorism*, pp. 163–180, Springer, 2013.

[32] X. Wang, M. S. Gerber, and D. E. Brown, "Automatic crime prediction using events extracted from twitter posts," in *Social Computing, Behavioral-Cultural Modeling and Prediction*, pp. 231–238, Springer, 2012.

[33] F. Salfner, "Predicting failures with hidden markov models," in *Proceedings of 5th European Dependable Computing Conference (EDCC '05)*, pp. 41–46, 2005.

[34] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Predicting flu trends using twitter data," in *Proceedings of the IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS '11)*, pp. 702–707, IEEE, Shanghai, China, April 2011.

[35] D. J. Gerner, P. A. Schrodt, O. Yilmaz, and R. Abu-Jabr, *Conflict and Mediation Event Observations (Cameo): A New Event Data Framework for the Analysis of Foreign Policy Interactions*, International Studies Association, New Orleans, La, USA.

[36] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, 2012.

[37] K. Murphy, "Hidden markov model (hmm) toolbox for matlab," 1998, https://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html.

[38] Gdelt 2.0: Our global world in realtime http://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/.