Hindawi
Discrete Dynamics in Nature and Society
Volume 2020, Article ID 3915036, 16 pages
https://doi.org/10.1155/2020/3915036

Research Article

Learning Evolutionary Stages with Hidden Semi-Markov Model for Predicting Social Unrest Events

Fengcai Qiao,[1] **Xin Zhang**,[2] **and Jinsheng Deng**[1]

[1]*College of Advanced Interdisciplinary Studies, National University of Defense Technology, Changsha 410073, Hunan, China*
[2]*College of System Engineering, National University of Defense Technology, Changsha 410073, Hunan, China*

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

Received 7 July 2020; Revised 14 September 2020; Accepted 18 September 2020; Published 9 October 2020

Academic Editor: Ricardo López-Ruiz

Social unrest events are common happenings in modern society which need to be proactively handled. An effective method is to continuously assess the risk of upcoming social unrest events and predict the likelihood of these events. Our previous work built a hidden Markov model- (HMM-) based framework to predict indicators associated with country instability, leaving two shortcomings which can be optimized: omitting event participants' interaction and implicitly learning the state residence time. Inspired by this, we propose a new prediction framework in this paper, using frequent subgraph patterns and hidden semi-Markov models (HSMMs). The feature called BoEAG (Bag-of-Event-Association-subGraph) is constructed based on frequent subgraph mining and the bag of word model. The new framework leverages the large-scale digital history events captured from GDELT (Global Data on Events, Location, and Tone) to characterize the transitional process of the social unrest events' evolutionary stages, uncovering the underlying event development mechanics and formulating the social unrest event prediction as a sequence classification problem based on Bayes decision. Experimental results with data from five main countries in Southeast Asia demonstrate the effectiveness of the new method, which outperforms the traditional HMM by 5.3% to 16.8% and the logistic regression by 11.2% to 43.6%.

1. Introduction

The era of information technology boosts the rapid development of the Internet of things, social media, and big data. As a data-intensive science, social computing is an emerging thing that leverages the capacity to collect and analyze data with an unprecedented breadth, depth, and scale. It represents a new computing paradigm and an interdisciplinary research and application field. Topics related to social computing have attracted the attention of more and more researchers.

The social unrest events such as protests, strikes, demonstrations, and occupy movements are important research focuses in the social computing area, which 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 for example; a series of political protests and three military coups happened between 1990 and 2015, resulting in the government being deposed, illustrating the power of the social unrest. Figure 1 depicts the activities that causally preceded the protest against the amnesty bill in Bangkok on 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.

Traditionally, the research in the area of social unrest was based on static analysis from the macroqualitative

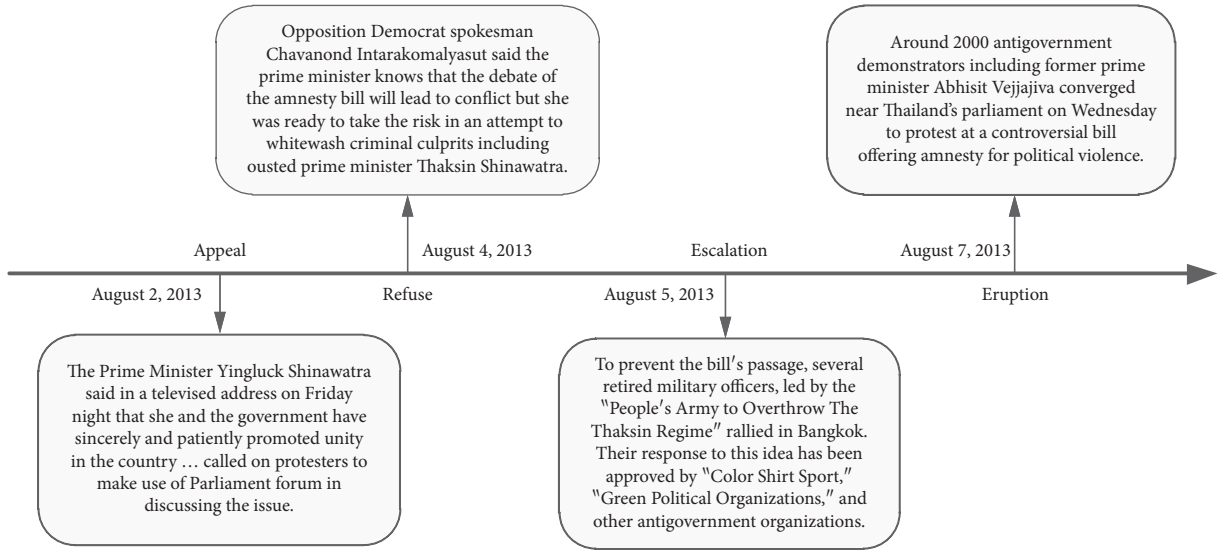

Figure 1: Event development stages before the protest against the amnesty bill on August 7, 2013, in Bangkok, Thailand.

perspective by the political researchers. Fortunately, with the development of data science, especially the rise of big data, there are more and more data-driven approaches proposed on microscopic insight into possible social unrest events. Last century, most researchers conducted the prediction work using human-coded data, including WEIS [2] and COPDAB [3]. In the recent two decades, several small-scale vertical machine-readable datasets [4, 5] and large-scale coded event data like ICEWS (Integrated Crisis Early Warning System) [6] and GDELT [7] appeared, fueling the development of computation methods for the analysis and prediction of social unrest.

Our previous work [8] published in Discrete Dynamics in Nature and Society built a hidden Markov model- (HMM-) based framework to predict indicators associated with country instability. The framework used the temporal burst patterns in GDELT event streams as features to train the hidden Markov models. There are two shortcomings in that work. First, the temporal burst pattern is essentially a simple feature in the number of coded events. The interaction characteristics between event participants are missing. Second, the probability of state residence time in the HMMs decreases exponentially with time, which is obviously not in line with the actual situation of social unrest events.

In response to the above shortcomings, we propose a new prediction framework in this paper, using frequent subgraph patterns and hidden semi-Markov models (HSMMs). The new framework also leverages the large-scale digital history events captured from GDELT to characterize the transitional process of the social unrest events' evolutionary stages. Our proposed framework converts the GDELT event streams to frequent subgraph patterns for capturing interaction features better. In addition, the mechanism of HSMM guarantees the prediction model can explicitly learn the probability distribution of state residence time from the historical data. Eventually, the social unrest event prediction is formulated as a sequence classification problem using Bayes decision. More concretely, our main contributions in this updated paper are four pronged:

(i) First, we identify a sequence of stages of events that potentially lead to a social unrest. Typical evolutionary stages of social unrest include appeal, accusation, refuse, escalation, and eruption, where each stage corresponds to a state in the hidden semi-Markov model. It should be noted that not all unrest events will go through all the four development stages before reaching the eruption stage.

(ii) Second, we propose the BoEAG (Bag-of-Event-Association-subGraph) features to capture the characteristics of frequent patterns instead of the temporal burst patterns used in our previous work [8]. The original GDELT data within a certain time are represented as an event element association graph, from which the frequent subgraph patterns are mined. In the end, the BoEAG features are constructed like the classic BoW (bag of word) model [9] used in the text processing.

(iii) Third, we propose a hidden semi-Markov model-based framework which contains four major components: ground set extraction, BoEAG feature construction, HSMM 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 BoEAG features of the GDELT stream are taken as the observations. Then, two HSMM 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.

(iv) Last, we conduct extensive experiment evaluations with GDELT event data from five main countries in

Southeast Asia. The proposed framework outperforms the traditional HMM by 5.3% to 16.8% and the logistic regression method by 11.2% to 43.6% for different countries. Sensitivity analyses are also conducted, revealing the impact of the parameters on the new framework's performance.

The paper is organized as follows. A coarse introduction of related work is provided in Section 2. Our HSMM-based social unrest event prediction framework is presented in Section 3. In Section 4, extensive experiments to evaluate the performance of the new method are conducted and analyzed. The work is summarized and conclusions are drawn in Section 5.

2. Related Work

2.1. Social Unrest Event Prediction. Predictive analysis of social unrest events has long stayed at the level of qualitative analysis relying on the experience of experts, especially political scientists. Since 2009, research studies on social unrest event prediction based on data mining have taken shape in some international political science journals [5, 10]. Especially since 2013, with the popularization of big data technology, big data-driven social unrest event prediction research has ushered in a period of vigorous development. In the conferences such as SIGKDD [11, 12], WWW [13, 14], SDM [15], AAAI [1, 16], and journals such as IEEE Trans. [17, 18], more than 30 related works have been published in succession, and the degree of attention is evident.

Event prediction has been explored in a variety of applications, including elections [19, 20], disease outbreaks [21], stock market movements [22, 23], social unrest event prediction [11, 13, 24–31], movie earnings [22], 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 will 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, 11, 25]. For example, Basnet et al. [34] used the GDELT data to propose a clustering method based on spatiotemporal k-dimensional structure trees to study the spatiotemporal distribution of conflict events in India in 2014.

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 [13], support vector machines [21], logistic regression [22, 24, 28, 35] and LASSO-based logistic regression [26, 27]. Wang et al. [36] used the LSTM model combined with GDELT's event data to predict the number of conflicting events. Yang et al. [37] used a two-stage sentiment analysis method based on deep neural networks to conduct early warning research on group aggregation behavior. Phillips [38] summarized the use of social media to predict future events, including applied research in the detection of political events and threat events. Parrish [39] used the recurrent neural network GRU sequence model and aggregated the GDELT event data by day, splicing them into feature vectors to determine whether a country has a social unrest event including domestic political crisis, riots, racial violence, and change of leadership. Zhao et al. [40] used the multitask learning of geographical spatial stratification, judging whether unrest events occurred on the specified date. Wu et al. [41] used the "Protest Participation Theory" proposed in the field of political science, combined with the SVM support vector machine model, to conduct early warning research on social unrest events. Deng et al. [12] extracted and learned graph representations from historical event documents. By employing the hidden word graph features, the model predicts the occurrence of future events and identifies sequences of dynamic graphs as event context.

Time series-based mining uses temporal correlation of relevant features such as tweet volume by adopting appropriate approaches. For example, Achrekar et al. [42] used autoregressive modeling to predict flu trends using twitter data. Radinsky et al. [29] 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.

So far, there are few works aiming at utilizing GDELT to make predictions about social unrest. Existing works attempted to use linear regression [43], time series forecasting [44], deep neural networks [36, 39], and frequent subgraphs [28, 35] to conduct the prediction work using GDELT. In [27], 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 evolutionary stages in the prediction models.

2.2. Hidden Semi-Markov Model. A hidden semi-Markov model (HSMM) is a statistical model with the same structure as a hidden Markov model except that the probability of there being a change in the hidden state depends on the amount of time that has elapsed since entry into the current state. This is in contrast to the original hidden Markov models where there is a constant probability of changing state given survival in the state up to that time.

HSMM was first proposed by Baum et al. [45] and has been successfully used in many applications, including word recognition task [46], daily return series modeling in financial market [47, 48], equipment health diagnosis and prognosis [49], activity recognition and abnormality detection [50], DNA analysis [51], and online failure prediction [52]. It is worth noting that in our work, we referred to the basic idea of online failure prediction in a commercial telecommunication system by Salfner et al. [33, 52]. The works motivate us to apply the hidden Markov model and hidden semi-Markov model to the social unrest event

prediction task. The prediction mechanism and Bayes decision-based classification are adopted specifically.

2.3. 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 4 AM 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 [53]. 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 (engage 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 discussion of that event, 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 Platform1 and can be accessed using Google BigQuery. In this paper, we export the following GDELT event data for the experiments from the Google BigQuery2 web service.

3. HSMM-Based Social Unrest Event 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, the fact is that 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 an unrest event or not.

The evolutionary stages of the social unrest event cannot be directly observed. However, the stages have been explicitly coded more or less on the Internet. The basic assumption of our approach is that the eruption of social unrest events can be identified by frequent subgraph patterns of the event sequence prior to the happening time point using HSMMs. Prediction mechanism of the 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 $t + \Delta t_l$ to $t + \Delta t_l + \Delta t_p$.

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, e.g., 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 the 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 stage before the occurrence by utilizing relevant GDELT event records observations. We propose a hidden semi-Markov model-based framework to characterize the underlying development of these events. Figure 3 illustrates the proposed HSMM-based social unrest event prediction framework, which contains four major components: ground set extraction, BoEAG feature construction, HSMM training, and 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 $D = \left\{\mathrm{ER}_{c,t}\right\}_{c\in\Omega,t\in\Gamma}$ 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(Action Geo_Country Code) 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 $s = \left\{\mathrm{DAER}_{c,t}\right\}_{t\in T\subseteq\Gamma}$, 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.

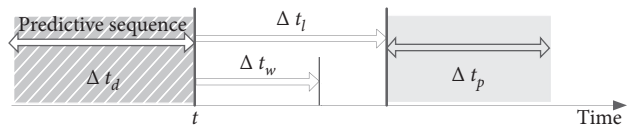

FIGURE 2: Prediction mechanism of upcoming social unrest events. t: present time; Δt_l: lead time; Δt_w: warning time; Δt_p: prediction period; Δt_d: data window size.

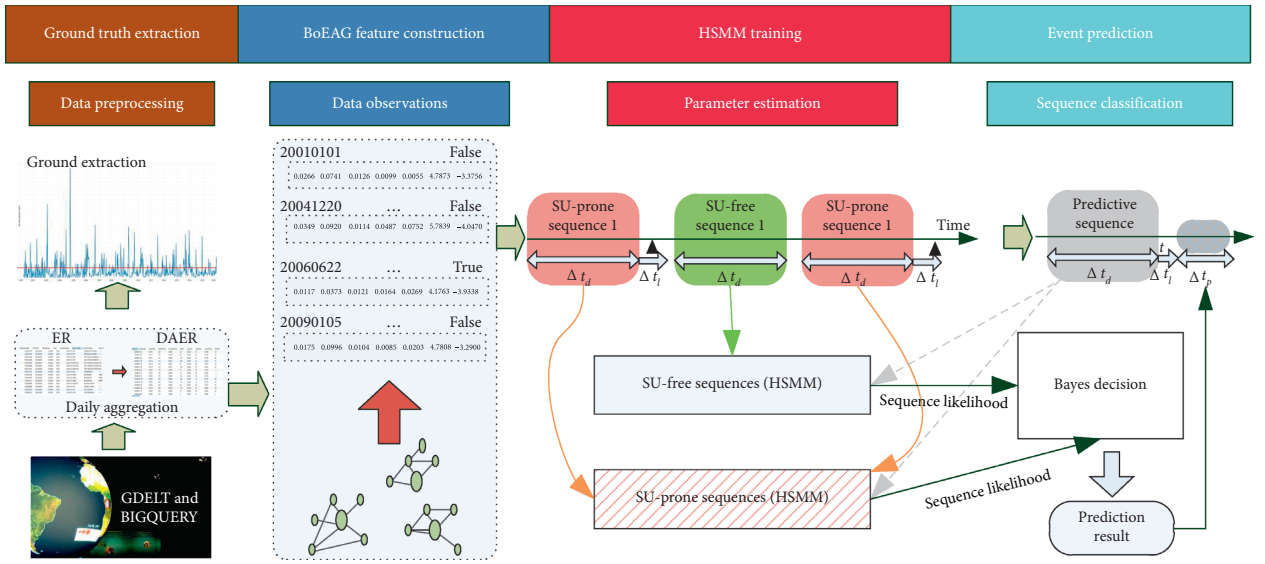

FIGURE 3: The proposed HSMM-based social unrest event prediction framework: two HSMMs are trained, with one for SU-prone sequences and one for SU-free sequences. SU-prone sequences consist of observations (BoEAG feature) within a time window of length Δt_d preceding a social unrest event (▲) by lead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

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 4).

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 means 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 count 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

$$M_{c,t} = \frac{\sum \mathrm{ER}_{c,t}\,(\text{Num Mentions}):\ \mathrm{ER}\,(\text{Event Root Code}) = 14}{(1/90)\sum_{j=t-90}^{t-1} \mathrm{ER}_{c,j}\,(\text{Num Mentions}):\ \mathrm{ER}\,(\text{Event Root Code}) = 14}, \tag{1}$$

where $M_{c,t}$ is the normalized total count of social unrest event mentions on the day t in country c and ER(Num Mentions) is the value of Num Mentions 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

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

where θ is the significance threshold.

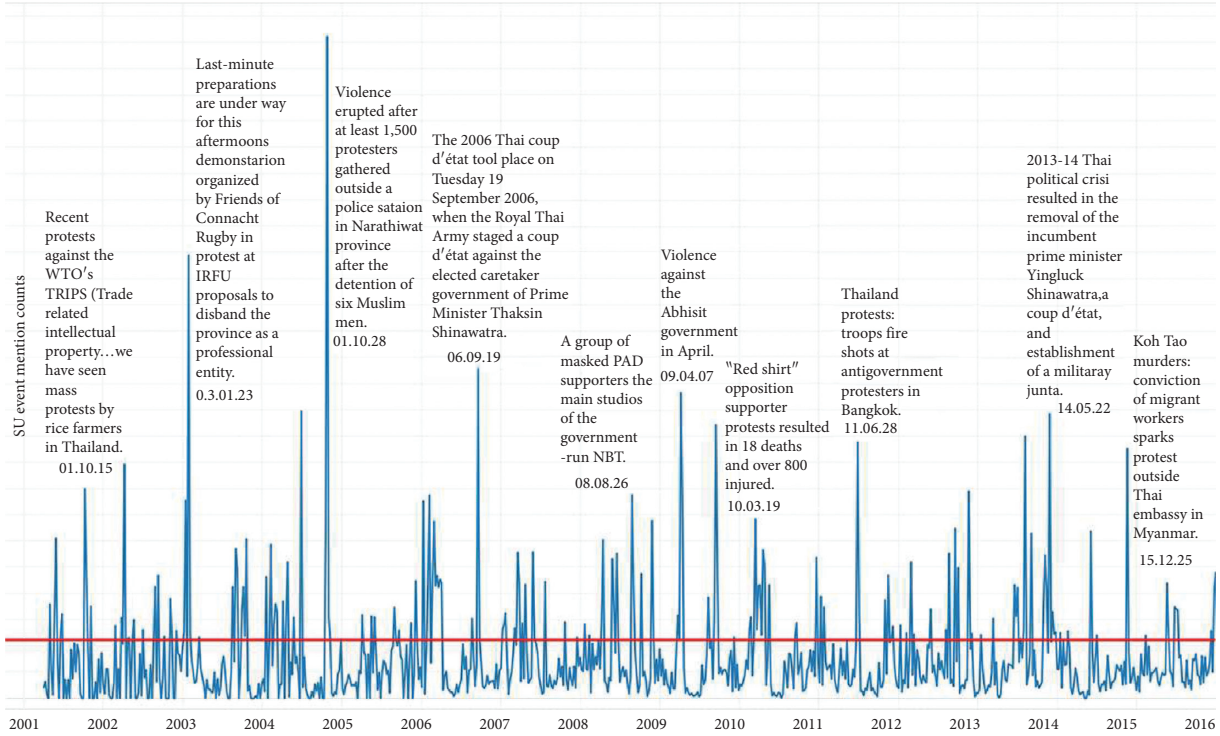

Figure 4: Normalized SU event mention counts of Thailand with annotations for top ten stretches above θ (red line).

3.3. BoEAG Feature Construction. The Bag-of-Event-Association-subGraph (BoEAG) feature is constructed from frequent subgraphs and the bag of word model. The original GDELT data within a certain time are first represented as a big single event element association graph. Then, the frequent subgraph patterns are mined from the big single graph. In the end, the BoEAG features are constructed like the classic BoW (bag of word) model.

The event element association graph draws on the SUBDUE system [54] which analyzed aviation safety events using graph mining. The system converts a series of aviation safety related event records into graph data for processing. The node labels represent the aviation safety event id and the attribute value. The edge labels represent the attribute name (such as location, time, and flight altitude) and the relationship between events. For example, "near_to" relationship means that the distance between the two accidents occurred is within 200 km.

Figure 5 gives a schematic diagram of the event element association graph of this paper. The figure contains two events numbered id1 and id2. The node label in the figure represents the number and attribute value of the GDELT event record, and the edge label represents the attribute name, such as event type, location, participants, and GoldsteinScale value. When two events contain at least one identical participant, there will be a "relate_to" relationship between the two events connected by an edge.

Bag of words model is a feature vectorization method commonly used in the field of text retrieval and text classification. In this paper, BoEAG feature construction is similar to BoW. The collection of GDELT event element association graphs aggregated by day corresponds to the corpus in the BoW model. Each event element association graph corresponds to a document and each frequent subgraph corresponds to a word in the BoW model. The $tf - idf$ weight of the frequent subgraph s of the $i-$th event element association graph can be calculated by the following formula:

$$\begin{cases} tf\,idf\,(s,i,D) = tf\,(s,i) \times idf\,(s,D), \\ tf\,(s,j) = \log\left(1 + f_{s,i}\right), \\ idf\,(s,D) = \log\left(\dfrac{N}{1+n_s}\right), \end{cases} \tag{4}$$

where $f_{s,i}$ denotes the frequency of subgraph s in the event association graph i. This value can be directly obtained through the single graph frequent subgraph mining algorithm SSIGRAM proposed in our previous work [55]. N denotes the number of event association graphs, that is, the time span of the dataset in days; n_s is the number of event association graphs that contain subgraphs s.

Algorithm 1 gives the process of BoEAG feature construction illustrated above. The input of the algorithm includes three parameters: the original GDELT event records, such as a set of event records within a certain period of time in a certain country, the support threshold, and the maximum number of subgraphs. The output is the BoEAG feature vector set. Lines 4 to 19 of the algorithm construct event association graphs. Lines 20–22 use the SSIGRAM algorithm for single large graphs for frequent subgraph mining. The maximum number of subgraphs $N_{\max}$ is to

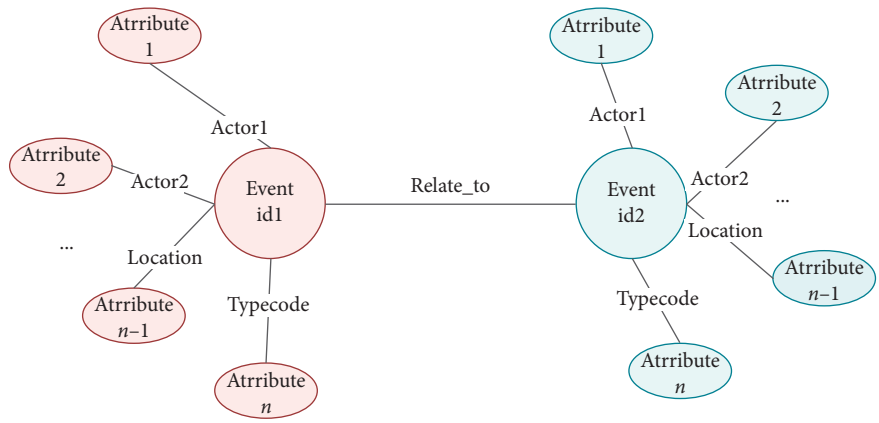

Figure 5: Schematic diagram of the event element association graph.

return the maximum number of subgraphs. That is, when the total number of frequent subgraphs found during the mining process reaches $N_{\max}$, it will stop iterating and arrange all subgraphs in descending order of frequency. Line 24 obtains the standard adjacency matrix coding sequence of each subgraph and uses it as the "Word." Line 25 calculates the $tf-idf$ feature vector corresponding to each event association graph according to formula (4).

3.4. HSMM Training

3.4.1. Structure of HSMM. Usually, the social unrest event has a series of evolutionary stages, 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 eruption. In this paper, a hidden semi-Markov model which contains five states with left and right structure is designed, whose structure is shown in Figure 6.

The structure contains five states, corresponding to the typical stages of the evolutionary process of social unrest events from left to right, such as appeal, accusation, refuse, escalation, and eruption. The state in this structure starts from $S1$ (appeal) and ends at state $S5$ (eruption). During the state transition, the number of the next transition state cannot be lower than the current state number. Correspondingly, the state transition probability matrix A has the following form:

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ 0 & a_{22} & a_{23} & a_{24} & a_{25} \\ 0 & 0 & a_{33} & a_{34} & a_{35} \\ 0 & 0 & 0 & a_{44} & a_{45} \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}. \tag{5}$$

In the traditional HMM model, the state residence time probability $P_i(d)$ shows an exponential downward trend with the number of residence time units [56], which is obviously not consistent with the state residence time of many application scenarios in the real world, especially the social unrest events. In order to improve this shortcoming, the state residence time probability distribution can be explicitly introduced into the HMM model so that it can automatically learn the probability distribution of the state residence time from historical data. This is the original intention of the hidden semi-Markov model.

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 BoEAG feature observations O, a standard continuous HSMM can be defined as $\lambda = (\pi, A, B, P)$, where the initial state probability π and output matrix B have the same meaning as HMM, while the state transition matrix A is defined as

$$a_{ij} = P\left(S_{t+d} = s_j \,\middle|\, S_t = s_i\right), \quad 1 \le i, j \le N. \tag{6}$$

This paper considers the discrete time probability, that is, the state residence time can only be an integer multiple of the residence time unit, e.g., day. Let D represent the maximum possible residence time; then, P can be denoted as a residence time probability matrix of $N \times D$, whose element value p_{id} represents the probability of the state s_i lasting d time units:

$$p_i(d) = P\left(d \,\middle|\, S_t = s_i\right), \quad 1 \le i \le N, 1 \le d \le D. \tag{7}$$

3.4.2. Sequence Likelihood. Given an observation sequence consisting of L days' BoEAG feature vector set $O = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_L)^T$. The goal of hidden semi-Markov model training is to optimize the model parameters π, A, B, and P so that the likelihood of the model generating sequence O is maximized. Given the HSMM model $\lambda = (\pi, A, B, P)$, the sequence likelihood of the observation sequence O is defined as

$$P(O \mid \lambda) = \sum_{s} \pi_1 b_{s_1}(\mathbf{o}_1) \prod_{t=2}^{N} P\left(S_t = s_t \,\middle|\, S_{t-1} = s_{t-1}\right) b_{s_t}(\mathbf{o}_t), \tag{8}$$

where $\mathbf{s} = [s_t]$ represents the hidden state sequence with length N. Similar to the traditional HMM, the sum over $\mathbf{s}$ can also be calculated by the forward-backward algorithm proposed in [57]. The difference is that the state residence time needs to be explicitly added during the derivation

Require: original event records ER, support threshold τ, and maximum subgraphs returned $N_{\max}$
Ensure: BoEAG feature set X_{set}
(1) $\text{EAG}_{\text{set}} \longleftarrow \varnothing$/* The set of event association graphs */
(2) $\text{Sub}\,G_{\text{set}} \longleftarrow \varnothing$/* The set subgraphs */
(3) $X_{\text{set}} \longleftarrow \varnothing$
(4) $\text{DAER}_{\text{list}} \longleftarrow$ ER: event records aggregated by day
(5) for DAER_t in $\text{DAER}_{\text{list}}$ do/* All the event records at date t */
(6) $\text{EAG}_t \longleftarrow \varnothing$/* All the event association graphs at date t */
(7) for e_i in DAER_t do
(8) if e_i is not traversed then
(9) $\text{EAG}_t \longleftarrow$ constructing the graph unit of event e_i
(10) for e_j in DAER_t do
(11) If e_j is not traversed then
(12) $\text{EAG}_t \longleftarrow$ constructing the graph unit of event e_j
(13) if e_i and e_j contain at least one identical participant then
(14) $\text{EAG}_t \longleftarrow$ generating "relate_to" edge between e_i and e_j
(15) end if
(16) end for
(17) end for
(18) $\text{EAG}_{\text{set}} \longleftarrow \text{EAG}_t$
(19) end for
(20) for EAG_t in EAG_{set} do
(21) $\text{Sub}\,G_t \longleftarrow$ SSIGRAM $(\text{EAG}_t, \tau, N_{\max})$/* Mining frequent subgraphs using the SSIGRAM algorithm */
(22) $\text{Sub}\,G_{\text{set}} \cdot \text{add}(\text{Sub}\,G_t)$
(23) end for
(24) Representing each subgraph in $\text{Sub}\,G_{\text{set}}$ as its standard adjacency matrix (CAM) coding sequence (for details of standard adjacency matrix, please refer to [55]).
(25) $X_{\text{set}} \overset{\text{TF-IDF}}{\longleftarrow} \text{Sub}\,G_{\text{set}}$/* Calculating feature set using formula (4) */
(26) Return X_{set}

Algorithm 1: The algorithm of BoEAG feature construction.

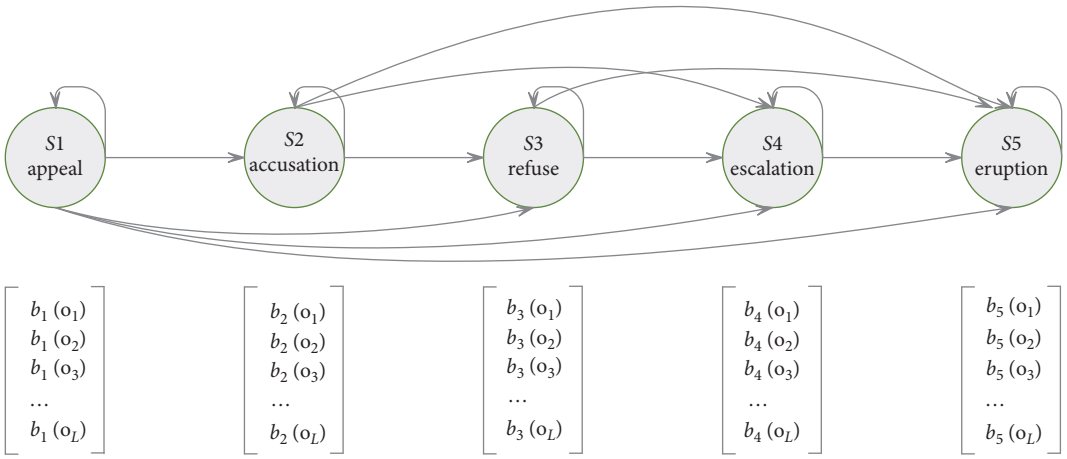

Figure 6: The structure of HSMM with five states.

process. Define $\alpha_t(j)$ as the forward variable, which means the probability of ending at the hidden state j at time t, given observation sequence $O = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_L)^T$:

$$\alpha_t(j) = P\left(\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_t, S_t = j \,\middle|\, S_{t+1} \neq j, \lambda\right), \quad 1 \le j \le N, 1 \le t \le L. \tag{9}$$

$\alpha_t(j)$ can be recursively calculated from front to back as follows:

$$\begin{cases} \alpha_0(i) = \pi_i, & 1 \le i \le N, \\ \alpha_t(j) = \sum\limits_{\substack{i=1 \\ i \ne j}}^{N} \sum\limits_{d=1}^{t} \alpha_{t-d}(i) a_{ij} p_j(d) \prod\limits_{s=t-d+1}^{t} b_j(\mathbf{o}_s), & t = 1, 2, \ldots, L, \quad 1 \le j \le N. \end{cases} \tag{10}$$

Finally, the sequence likelihood can be efficiently computed by

$$P(O \mid \lambda) = \sum_{i=1}^{N} \alpha_L(i). \tag{11}$$

The backward variable is defined as $\beta_t(i)$, which means the probability of starting at the hidden state i at time t, given observation sequence $O = (\mathbf{o}_1, \mathbf{o}_2, \ldots, \mathbf{o}_L)^T$:

$$\beta_t(i) = P\left(\mathbf{o}_{t+1}, \mathbf{o}_{t+2}, \ldots, \mathbf{o}_L, S_t = i \,\middle|\, S_{t+1} \ne i, \lambda\right), \quad 1 \le i \le N, 1 \le t \le L. \tag{12}$$

$\beta_t(i)$ can be recursively calculated from front to back as follows:

$$\begin{cases} \beta_L(i) = 1, & 1 \le i \le N, \\ \beta_t(i) = \sum\limits_{\substack{j=1 \\ j \ne i}}^{N} \sum\limits_{d=1}^{L-t} a_{ij} p_j(d) \prod\limits_{s=t+1}^{t+d} b_j(\mathbf{o}_s) \beta_{t+d}(j), & t = L-1, L-2, \ldots, 1, \quad 1 \le i \le N. \end{cases} \tag{13}$$

The sequence likelihood can be efficiently computed by

$$P(O \mid \lambda) = \sum_{i=1}^{N} \beta_1(i). \tag{14}$$

3.4.3. Parameter Estimation. There are 4 parameters to be estimated for the model training, including initial probability distribution π, state transition probability a_{ij}, observed probability density function $b_i(\mathbf{o}_t)$, and state residence time probability density function $p_j(d)$. π and a_{ij} can be calculated directly. $b_i(\mathbf{o}_t)$ and $p_j(d)$ need to specify the description form of probability density function in advance. We use multivariate mixed Gaussian probability density to describe the probability density of observations $b_i(\mathbf{o}_t)$:

$$b_i(\mathbf{o}_t) = \sum_{m=1}^{M} c_{\text{im}} \mathcal{N}(\mathbf{o}_t; \mu_{\text{im}}, \mathbf{U}_{\text{im}}), \tag{15}$$

where M represents the number of mixed Gaussian elements; c_{im} is the weight of the m mixed Gaussian elements in the state i; $\sum_{m=1}^{M} c_{\text{im}} = 1$; and μ_{im} and $\mathbf{U}_{\text{im}}$ are the mean and variance of the $i-$th Gaussian element, respectively.

We use a single Gaussian distribution to describe the probability density of state residence time $p_j(d)$:

$$p_i(d) = \mathcal{N}\left(d; m_i, \sigma_i^2\right), \tag{16}$$

where m_i and σ_i^2 are the mean and variance, respectively.

Denote the variable $\xi_t(i, j)$ as the probability of transferring from state i to state j after residing in d time units at the time t. Given the observation sequence O and the model parameters λ, then

$$\xi_t(i, j) = P\left(S_t = i, S_{t+d} = j \mid O, \lambda\right). \tag{17}$$

Given the definitions of the forward variable and backward variable, $\xi_t(i, j)$ can be calculated as

$$\xi_t(i, j) = \frac{\alpha_t(i) a_{ij} \sum_{d=1}^{t} \beta_{t+d}(j) p_j(d) \prod_{s=t+1}^{t+d} b_j(o_s)}{\beta_0}. \tag{18}$$

So far, the parameter estimation can be achieved by the expectation maximization (EM) algorithm, also known as the Baum–Welch algorithm in HMM [57]. The E step of the EM algorithm is to construct a Q function and then maximize the Q function in the M step. Thus, we can obtain the re-estimated model parameters π, a_{ij}, $b_i(\mathbf{o}_t)$, and $p_j(d)$. Then, the process iterates continuously until the parameters converge or the maximum number of iterations is reached, formulated as

$$\widehat{\lambda} = \arg\max_{\lambda} P(O \mid \lambda). \tag{19}$$

As the ground truth contains multiple positive samples and negative samples, we need to use multiple sets of observation data to train the model. Denote $O = [O^{(1)}, O^{(2)}, \ldots, O^{(k)}, \ldots, O^{(K)}]$ as the training data containing K

observation sequences. All observation sequences have the same length L. We assume that each observation sequence is independent with each other. $P(O \mid \lambda)$ represents the probability of the combination of observation sequences under a given model; then,

$$P(O \mid \lambda) = \prod_{k=1}^{K} P\left(O^{(k)} \mid \lambda\right). \tag{20}$$

Finally, we trained two HSMMs based on two corresponding set 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 evolution 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, we formalize the social unrest event prediction 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 the limit of double-precision floating point operations may be reached. The scaling technique log-likelihood 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

$$\log\left[P\left(O \mid \lambda_{\mathrm{SU}}\right)\right] - \log\left[P\left(O \mid \lambda_{\overline{\mathrm{SU}}}\right)\right] > \underbrace{\log\left[c_{\overline{\mathrm{SU}},\mathrm{SU}} - c_{\overline{\mathrm{SU}},\overline{\mathrm{SU}}}/c_{\mathrm{SU},\overline{\mathrm{SU}}} - c_{\mathrm{SU},\mathrm{SU}}\right] + \log\left[P(\overline{\mathrm{SU}})/P(\mathrm{SU})\right]}_{\varepsilon \in (-\infty,\infty)}, \tag{21}$$

where c_{ta} denotes the associated cost for assigning a sequence of type t to class a, e.g., $c_{\mathrm{SU},\overline{\mathrm{SU}}}$ denotes the cost for falsely classifying a SU-prone sequence as SU-free. $P(\overline{\mathrm{SU}})$ and $P(\mathrm{SU})$ are constants representing the prior probabilities of SU sequences and $\overline{\mathrm{SU}}$ sequences, respectively (see, e.g., [58] for a derivation of the formula).

Thus, given the costs of misclassification, the right hand side of this inequality determines a constant threshold on the difference of sequence log-likelihood, 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 HSMM-based prediction framework based on five countries from Southeast Asia.

4.1. Experiment Design

4.1.1. Dataset. Our focus area is distributed across five major nations in Southeast Asia: Thailand, Malaysia, Philippines, Indonesia, and Cambodia. These countries have experienced mass protests of varying degrees over the past decade, so they are ideal sources of research data. As mentioned above, GDELT uses the CAMEO coding system [53], where root event code 14 represents social unrest. Figure 7 illustrates the mention counts of protest event occurring in these countries retrieved from GDELT between January 1, 2001, and February 29, 2016. Among them, Thailand (25877 times) was mentioned the most in protest reports, followed by the Philippines (23381 times), and Cambodia (7322 times) being the least. In consideration of the quarterly normalization in Section 3.2, the actual training data were from April 1, 2001, to December 31, 2013, and the test data were from January 1, 2014, to February 29, 2016.

4.1.2. Comparison Methods. As a comparison, three methods are selected in this paper. One is the traditional hidden Markov model (HMM), and its structure is also a form of left to right as Figure 6, except that there is no explicit state residence time probability distribution estimation during the model training process; the remaining steps are the same as the HSMM method. The second is the logistic regression method. Two logistic regression models are trained, and sequence classification is conducted based on this. The third is baseline which does not train any model. It directly uses the probability of protest event records in a country in history as the future social unrest events' probability.

4.1.3. Performance Metrics. We evaluate our social unrest event prediction framework using metrics similar to those described in Kallus et al. [13]. We quantify the success of the proposed predictive mechanism and comparison methods based on their balanced accuracy. Let $T_{\mathrm{ct}} \in \{0, 1\}$ and $P_{\mathrm{ct}} \in \{0, 1\}$, 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 ($T_{\mathrm{ct}} = 1$) 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:

$$\mathrm{BACC} = \frac{\mathrm{TPR} + \mathrm{TNR}}{2}. \tag{22}$$

BACC, unlike the marginal accuracy, cannot be artificially inflated. In fact, due to the unbalanced distribution of positive and negative examples in our dataset, always

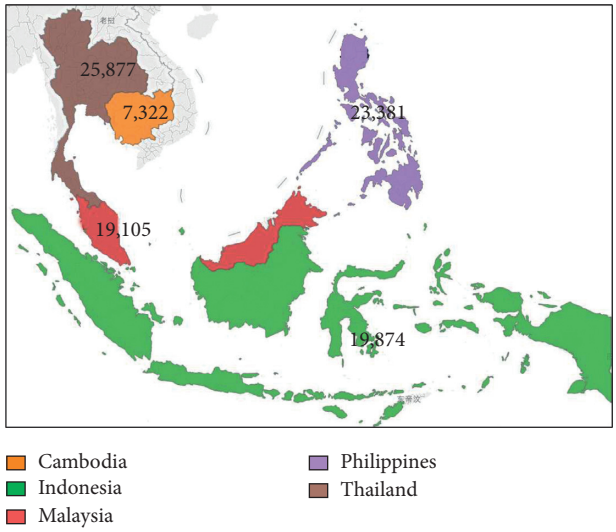

Figure 7: Figure 7Mention counts of protest events occurring in Cambodia, Indonesia, Malaysia, Philippines, and Thailand extracted from GDELT between January 1, 2001, and February 29, 2016.

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.

4.1.4. Parameter Settings. In the extraction stage of ground truth, the threshold value of θ is set to 2.3. This value is approximately equal to the 90% quantile of the standard exponential distribution, that is, approximately 10% of the 7-day time windows in the ground truth will be marked positive.

In the BoEAG feature extraction stage, the maximum number of returned frequent subgraphs $N_{\max}$ is set to 10000. The logistic regression has one parameter: the iteration convergence threshold, which is set to 10^{-6} in the experiment. The baseline method does not require any parameter values to be set in advance. The HMM model and the HSMM model both have 6 parameters that need to be set, including the hidden state number N, the number of mixed Gaussian elements used in the estimation of the probability density of the observation value M, the prediction interval Δt_p, the lead time Δt_l, the prediction data time window Δt_d, and the likelihood threshold ε. In experiments, N, M, and Δt_p are used as fixed parameters, that is, the three values are the same when the experiment is performed on the dataset of five countries. We set $N = 5$, $M = 3$, and $\Delta t_p = 7$, respectively. The meaning of $\Delta t_p = 7$ is to determine whether there will be a social unrest event during the 7-day (one week) time window. In addition, Δt_d, Δt_l, and ε are adjustable parameters, and the optimal value is obtained by performing 10-fold cross-validation on the training set of each country. The value interval of Δt_l is one day to seven days. The values of Δt_d is 10, 20, 30, and 40 days, and the value interval of ε is $[-2, 2]$, with a step of 0.1. The final value details are shown in Table 1.

Table 1: Adjustable parameters (Δt_l, Δt_p, and ε).

Country	HMM			HSMM		
	Δt_l	Δt_p	ε	Δt_l	Δt_p	ε
Thailand	1	10	0.2	1	10	0.1
Malaysia	1	10	0	1	10	0.1
Philippines	1	20	0	1	20	-0.2
Indonesia	1	30	0.2	1	30	0.3
Cambodia	1	10	0.1	1	10	0.1

4.1.5. Ground Set. Table 2 gives the ground-truth results on five datasets of Thailand, Malaysia, Philippines, Indonesia, and Cambodia. The experiment uses 7-day time stretches as the time unit. The time span of the dataset (2001.04.01–2016.02.29) contains a total of 778 7-day time windows. The training period includes 666 7-day stretches while the testing period includes 112. The number of positive stretches in the training set (2001.04.01–2013.12.31) and the test set is listed in the table.

Figure 4 takes Thailand as an example, giving its normalized number of protest reports (the red line represents the threshold θ). We mark the top ten 7-day time windows with the most reports and give a brief description. These are the social unrest events that have really happened in Thailand in the history, such as the "Tak Bai incident" with about 1500 protesters on October 28, 2004, which occurred in Tak Bai district in Southern Thailand, caused by the detention of 6 Muslim believers. And the protest conflict against the Abhisit government which broke out in Bangkok on April 7, 2009, is also included. This also shows the effectiveness of the proposed method of extracting ground truth from GDELT data.

4.2. Event Prediction Results. Table 3 gives the balanced accuracy (BACC) values of the hidden semi-Markov model (HSMM), the traditional hidden Markov model (HMM), the logistic regression, and the baseline method on the test set. Based on the BoEAG feature pattern, it can be seen that in the test datasets of various countries, the performance of the prediction method based on the hidden semi-Markov model proposed in this paper is the best, which shows that the HSMM model can indeed better model the characteristics of mass protest events due to explicitly considering the residence time of the event development evolution stage. The performance of the HMM model is the second best, followed by the logistic regression, and the baseline performs the worst, which is basically random guessing. A longitudinal comparison of the five countries shows that each method performs best in the Thailand test set, especially the HSMM method, which achieves a BACC value of 95.9%. For all the five countries, our proposed HSMM-based approach achieved the best overall performance in balanced accuracy, outperforming the HMM model by 12.7%, 7.1%, 4%, 16.8%, and 5.3% and the logistic regression by 43.6%, 25.8%, 11.2%, 33.5%, and 12.6% for Thailand, Indonesia, Philippines, Malaysia, and Cambodia, respectively.

In addition, comparing each method's performance based on the BoEAG pattern and the temporal burst pattern

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

Country	# of positive 7-day stretches	
	Training	Testing
Thailand	95	12
Malaysia	78	8
Philippines	85	9
Indonesia	83	7
Cambodia	75	7

Table 3: The BACC value of each method.

	BoEAG pattern				Temporal burst pattern			
	HSMM	HMM	Logistic	Baseline	HSMM (%)	HMM (%)	Logistic (%)	Baseline (%)
Thailand	95.9	85.1	69.8	52.8	87.1	86.1	67.8	53.1
Indonesia	75.5	72.5	59.5	50.5	71.0	70.5	60.0	50.8
Philippines	72.6	71.4	68.3	49.5	70.1	69.7	65.3	48.1
Malaysia	78.9	72.3	63.2	50.6	72.3	68.3	59.1	47.6
Cambodia	73.1	75.5	67.5	50.6	71.8	69.4	64.9	51.6

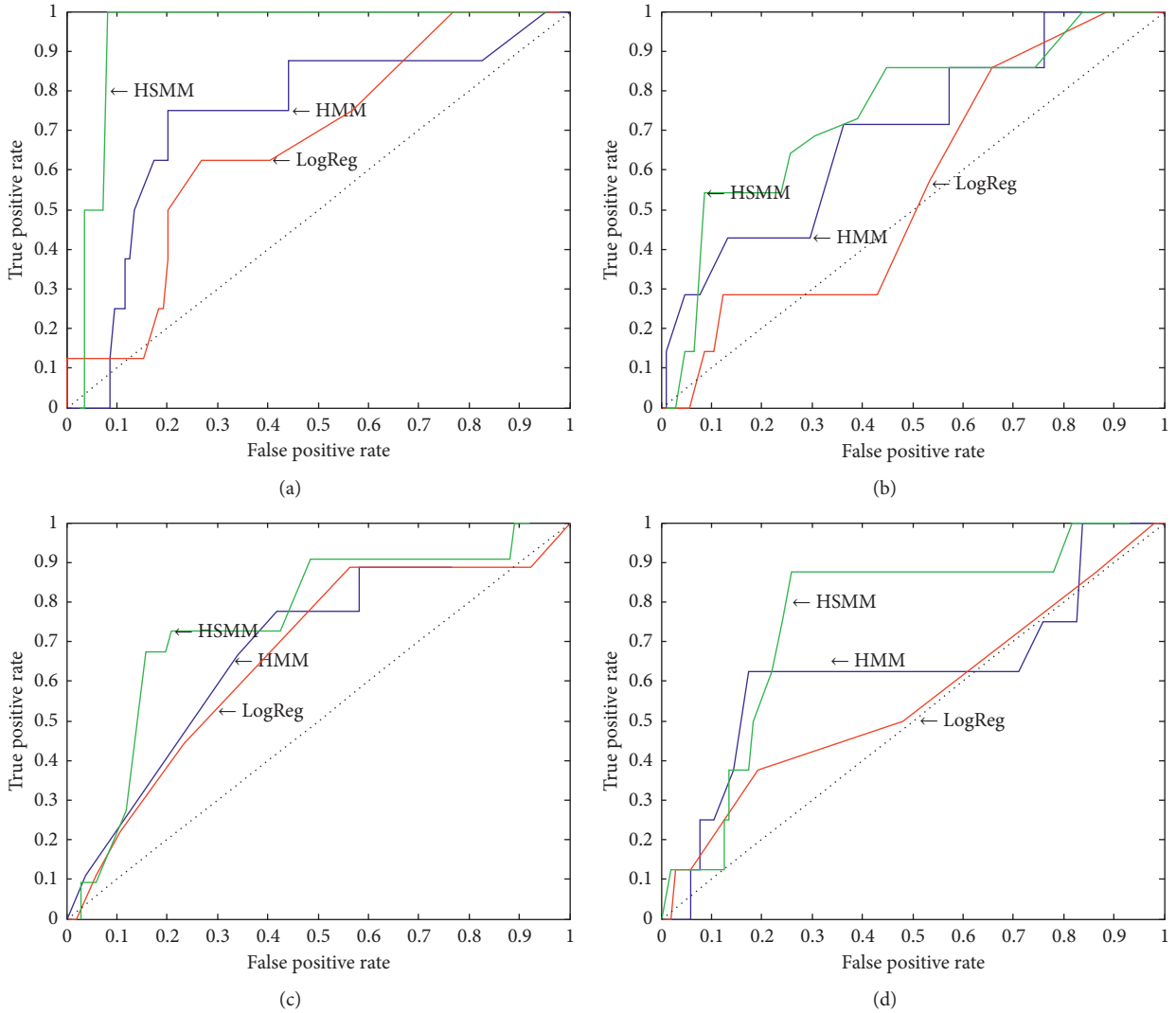

Figure 8: Continued.

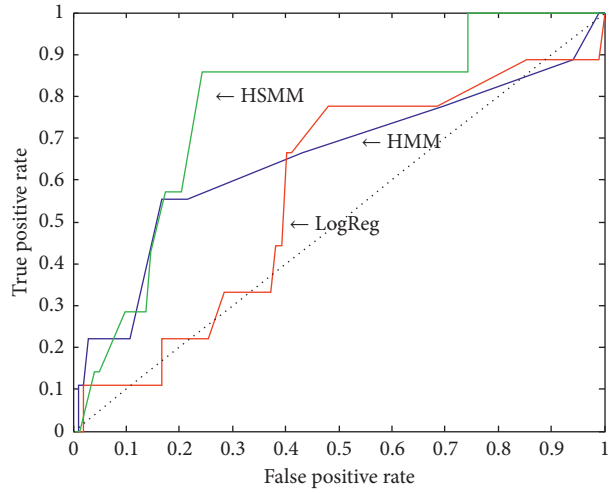

Figure 8: ROC curves for the compared prediction models: HSMM, HMM, and logistic regression. (a) Thailand. (b) Indonesia. (c) Philippines. (d) Malaysia. (e) Cambodia.

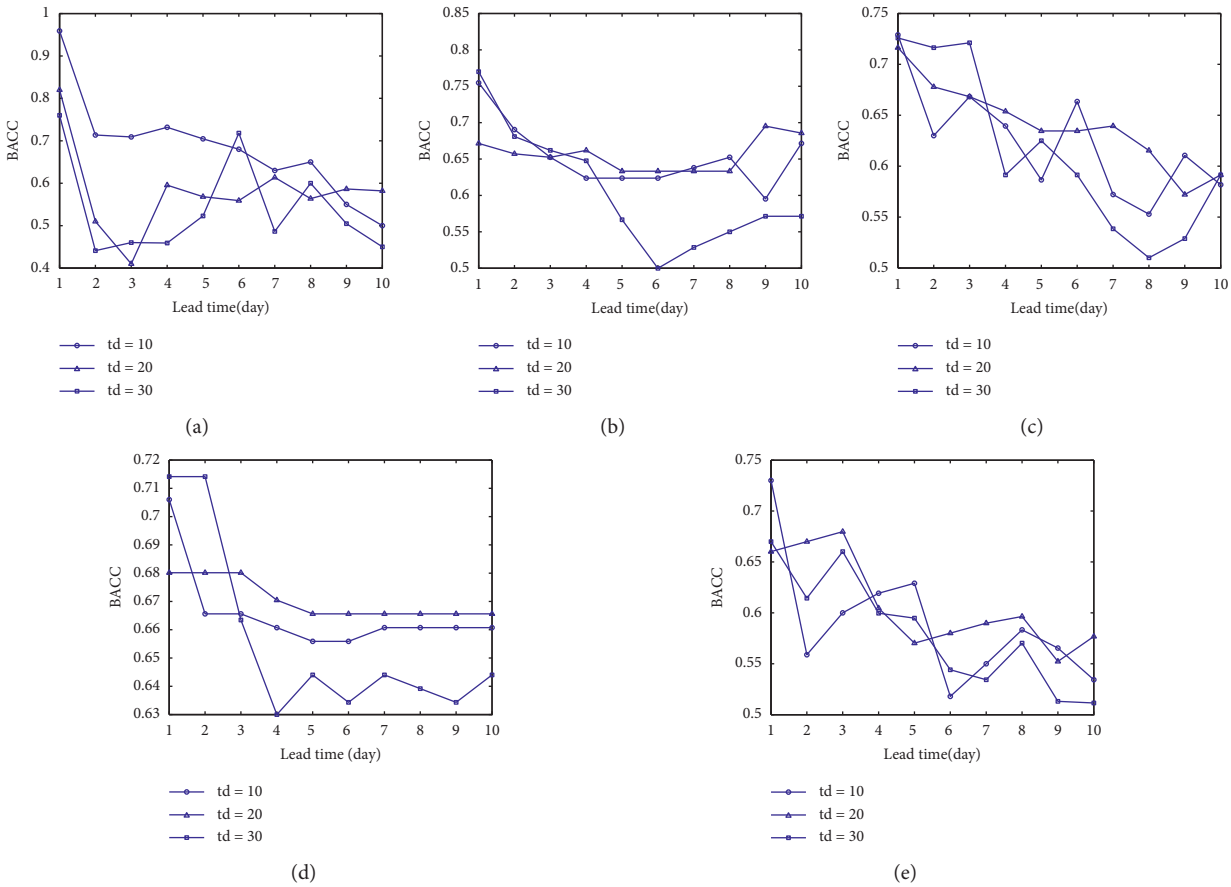

Figure 9: Sensitivity analysis on lead times Δt_l and data window size Δt_d. (a) Thailand. (b) Indonesia. (c) Philippines. (d) Malaysia. (e) Cambodia.

used in our previous work [8], we can see that the BoEAG pattern constructed from frequent subgraphs can better model the stages of social unrest events, as the BACC values of HSMM, HSMM, and logistic regression all improve when the BoEAG patterns are used. This is because the BoEAG pattern considers both temporal burst and the interaction between event participants.

By adjusting the likelihood ratio threshold ε, a series of correspondences between the true positive rate (TPR) and the false positive rate (FPR) can be obtained, and then ROC analysis can be performed for each method. Figure 8 shows the ROC curve of the three methods of HSMM model, HMM model, and logistic regression. The larger the area under the curve (AUC) under the ROC curve, the better the prediction performance of the model. Obviously, among the three methods shown, the AUC of the hidden semi-Markov model (HSMM) is the largest on each test set, and its performance is the best among the methods.

4.3. Sensitivity Analysis on Δt_l and Δt_d. Although the model parameters are fixed on the training set by 10-fold cross-validation, it is still necessary to investigate the performance of the prediction model at different leading time Δt_l and prediction time window Δt_d, which also has guiding significance to the actual application model.

Figure 9 shows the trend of the prediction performance of the HSMM model on each test set with Δt_l and Δt_d. The leading time Δt_l is 1 day to 10 days, and the value of Δt_d is 10 days, 20 days, and 30 days. Two phenomena can be found: First, as the leading time Δt_l increases, the overall prediction accuracy of the model decreases. In most cases, when $\Delta t_l = 1$, the BACC value is the highest. This is consistent with our common sense, that is, the closer the observation data are to the time point of the event, the more accurate the event can be predicted in the future. Second, the performance of the model is not necessarily related to the length of time windows of the observation sequence data used. It is not that the longer the observation sequence used, the higher the prediction accuracy, and the more the data, the more the interference. Given the trained prediction model and the lead time parameters, different test sets require different time windows for prediction data to achieve optimal prediction accuracy.

5. Discussion

This paper presents a hidden semi-Markov model-based framework for leveraging large-scale digital history coded events captured from GDELT to utilize the frequent subgraph patterns mined from the GDELT event streams to uncover the underlying event evolution 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 traditional HMM, the logistic regression model, and the baseline model. It shows that the GDELT dataset does reflect some useful precursor indicators that reveal the causes or evolution of future events.

We plan to conduct our future work in the following three aspects. First, we plan to introduce a multilevel prediction mechanism to our framework, such as city level or province level. Second, in GDELT 2.0, event mention details and global knowledge graphs [59] are also provided in real time, which can bring us with detail insights to the events. More machine learning and deep learning methods like the graph neural networks [60] can be developed with more events' elements. Third, the prediction framework may be improved by distinguishing widespread news coverage from localized coverage.

Data Availability

The GDELT data used to support the findings of this study are included within the article in Section 2.

Conflicts of Interest

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

Acknowledgments

This study was supported by the National Social Science Fund of China under grant no. 2019-SKJJ-C-078.

References

[1] S. Muthiah, B. Huang, J. Arredondo et al., "Planned protest modeling in news and social media," *pages*, in *Proceedings of the AAAI*, pp. 3920–3927, Austin, Tx, USA, January 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, Los Angeles, CA, USA, 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," *Journal of Conflict Resolution*, vol. 41, no. 4, pp. 553–579, 1997.

[5] S. P. Orien, "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 *Proceedings of the 2nd International Conference on Cross-Cultural Decision Making: Focus*, San Francisco, CA, USA, July 2012.

[7] K. Leetaru and P. A. Schrodt, "GDELT: global data on events, location, and tone, 1979–2012," in *Proceedings of the ISA Annual Convention*, vol. 2, Citeseer, San Francisco, CA, USA, April 2013.

[8] F. Qiao, P. Li, X. Zhang, Z. Ding, J. Cheng, and H. Wang, "Predicting social unrest events with hidden Markov models using GDELT," *Discrete Dynamics in Nature and Society*, vol. 2017, Article ID 8180272, 13 pages, 2017.

[9] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas, "Short text classification in twitter to improve information filtering," in *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in*

Information Retrieval, pp. 841-842, ACM, Geneva, Switzerland, July 2010.

[10] B. E. Bagozzi, *Forecasting Civil Conflict with Zero-Inflated Count Models*, Pennsylvania State University, State College, PA, USA, 2011.

[11] 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*, pp. 1799–1808, ACM, New York, NY, USA, August 2014.

[12] S. Deng, H. Rangwala, and Y. Ning, "Learning dynamic context graphs for predicting social events," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1007–1016, ACM, Anchorage, AK, USA, August 2019.

[13] K. Nathan, "Predicting crowd behavior with big public data," in *Proceedings of the 23rd International Conference on World Wide Web*, ACM, Seoul, Korea, pp. 625–630, April 2014.

[14] W. Yang, X. Liu, J. Liu, and X. Cui, "Prediction of collective actions using deep neural network and species competition model on social media," *World Wide Web*, vol. 22, no. 1, 2018.

[15] L. Zhao, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Spatiotemporal event forecasting in social media,"vol. 15, pp. 963–971, in *Proceedings of the SIAM International Conference on Data Mining*, vol. 15, pp. 963–971, SIAM, Vancouver, Canada, April 2015.

[16] S. Ranganath, F. Morstatter, X. Hu, J. Tang, S. Wang, and H. Liu, "Predicting online protest participation of social media users," in *Proceedings of the AAAI*, pp. 208–214, Phoenix, AZ, USA, February 2016.

[17] C. Wu and M. S. Gerber, "Forecasting civil unrest using social media and protest participation theory," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 1, pp. 82–94, 2018.

[18] L. Zhao, F. Chen, and Y. Ye, "Efficient learning with exponentially-many conjunctive precursors to forecast spatial events," *IEEE Transactions on Knowledge and Data Engineering*, vol. 32, no. 10, pp. 1923–1935, 2019.

[19] 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 ICWSM*, vol. 11, no. 122–129, pp. 1-2, Washington, DC, USA, May 2010.

[20] A. Tumasjan, T. Oliver Sprenger, P. G. Sandner, and M. W. Isabell, "Predicting elections with twitter: what 140 characters reveal about political sentiment," in *Proceedings of the ICWSM*, vol. 10, pp. 178–185, Washington, DC, USA, May 2010.

[21] 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, Exeter, UK, October 2009.

[22] Marta Arias, A. Arratia, and R. Xuriguera, "Forecasting with twitter data," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 5, no. 1, p. 8, 2013.

[23] 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.

[24] 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.

[25] S. Muthiah, "Forecasting protests by detecting future time mentions in news and social media," Masters thesis, Virginia Tech, Blacksburg, VA, USA, 2014.

[26] C. Jose, 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.

[27] G. Korkmaz, C. Jose, C. J. Kuhlman, A. Marathe, A. Vullikanti, and N. Ramakrishnan, "Combining heterogeneous data sources for civil unrest forecasting," in *Proceedings of the 2015 International Conference on Advances in Social Network Analysis and Mining, ASONAM*, Paris, France, August 2015.

[28] Y. Keneshloo, C. Jose, G. Korkmaz, and N. Ramakrishnan, "Detecting and forecasting domestic political crises: a graph-based approach," in *Proceedings of the 2014 ACM Conference on Web Science*, ACM, Bloomington, IN, USA, pp. 192–196, June 2014.

[29] K. Radinsky and E. Horvitz, "Mining the web to predict future events," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, ACM, Rome, Italy, pp. 255–264, February 2013.

[30] B. P. Vladimir, J. H. Bond, and D. H. Bond, *Using Hidden Markov Models to Predict Terror before it Hits (Again)*, Springer, Berlin, Germany, 2013.

[31] M. K. Islam, M. M. Ahmed, K. Z. Zamli, and S. Mehbub, "An online framework for civil unrest prediction using tweet stream based on tweet weight and event diffusion," *Journal of Information and Communication Technology*, vol. 19, no. 1, pp. 65–101, 2020.

[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, Berlin, Germany, 2012.

[33] F. Salfner, "Predicting failures with hidden Markov models," in *Proceedings of the 5th European Dependable Computing Conference (EDCC-5)*, pp. 41–46, Budapest, Hungary, April 2005.

[34] S. Basnet, L.-K. Soh, A. Samal, and D. Joshi, "Analysis of multifactorial social unrest events with spatio-temporal k-dimensional tree-based dbscan," in *Proceedings of the 2nd ACM SIGSPATIAL Workshop on Analytics for Local Events and News*, pp. 1–10, Seattle, WA, USA, November 2018.

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

[36] X. Wang, H. Chen, Z. Li, and Z. Zhao, "Unrest news amount prediction with context-aware attention lstm," in *Proceedings of the Pacific Rim International Conference on Artificial Intelligence*, Springer, Nanjing, China, pp. 369–377, August 2018.

[37] W. Yang, X. Liu, J. Liu, and X. Cui, "Prediction of collective actions using deep neural network and species competition model on social media," *World Wide Web*, vol. 22, no. 6, pp. 2379–2405, 2019.

[38] L. Phillips, D. Chase, K. Shaffer, H. Nathan, and S. Volkova, "Using social media to predict the future: a systematic literature review," 2017, https://arxiv.org/abs/1706.06134.

[39] H. P. Nathan, A. L. Buczak, J. T. Zook, P. H. James, B. J. Ellison, and B. D. Baugher, "Crystal cube: multidisciplinary approach to disruptive events prediction," in *Proceedings of the International Conference on Applied Human Factors and Ergonomics*, Springer, Orlando, FL, USA, pp. 571–581, July 2018.

[40] L. Zhao, Q. Sun, J. Ye, F. Chen, C.-T. Lu, and N. Ramakrishnan, "Feature constrained multi-task learning models for spatiotemporal event forecasting," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 5, pp. 1059–1072, 2017.

[41] C. Wu and M. S. Gerber, "Forecasting civil unrest using social media and protest participation theory," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 1, pp. 82–94, 2017.

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

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

[44] E. Y. James, *Predicting Future Levels of Violence in Afghanistan Districts Using Gdelt*, BibSonomy, Kassel, Germany, 2013.

[45] L. E. Baum and T. Petrie, "Statistical inference for probabilistic functions of finite state Markov chains," *The Annals of Mathematical Statistics*, vol. 37, no. 6, pp. 1554–1563, 1966.

[46] Y. Guédon and C. Cocozza-Thivent, "Explicit state occupancy modelling by hidden semi-Markov models: application of Derin's scheme," *Computer Speech & Language*, vol. 4, no. 2, pp. 167–192, 1990.

[47] J. Bulla and I. Bulla, "Stylized facts of financial time series and hidden semi-Markov models," *Computational Statistics & Data Analysis*, vol. 51, no. 4, pp. 2192–2209, 2006.

[48] A. Maruotti, A. Punzo, and L. Bagnato, "Hidden Markov and semi-Markov models with multivariate leptokurtic-normal components for robust modeling of daily returns series," *Journal of Financial Econometrics*, vol. 17, no. 1, pp. 91–117, 2019.

[49] M. Dong, D. He, P. Banerjee, and J. Keller, "Equipment health diagnosis and prognosis using hidden semi-Markov models," *The International Journal of Advanced Manufacturing Technology*, vol. 30, no. 7-8, pp. 738–749, 2006.

[50] V. D. Thi, H. B. Hung, Q. P. Dinh, and S. Venkatesh, "Activity recognition and abnormality detection with the switching hidden semi-Markov model,"vol. 1, pp. 838–845, in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 838–845, IEEE, San Diego, CA, USA, June 2005.

[51] V. S. Barbu and N. Limnios, *Semi-Markov Chains and Hidden Semi-Markov Models Toward Applications: Their Use in Reliability and DNA Analysis*, Vol. 191, Springer Science & Business Media, Berlin, Germany, 2009.

[52] F. Salfner and M. Malek, "Using hidden semi-Markov models for effective online failure prediction," in *Proceedings of the 2007 26th IEEE International Symposium on Reliable Distributed Systems (SRDS 2007)*, IEEE, Beijing, China, pp. 161–174, October 2007.

[53] J. G. Deborah, 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, 2002.

[54] S. K. Nikhil, L. B. Holder, and D. J. Cook, "Subdue: compression-based frequent pattern discovery in graph data," in *Proceedings of the 1st International Workshop on Open Source Data Mining: Frequent Pattern Mining Implementations*, ACM, Maebashi, Japan, pp. 71–76, August 2005.

[55] F. Qiao, X. Zhang, P. Li, Z. Ding, S. Jia, and H. Wang, "A parallel approach for frequent subgraph mining in a single large graph using Spark," *Applied Sciences*, vol. 8, no. 2, p. 230, 2018.

[56] Y. Peng, *Research on equipment health prediction and systematic maintenance strategy based on degenerate implicit semi-Markov model*, Ph.D. thesis, Shanghai Jiao Tong University, Shanghai, China, 2011.

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

[58] R. O. Duda, P. Hart, and D. G. Stork, *Pattern Classification*, John Wiley & Sons, Hoboken, NJ, USA, 2012.

[59] Gdelt 2.0: Our Global World in Realtime, The Official GDELT Project Blog, USA, 2015, http://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/.

[60] T. N. Kipf, *Deep learning with graph-structured representations*, Ph.D thesis, University of Amsterdam, Amsterdam, Netherlands, 2020.